Université de Lille

Thèse

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

En vue de l'obtention du

DOCTORAT EN MECANIQUE

Délivré par l'Université de Lille
Préparée au sein du laboratoire LaMcube

Présentée et Soutenue le 08/12/2025 par :

Guillaume Bauman

Sous la direction de Vincent Magnier et Hazem Wannous

Machine Learning pour la tribologie des freins: prédiction de la pollution, analyse des mécanismes et stratégies de contrôle

Machine Learning in Brake Tribology:
Pollution Prediction, Mechanism Analysis and Control Strategies

Jury :

Nom	Grade	Fonction	Établissement
Xavier Chiementin	Président	Professeur des universités	Université de Reims
Faten Chakchouk	Rapporteuse	Professeure des universités	Paris Panthéon–Assas Université
Franck Massa	Rapporteur	Professeur des universités	Université Polytechnique Hauts de France
Alizée Bouchot	Examinatrice	Maîtresse de conférences	INSA de Lyon
Maxime Devanne	Examinateur	Maître de conférences	Université de Haute–Alsace
Vincent Magnier	Directeur	Professeur des universités	Université de Lille
Hazem Wannous	Co-directeur	Professeur des universités	IMT Nord Europe
Nikzad Motamedi	Invité	Maître de conférences	IMT Nord Europe

ENGSYS 632
Ecole Doctorale

LaM CUBE
Laboratoire de mécanique, multiphysique, multiéchelle

Région Hauts-de-France

Remerciements

Je tiens à exprimer toute ma gratitude à mes directeurs de thèse pour leur encadrement, leurs conseils précieux et leur soutien tout au long de ces années de recherche. Vincent Magnier, par son accompagnement rigoureux, sa pédagogie et sa disponibilité, et Hazem Wannous, par ses idées stimulantes, son regard critique et sa maîtrise technique, ont chacun contribué de manière essentielle à la réalisation de cette thèse. Je remercie chaleureusement les rapporteurs, Faten Chakchouk et Franck Massa, pour le temps consacré à l'évaluation de ce travail et pour la richesse de leurs commentaires. J'adresse également mes remerciements aux membres du jury, Alizée Bouchot, Maxime Devanne et Xavier Chiementin, pour l'honneur qu'ils me font en acceptant de participer à la soutenance. Enfin, je souhaite remercier le LaMcube, l'université de Lille et la région Hauts-de-France pour avoir rendu cette recherche possible.

Je suis reconnaissant envers mes collègues et ami·e·s du laboratoire, en particulier Nikzad, Nagesh, Maël, Rongfei et Sacha, pour leurs échanges stimulants, leur aide précieuse et leur bonne humeur quotidienne.

Je voudrais adresser mes plus profonds remerciements à ma famille — en particulier à mes parents, Étienne et Sophie, ainsi qu'à ma sœur Maude, pour leur soutien indéfectible et leurs encouragements constants tout au long de ce parcours. Je pense aussi à mes ami·e·s, Adèle, Amandine, Charles, Gaïa, Geoffrey, Ilya, Martin et Robert, qui ont su m'apporter leur soutien, leurs encouragements et de précieux moments de légèreté en dehors du laboratoire. Enfin, une pensée toute particulière pour Héloïse, dont la gentillesse, le soutien et la douceur ont été une source précieuse d'équilibre tout au long de ces années. Sa présence bienveillante, sa capacité à m'encourager dans les moments de doute et à célébrer les petites victoires ont nourri ma motivation tout au long de cette thèse.

Abstract

Disk brake systems exemplify the complexity of tribological contacts, where mechanical, thermal, acoustic, and chemical phenomena interact in strongly coupled and evolving ways. These interactions give rise to practical concerns such as noise and particulate emissions, which remain difficult to capture with physics-based approaches alone. This thesis investigates how machine learning can extend the analysis and optimization of brake tribology.

Using a multimodal experimental dataset that records acoustic, thermal, mechanical, and particle-emission signals, we first examine predictive models for braking noise and pollution. We then move beyond prediction by applying clustering and classification methods to acoustic spectrograms, complemented by interpretability techniques, which reveal recurring signal states and the variables that influence them. Finally, we couple a finite-element brake-like simulator with reinforcement learning in a proof-of-concept study, demonstrating that reinforcement learning can discover adaptive control policies that enhance the control and stability of disk brakes.

While these contributions remain exploratory and not yet at the level of industrial standards—whether for real-time prediction, mechanism understanding, or control—they highlight promising directions and suggest that data-driven approaches could ultimately be generalized to vehicle-scale applications. Taken together, the work addresses open challenges of granularity, interpretability, and methodological breadth in machine learning applied to brake tribology research, illustrating how machine learning can complement experimental and numerical tools in advancing the design and operation of braking systems.

Résumé

Les systèmes de freins à disque illustrent la complexité des contacts tribologiques, où des phénomènes mécaniques, thermiques, acoustiques et chimiques interagissent de manière fortement couplée et évolutive. Ces interactions soulèvent des enjeux pratiques, notamment le bruit généré et les émissions particulaires, qui sont difficiles à appréhender uniquement par des approches physiques classiques. Cette thèse explore comment le Machine Learning peut enrichir l'analyse et l'optimisation de la tribologie des freins.

À partir d'un jeu de données multimodal combinant des mesures acoustiques, thermiques, mécaniques et d'émissions de particules, nous examinons d'abord des modèles prédictifs pour le bruit émis et les émissions particulaires. Nous allons ensuite au-delà de la prédiction en appliquant des méthodes de clustering et de classification à des spectrogrammes acoustiques. Associées à des techniques d'interprétabilité, elles révèlent des états de signal récurrents ainsi que les variables qui les influencent. Enfin, nous couplons un simulateur par éléments finis d'un système de frein avec du reinforcement learning dans une étude de faisabilité, montrant que cette approche peut découvrir des politiques de contrôle adaptatives améliorant la maîtrise et la stabilité des freins à disque.

Bien que ces travaux demeurent exploratoires et encore éloignés des standards industriels — qu'il s'agisse de la prédiction en temps réel, de la compréhension des mécanismes ou du contrôle — ils ouvrent des perspectives prometteuses et laissent entrevoir la possibilité de généraliser de tels modèles à l'échelle du véhicule réel. Dans leur ensemble, ces contributions abordent des challenges liés à la granularité, à l'interprétabilité et à la diversité méthodologique dans le domaine du Machine Learning appliqué à la tribologie du freinage, illustrant comment le Machine Learning peut compléter les outils expérimentaux et numériques pour améliorer la conception et le fonctionnement des systèmes de freinage.

Contents

Introduction

Contact between solids is a quintessential tribological problem: it is *multiscale*, *multiphysics*, and *evolutionary*. From asperity contacts to full assemblies, mechanical loading, frictional heating, and interfacial chemistry continually reshape the surfaces in contact. A sound understanding of these coupled processes enables better designs, reduces energy and material losses, and prevents the operational disturbances that contact can trigger.

Friction brakes are a prime example. Widely deployed in automotive and railway systems, disk brakes must satisfy competing goals: *safety* under extremes, *comfort* for users (noise, vibration, harshness), *durability* and cost for manufacturers, and increasingly, *air-quality* constraints due to non-exhaust particulate emissions. Two concerns dominate practice: (i) squeal and related acoustic phenomena that degrade comfort and may indicate interfacial instabilities, and (ii) the release of fine and coarse particles during braking, with associated environmental and maintenance costs. These challenges are exacerbated by the intrinsic difficulty of the contact: rough, heterogeneous, and continually evolving third-body layers; strong coupling among mechanical, thermal, chemical, and vibro-acoustic fields; and a persistent gap between laboratory benches and in-service conditions.

This introduction sets the stage for the thesis. First, we consolidate the non-ML physics of **disk brake tribology**, reviewing interfacial phenomena, what test benches capture versus in-situ reality, and why multiscale/multiphysics couplings matter. Second, we survey **machine learning for disk brake tribology**, highlighting common formulations, targets, sensors, and current limitations. We then briefly introduce this thesis contributions.

Disk Brake Tribology

Before turning to data-driven methods, we anchor the discussion in the physics of the pad–disc interface. Disk brake tribology governs what any model can (and cannot) learn in disk brake applications: friction emerges from an evolving third body on rough, heterogeneous surfaces; thermal fields reshape contact; and composition, geometry, and environment steer wear and noise.

Key tribological phenomena at the pad–disc interface

Frictional braking is governed by the coupled evolution of friction, wear, vibration and chemistry at the sliding interface. Early foundational work established how pad surfaces transform under load into complex "tribological surfaces" with third-body plateaus that stabilize friction and modulate wear [16, 15]. Subsequent system-oriented views linked mesoscale contact dynamics to macroscopic brake behavior, including intermittency and self-organization of hard patches within the friction layer [45].

Material formulation strongly conditions these processes. Micro-/nano-additives and reinforcements (e.g., CNTs, graphite forms, short carbon fibers) tune friction stability, fade, and wear through their effects on film formation, thermal transport, and debris mechanics [26, 53, 1]. For high-energy rail brakes, copper-based pads develop distinct tribolayers and wear mechanisms under extreme speeds (e.g., 380 km/h) [67], while the contact-surface state of sintered Cu materials evolves dramatically under severe braking, with performance tied to the developing third-body architecture [68]. Environment and operating conditions also matter: low temperatures and icing change friction/wear pathways and stability [47]. Surface engineering of discs (e.g., WC-reinforced laser claddings) is a parallel route to reduce wear and airborne particle emissions while maintaining frictional performance [38].

Highly instrumented test benches vs. in-situ reality

Instrumentation in laboratory tribology benches has become increasingly sophisticated, enabling precise observation of coupled thermal, vibrational, and wear processes. For example, reduced-scale inertia benches have been equipped with torque sensors and high-resolution thermal monitoring (e.g.[36]), high-speed infrared thermography has been applied to map transient rotor temperatures (e.g.[61]), and combined transducers and thermocouples have been used to construct standardized pvT friction maps (e.g.[63]). Severe thermal conditions have been studied with exhaust gas analysis and in-situ mass tracking (e.g.[37]), while acoustic emission sensors on dynamometers have revealed links between squeal-related vibration bands and material degradation (e.g.[74]). Finally, particulate monitoring with optical counters and laser diffraction has quantified airborne particle distributions during accelerated wear (e.g.[55]).
Yet, despite such advances, laboratory tribometers and dynamometers only approximate service conditions. Careful comparisons show systematic differences between reduced pin-on-disc (PoD) and inertia-dynamometer tests—PoD often achieves steady states and can overexpress certain vibration bands, whereas dynamometers capture more realistic transients and thermal trajectories [50]. Rail tribology studies highlight that reduced-scale benches can be representative if similitude rules are respected, though compromises are inevitable (contact load histories, ventilation, scale-coupled wear morphologies) [14]. Beyond the rig choice, what is measured (and how) is crucial; standardization of friction metrics and protocols remains an active topic [30]. Taken together, these examples illustrate that the debate is not only "test benches vs. in-situ reality," but also how far instrumentation can bridge the gap between them.

Multiphysics coupling

The brake interface is quintessentially multiphysics: frictional heating modifies material response (expansion, softening, oxidation), which redistributes contact, which in turn alters heat generation and debris flux. In high-energy service scenarios such as emergency braking in high-speed TGV trains, interface temperatures can reach up to 1000 °C; combined with the inherently coupled nature of the thermo-mechanical-wear fields, this makes braking a genuinely high-intensity and therefore complex process. Modern FE-based studies therefore treat thermal, mechanical, and wear fields in a *fully coupled* manner to predict temperature rise, contact pressure migration, and wear evolution consistently [11]. Comprehensive strategies now integrate wear remeshing, ALE schemes, and temperature-dependent laws to capture degradation pathways in railway disc brakes [71]. Omitting any of these couplings (e.g., wear or thermal expansion) can yield large prediction errors under high-energy events.

Multiscale and multilevel viewpoints

From asperity-scale plasticity and tribofilm chemistry to component-level thermoelastic instabilities and system-level vibroacoustics, brake tribology is intrinsically multiscale. Cellular/patch-based perspectives connect microstructural evolution of third bodies to macroscopic friction laws and variability [45]. On the dynamics side, incorporating contact-tribology parameters into stability analyses clarifies the routes by which frictional interfaces excite squeal and other self-excited vibrations [23]. In parallel, materials/process studies at the micro/meso scale (e.g., reinforcement architecture, surface engineering) are increasingly mapped to macro responses (fade, wear rate, NVH) using coupled models and well-designed bench campaigns [1, 38].

Transition to Machine Learning.

From the previous sections, it becomes clear that modern tribological experiments generate rich and heterogeneous data, reflecting a system whose mechanical, thermal, and acoustic responses are strongly coupled, nonlinear, and sensitive to operating history. Traditional analyses—often based on averaged quantities or simplified correlations—struggle to capture this complexity. Machine learning offers a complementary path forward: it can reveal structure within multimodal datasets, quantify dependencies among signals, and provide interpretable predictions that bridge experimental and numerical perspectives. The next section reviews how such approaches have been applied to disk-brake tribology and how the present work builds upon and extends these efforts.

Machine Learning applied to Disk Brake Tribology

Machine learning (ML) is emerging as a bridge between traditional tribological experimentation and data-driven analysis. In friction brakes, it enables the discovery of patterns in complex datasets that couple mechanical, thermal, and acoustic phenomena with operating conditions. By capturing interactions that are often intractable for purely physics-based models, ML provides complementary tools for prediction, interpretation, and control within tribological systems.

In the literature, the dominant approach is *supervised learning*, where one or several tribological quantities are predicted from measurable inputs such as speed, pressure, temperature history, material composition, surface metrics, or sensor signals. Training data are typically obtained from controlled laboratory rigs (pin-on-disc or dynamometer), where loads and speed profiles are repeatable, while a smaller number of studies rely on instrumented vehicles or racing telemetry to capture in-use conditions. Within this landscape, existing work mainly falls into the following application areas:

Friction coefficient and wear. Early demonstrations established that relatively simple neural networks already outperform linear baselines on dynamometer measurements for friction coefficient and wear [3]. Subsequent studies generalized this approach to pad recipes and composites, using regression algorithms to map braking conditions and formulation variables to μ and wear endpoints [54, 48, 17, 52, 6]. A key step forward was introduced by Sellami *et al.* [52], who proposed a **multi-target** framework that predicts μ and wear simultaneously and reports feature-level explanations—an important move toward interpretability. Similar supervised regressors have been used for metallic systems, such as random forests for beryllium bronze and aluminium alloys [34], and for "triboinformatic" modelling of aluminium

alloys [22].

Wear life prediction. Several papers target **service life** rather than instantaneous properties. Choudhuri and Shekhar [13] model pad-wear progression from dynamometer data, and Cao *et al.* [10] use ANNs to forecast pad-wear life across braking cycles. While nominally time-resolved, these wear data are typically aggregated over particle-size channels, meaning that the models track total material loss rather than discriminating between underlying wear mechanisms.

Torque and temperature. Machine learning has also been employed as a **virtual sensor**. Han *et al.* [21] compared algorithms and reported accurate disc-brake **temperature** prediction in commercial vehicles. Bonini *et al.* [7] estimated **braking torque** from measured variables, and Bonini *et al.* [8] extended torque estimation to MotoGP contexts using race telemetry. A noteworthy departure from the aggregate-target trend is the doctoral work of Motamedi [44], which trained ML models to **forecast time- and sensor-resolved temperature signals** given braking force.

Surrogates and virtual sensing in design. Beyond direct tribological targets, ML is increasingly embedded as a **surrogate** within engineering workflows. Bao *et al.* [29] outlined an intelligent tribological forecasting system, while Antanaitis [4] applied ML to enable virtual development for high-performance braking. Yang *et al.* [70] integrated high-precision ML surrogates into **reliability-based robust optimization**. In many of these studies, training data combine **controlled tests and simulations**, with targeted vehicle-level checks to verify transferability.

Brake squeal detection. Machine-learning methods have also been widely applied to the study of vibration and acoustic phenomena, particularly **brake squeal**. Stender *et al.* [60] trained deep models on vibration spectrograms to detect squeal events. In related work, Stender *et al.* [59] combined recurrence quantification with ML to characterize vibration states, while Geier *et al.* [18] constructed ML-based **state maps** linking modal dynamics to instability. Yang *et al.* [69] optimized modal behavior using ANNs, and Song *et al.* [57] enhanced CAE noise predictions through ML-assisted modeling.

Fault and anomaly detection. Closely related studies address the broader problem of system health monitoring. Li *et al.* [35] developed an ML-based method for mine-hoist brake fault identification. Jegadeeshwaran and Sugumaran [28] classified hydraulic-brake faults from vibration features, while Yin *et al.* [72] combined signal processing and ML to detect and forecast anomalies under varying operating conditions.

Overall, studies spanning from Friction, wear torque, temperature, Brake squeal and fault diagnosis show machine-learning's ability to reproduce tribological responses and to capture cross-couplings among measured variables. Recent reviews likewise conclude that supervised prediction dominates the field, whether models are trained on controlled rigs, simulation surrogates, or instrumented-vehicle datasets [58, 73, 51].

Synthesis

Despite a wide range of modeling approaches (RF, XGB, SVR, ANN, CNN, RNN, GPR) and sensing modalities (vibration, AE, IR, telemetry), most studies share a similar formulation: supervised mappings from inputs to scalar or binary (sometimes time-resolved) targets,

typically trained on rig/dynamometer data. Three limitations stand out: (i) **Granularity:** studies tend to favor predicting aggregate endpoints over *time- and resolution-resolved* signals; (ii) **Interpretability:** feature attribution/importance is rarely undertaken; and (iii) **Methodological breadth:** unsupervised learning is uncommon and reinforcement learning is largely absent.

Thesis overview and contributions

Building directly on the limitations summarized in above, this thesis advances machine learning for disk-brake tribology along three axes.

Chapter 2 : We develop models that predict brake *sound* and *particulate emissions* both *time-resolved* and *resolution-aware*—i.e., on *frequency bins* for acoustics and *particle-size bins* for particulate emissions.

Chapter 3 : We move beyond raw prediction to *understanding* the mechanisms behind acoustic responses. Spectrogram segments are clustered into recurring states and then linked to auxiliary variables via classification with explainability (permutation feature importance and Integrated Gradients).

Chapter 4 : In this chapter, we move from understanding tribological responses to exploring active control strategies. Since experimental setups cannot easily support control-policy exploration, a simplified FEM environment was developed to emulate thermo-mechanical coupling under controllable boundary conditions. We frame a proof-of-concept control problem on an FE-based brake-like simulator that couples mechanical loading and thermal evolution.

To our knowledge, these three contributions are novel within brake-disc tribology and respectively address the *granularity*, *interpretability*, and *methodological breadth* research gaps.

Chapter 1

Multimodal Test Bench and Measurement Protocol

This chapter introduces the experimental dataset that serves as the foundation for the machine learning analyses developed in Chapter 2 and 3. We first present the experimental setup and the protocol used to generate the data (Section 1.1 and 1.2). Then, we explore statistical properties of the resulting signals (Section 1.4), focusing on their variability, potential correlations, and how they inform feature selection and modeling strategies in subsequent chapters.

1.1 Experimental setup

The experimental platform used in this study is a custom-designed tribological test bench developed for the in-depth analysis of braking systems. It is equipped with multi-modal instrumentation to enable synchronized acquisition of mechanical, thermal, acoustic, and environmental data, as well as particle emissions. This configuration provides a controlled environment for studying friction-induced phenomena during braking events in detail. An overview of the experimental platform is presented in Fig. 1.1.

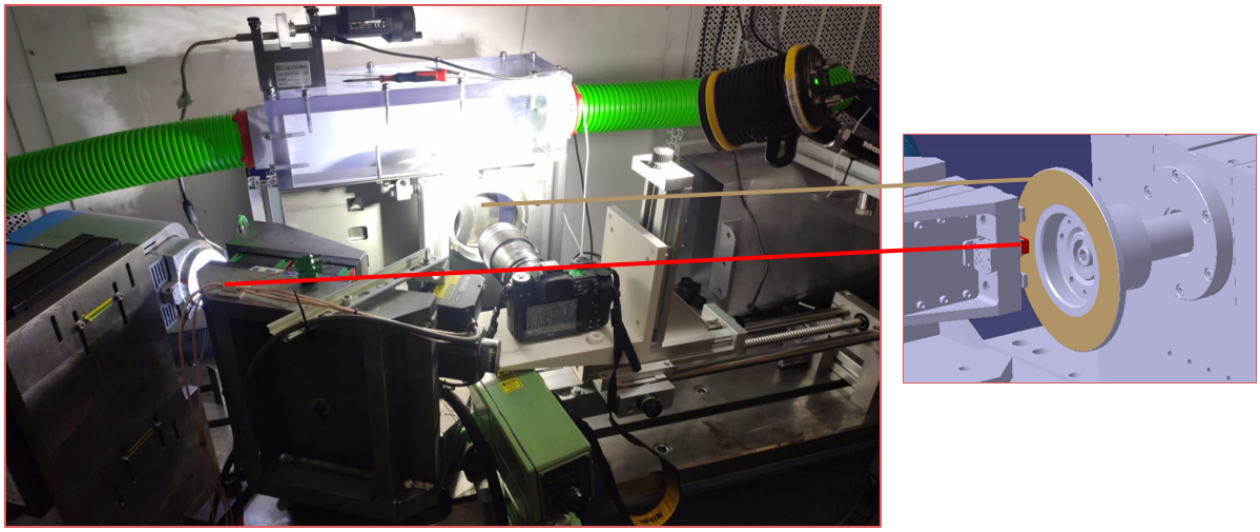


Figure 1.1: Tribological test bench

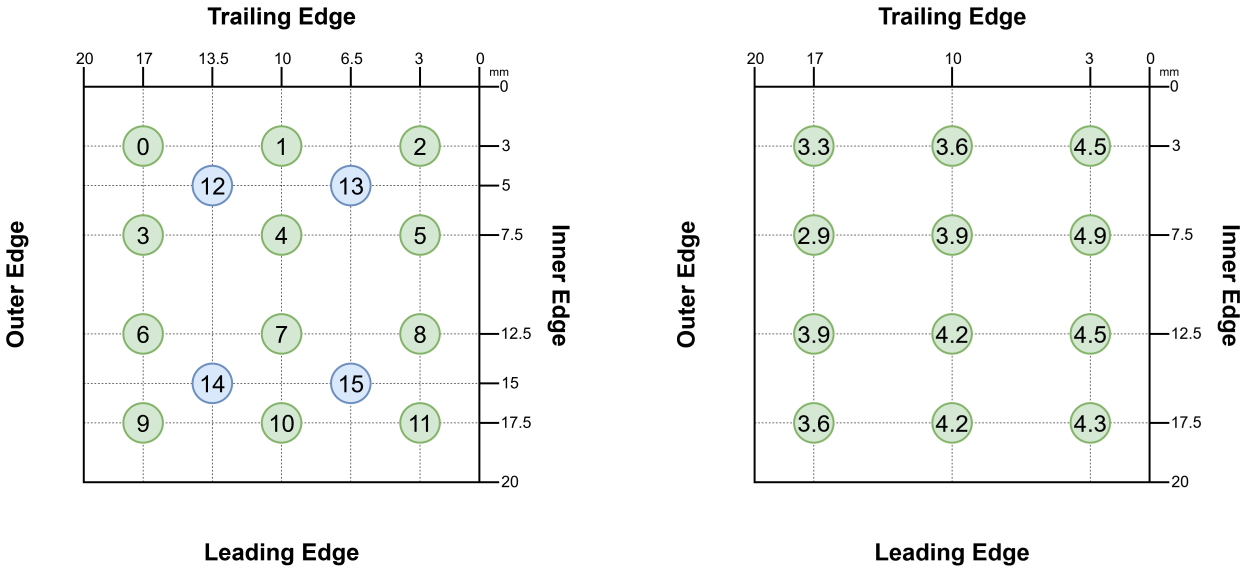
At its core, the test bench operates on a pin-on-disc configuration, where a brake pad is pressed against a rotating disc under controlled conditions. The disc is driven by a motor,

and the normal force applied between pad and disc is precisely regulated. Tangential forces (due to friction), torque, and relative displacements are continuously monitored using high-precision sensors, enabling detailed analysis of the mechanical interactions during braking.

A thermal instrumentation network is integrated into the platform to monitor the heat generated from the contact:

- Thermocouples embedded within the brake pad capture subsurface temperature gradients,
- A pyrometer measures the disc surface temperature in real time during braking events.

Figure 1.2 presents the spatial arrangement of the thermocouples embedded in the brake pad and indicates their grouping by depth. The depths of the shallow thermocouples are known and shown in the second figure.



Thermocouple locations Shallow thermocouple depths (mm)

Figure 1.2: Spatial arrangement of the thermocouples (TC0–TC15) inside the brake pad. Green thermocouples are shallow and the blue ones are deeper.

Acoustic emissions are recorded using a high-sensitivity microphone positioned near the contact interface, allowing for the detection and analysis of noise phenomena such as brake squeal and low-frequency vibrations.

The test environment includes a humidity control system to ensure stable and reproducible ambient conditions, which is essential for consistent and comparable test results.

Finally, a major feature of the setup is the particle capture enclosure system, designed to contain and analyze wear debris generated during braking. A filtered air supply directs airborne particles to sampling points equipped with two complementary instruments:

- An Engine Exhaust Particle Sizer (EEPS), covering a particle size range from 5.6 to 523 nm,

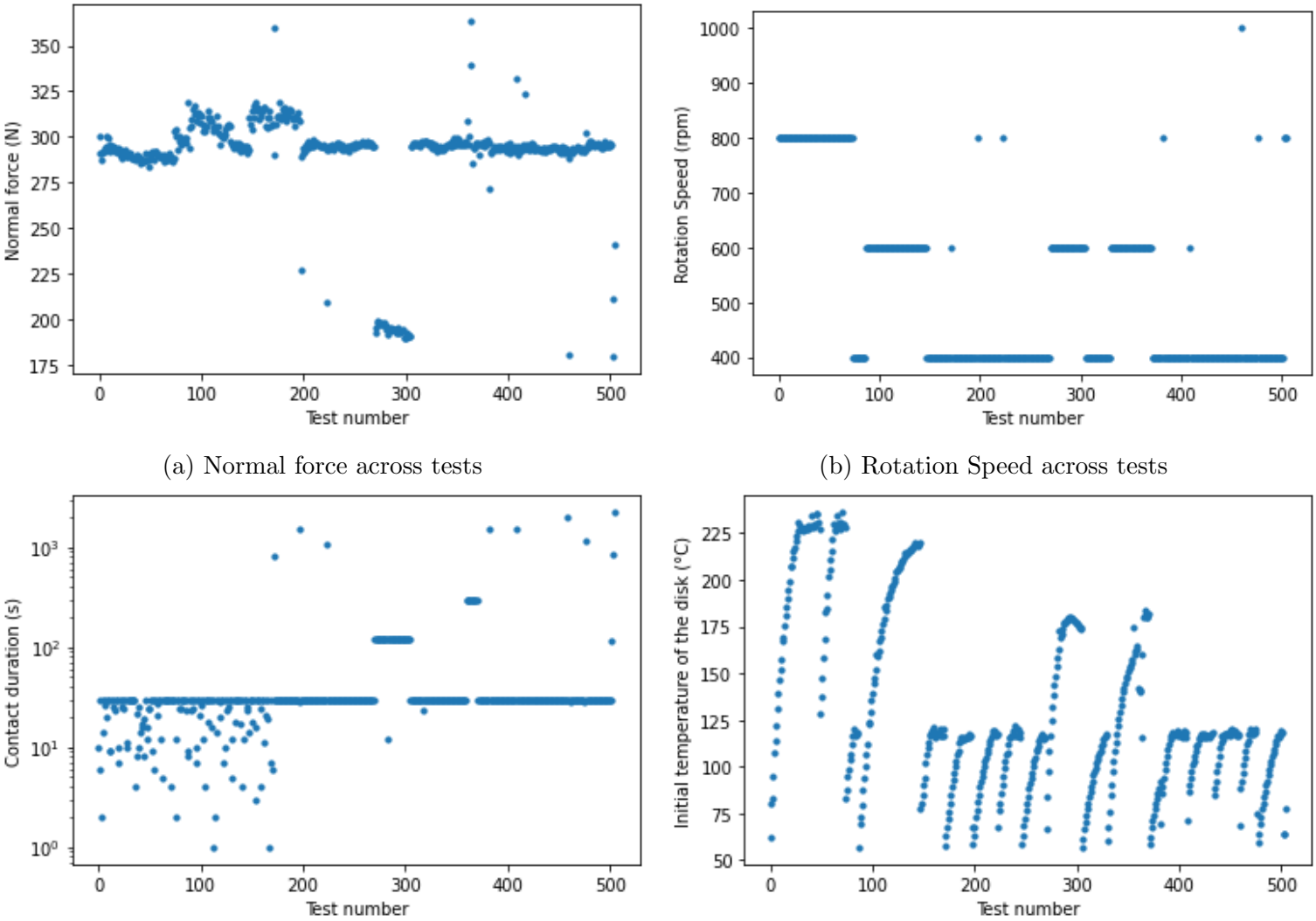
- An Optical Particle Sizer (OPS), covering a size range from 300 nm to 10 µm.

These devices enable real-time characterization of particle size distribution and concentration, providing key data for assessing the environmental impact of braking.

This comprehensive experimental platform thus delivers a rich, multi-dimensional dataset from each braking event. The collected signals form the basis for the machine learning models and physical insights presented in the subsequent sections of this work.

1.2 Experimental protocol

The database used in this work was generated by other members of the research group using the experimental setup detailed above. It comprises a series of braking test sets, each containing multiple individual tests. All tests were conducted using a NAO brake pad and a steel disc. Each test is defined by a quadruplet of input parameters: normal force, rotational speed, contact duration, and initial disc temperature. The testing protocol is illustrated in Figure 1.3.



(a) Normal force across tests

(b) Rotation Speed across tests

(c) Contact duration across tests (log-scaled)

(d) Initial temperature of the disk

Figure 1.3: Distribution of test parameters across the experimental database (a) Normal force, (b) rotational speed, (c) contact duration (log scale), and (d) initial disk temperature.

We observe that certain input parameters—such as contact duration—exhibit saturation, with the majority of values concentrated around a central range. This arises from the experimental protocol being purposefully designed to target specific phenomena known to occur under particular test conditions.

This focus leads to a sparsely explored parameter space, which may hinder the ability of downstream models to generalize effectively. An additional concern is that parameter saturation can introduce bias during model evaluation. For example, only 9 tests in the dataset exceed a contact duration of 4 minutes. If certain behaviors emerge only after this threshold, they will be captured by just those 9 tests. In a standard cross-validation setup with 6 splits, it becomes likely that some splits will entirely miss these rare but critical cases.

To address this, we implemented a cross-validation strategy based on the Wasserstein distance (see [65]), which promotes a more balanced representation of the parameter space across splits. The details of this approach are presented in Appendix A.2.

1.3 Post-Processed Variables

After data acquisition, the raw signals from the experimental setup were processed to extract a set of interpretable and physically meaningful variables. This post-processing involved unit normalization, and—in some cases—model-based estimations. The resulting variables provide a condensed yet comprehensive description of each braking test and serve as the foundation for the subsequent analyses and machine learning models developed in this work.

The final set of post-processed variables includes the following categories:

- **Acoustic signal:** raw sound pressure levels recorded by a high-sensitivity microphone positioned near the contact zone.
- **Mechanical signals:** force and motion measurements derived from various sensors, including:
  - Normal force (F_Z),
  - Tangential force (F_Y),
  - Rotation speed,
  - Torque,
  - Friction coefficient (μ),
  - Cumulative dissipated energy,
  - Tangential and radial angular displacements.
- **Thermal measurements:**
  - Surface temperature of the disc measured by a pyrometer,
  - Subsurface brake pad temperatures from embedded thermocouples, grouped into two depth zones: TC0–TC11 and TC12–TC15.
- **Particle emissions:**
  - EEPS (Engine Exhaust Particle Sizer) data across 32 particle size channels,
  - OPS (Optical Particle Sizer) data across 16 particle size channels.

1.4 Exploratory Data Analysis

The dataset generated by the experimental protocol is high-dimensional, multimodal, and exhibits nonlinear behavior. This section aims to highlight the complex and sometimes chaotic relationship between test parameters and resulting measurements. In addition, we perform standard exploratory data analysis (EDA) to assess variability, distributional properties, and interdependencies among the recorded variables.

1.4.1 Outcome variability under near-identical test parameters

Within the experimental campaign, several groups of tests were conducted under nearly identical parameters. Results from one of these groups are presented in Fig. 1.4 below:

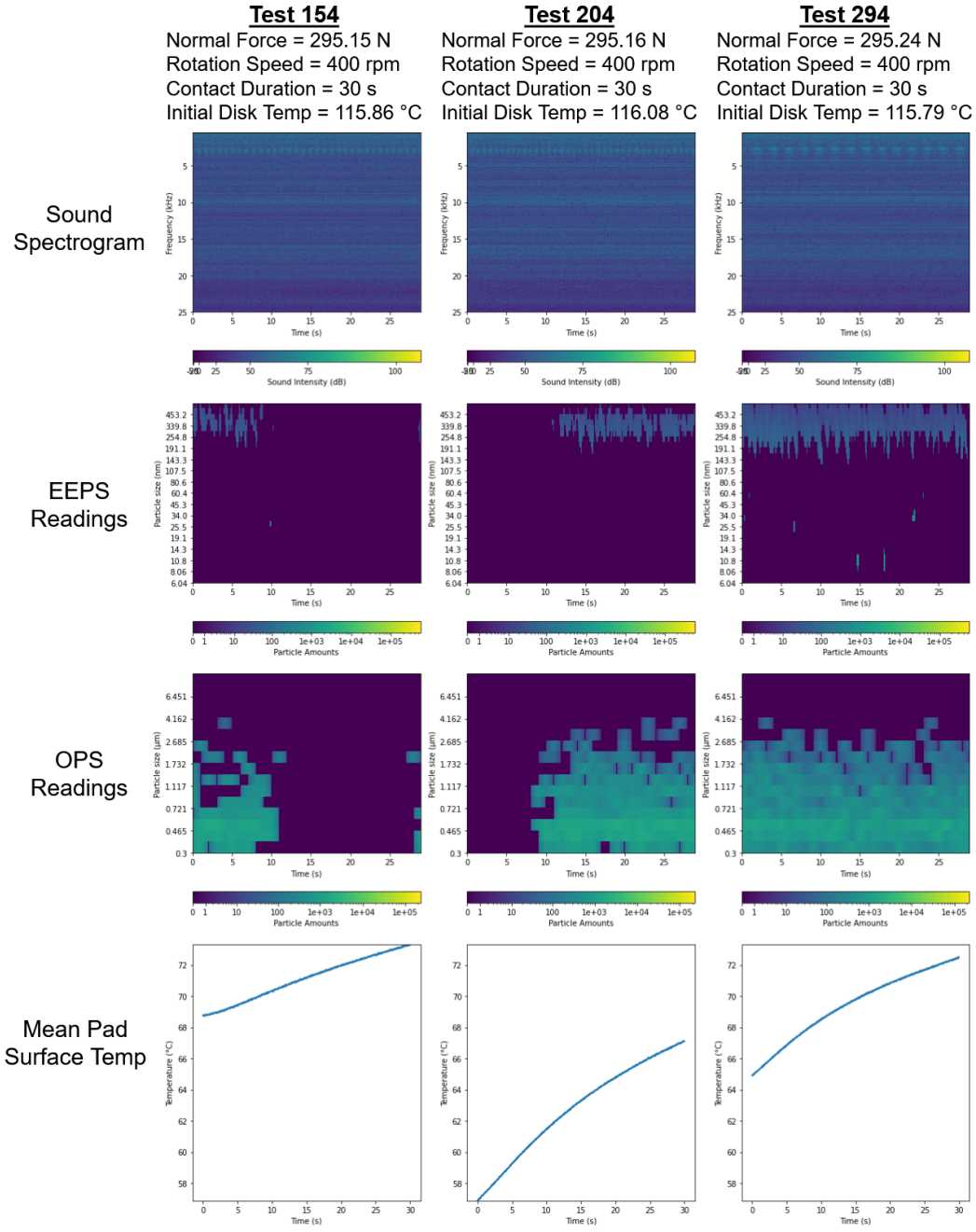


Figure 1.4: Results from three tests with nearly identical test parameters

Despite the similarity in test parameters, the resulting measurements vary significantly:

- **Acoustic response:** All tests exhibit a cyclic pattern, but the nature of these patterns differs. Also, certain frequency bands appear only in specific tests, such as around 14 kHz in Test 204 or 4 kHz in Test 294.
- **EEPS readings:** While particle sizes and amounts are broadly similar, the timing of emissions differs—test 154 exhibit emissions at the beginning, test 204 at the end, and test 294 throughout the duration.
- **OPS readings:** Similar observations apply as with EEPS—timing and consistency of emissions vary across tests.
- **Mean pad surface temperature:** Although all show an increasing trend, the rate of increase and final temperature differ.

This variability under near-identical conditions is well documented in the literature. It is often attributed to factors such as surface imperfections in pads and disks, third-body material dynamics (accumulation and ejection), and gradual fatigue or wear of the components.

These results highlight the complex and nonlinear nature of the system being studied. Crucially, they emphasize that the system's behavior cannot be fully predicted from test parameters alone. As such, this motivates the need for more advanced modeling approaches—such as the ones developed in this thesis—to account for hidden or emergent effects inherent in frictional contact systems.

To enable the development of such models, we now perform a global analysis of the processed variables.

1.4.2 Distribution Analysis

To better understand the overall behavior of the experimental system, we analyze the distribution of the measured variables. We begin by focusing on the mechanical-oriented variables, as shown in Figure 1.5.

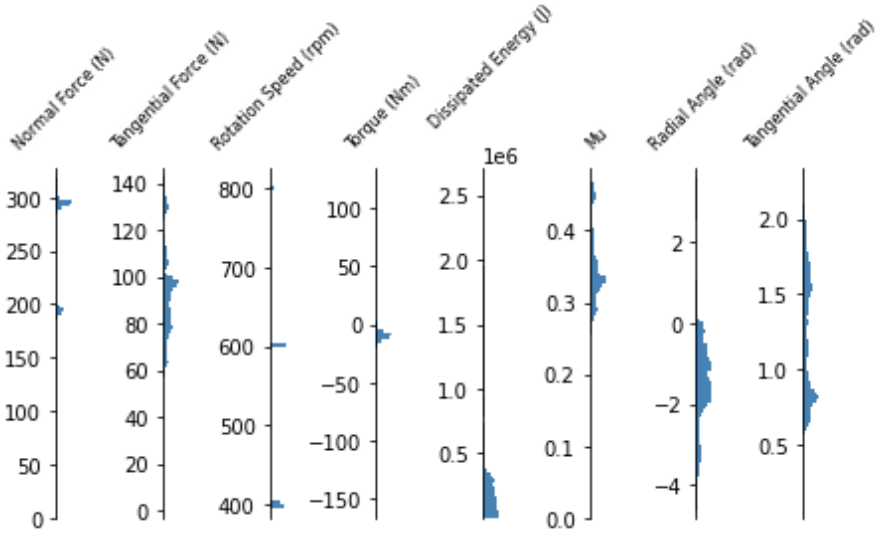


Figure 1.5: Mechanical-oriented variables histograms

It is interesting to observe that almost all mechanical variables exhibit multi-modal distributions. While this behavior is expected for variables that are directly tied to the test parameters (such as Normal Force and Rotation Speed) due to the experimental protocol, it is not necessarily expected for the other variables. The multi-modality in these variables is likely explained by the multi-modal nature of the distributions of the test parameters, which in turn leads to multi-modal outcomes. In other words, different test conditions produce distinct sets of system behaviors, which result in varying sets of measurements.

A similar analysis for the thermal variables is presented in Figure 1.6.

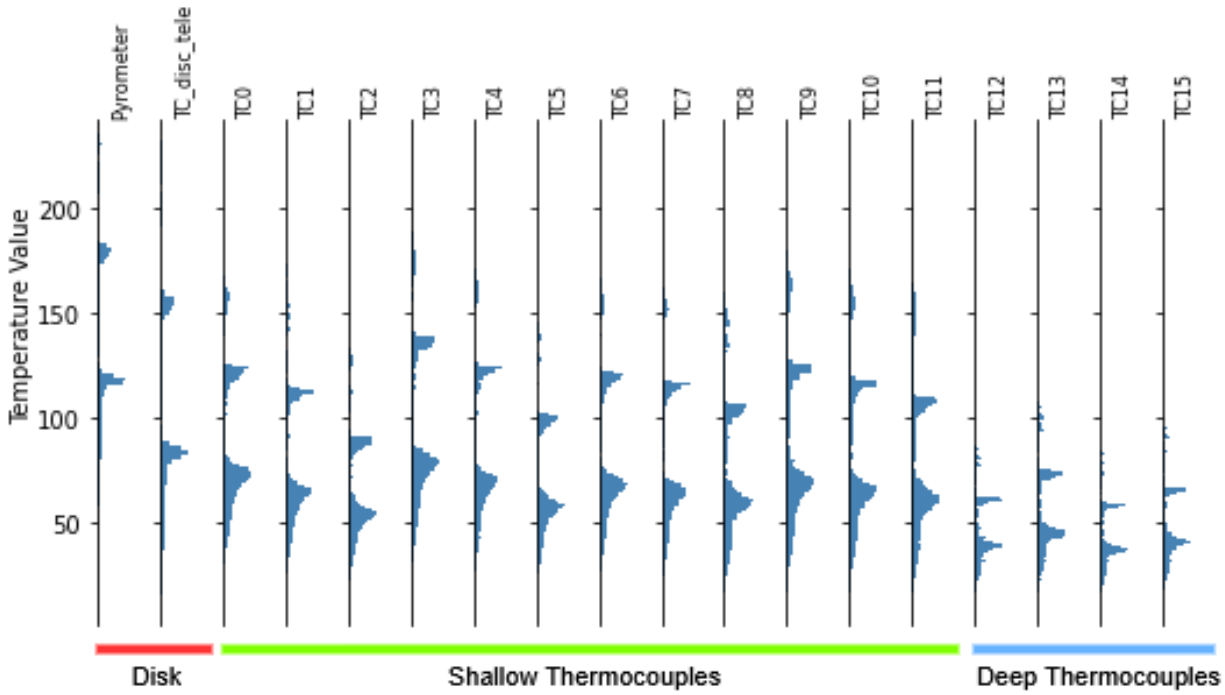


Figure 1.6: Thermal variables histograms

Similar to the mechanical variables, all thermal variables also exhibit multi-modal behavior. For disk-related variables, two modes are typically observed, while pin-related variables tend to show three distinct modes. Interestingly, we note that the deep thermocouples generally register lower temperatures than the shallow ones, which aligns with expectations, as they are further away from the contact zone. However, despite all shallow thermocouples being positioned nearly identically in relation to the contact, their histograms show significant variability. This suggests that there may be a tendency toward localized contact, which has been previously reported in the literature. This variability in thermal behavior further underscores the complexity of the system.

Finally, to complete the distribution analysis, we examined the sound and particle emission data through histogram heatmaps, as shown in Figure 1.7.–1.8.-1.9.

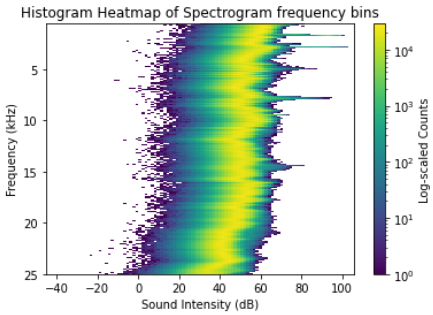


Figure 1.7: Sound Spectrogram per-frequency bin histogram heatmaps

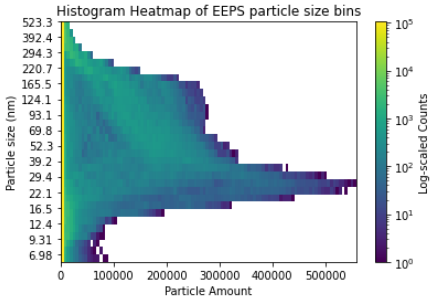


Figure 1.8: EEPS per-size bins histogram heatmaps

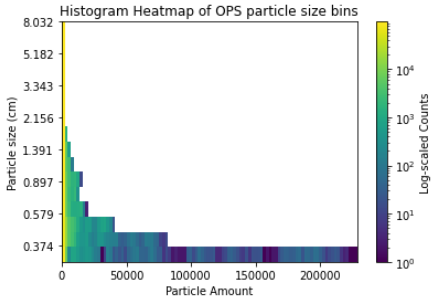


Figure 1.9: OPS per-size bins histogram heatmaps

We can make a few observations here:

- **Sound Spectrograms:** All frequency bins show a single mode pattern, but the mode tends to decrease with frequency. Few frequencies exhibit observations significantly higher than their modes, typically above 80 dB, likely indicating squeal events at specific frequency bins.
- **EEPS Emissions:** Particle size bins predominantly show a mode at zero, as the system does not emit particles most of the time. However, there are differences in the emission distributions across particle size bins, with some being emitted in far bigger quantities than others.
- **OPS Emissions:** Similar to EEPS emissions, all bins have a clear mode at zero, with large variations in the amounts emitted between particle sizes. However here, a clear trend emerges where the larger particles are emitted less frequently and in lesser quantities.

The preceding analysis provided valuable insight into the individual behavior of each measured variable. To gain a more comprehensive understanding of the system, it is now important to examine how these variables relate to one another. In the following section, we therefore investigate the correlations and redundancies among the collected variables.

1.4.3 Correlation and Redundancy Analysis through Hierarchical Clustering

This section focuses on computing correlations and studying redundancy to facilitate reliable explainability methods and inform data engineering in subsequent chapters. In this context, we limit our analysis to the Mechanical and Thermal variables, excluding the Sound and Particle Emissions variables, which will be treated as prediction targets rather than inputs in future analyses.

The dataset introduced here is not suitable for standard Pearson correlation, as it consists of time series. This time-series structure leads to multi-modal distributions, as previously observed, where all observations within a given test tend to cluster around the same mode. Consequently, computing Pearson correlation across the entire dataset could yield misleading results, as the experimental mean would be skewed. To address this, we compute the absolute Pearson correlation for each individual time series and then average the results. This

approach provides an estimate of the expected mean absolute correlation between variables in a typical test. By using the absolute value, we ensure that negative and positive correlations do not cancel each other out when calculating the mean. In this manner, we focus on quantifying the strength of the correlation, rather than its direction.

The resulting correlation matrix is shown below:

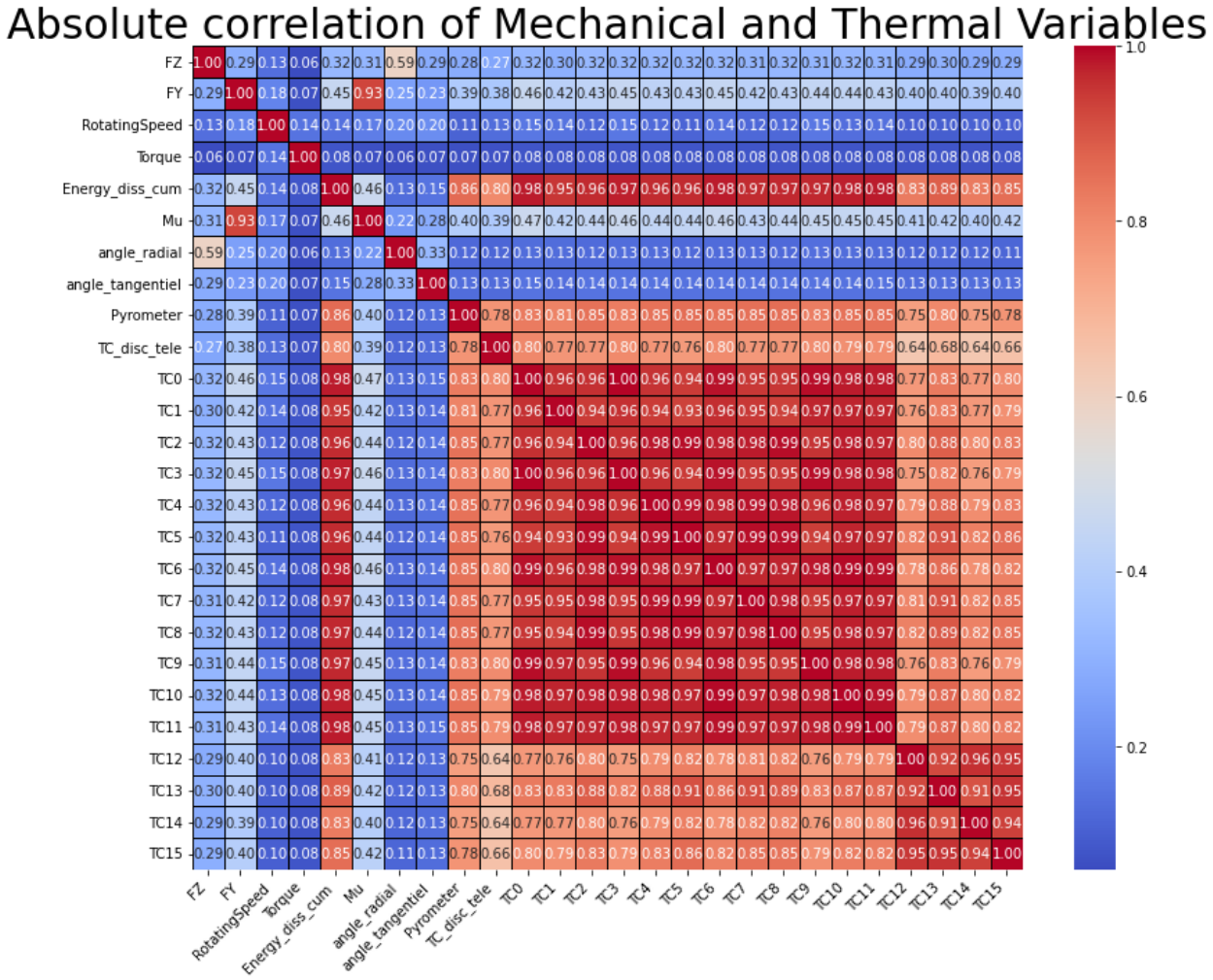


Figure 1.10: Pearson correlation strength between Mechanical and Thermal variables

Examining the correlation matrix, we observe numerous high correlations within the dataset. For instance, all temperature-related variables exhibit strong correlations with each-other, which aligns with our expectations. To enable the development of better models and more reliable explainability methods in future work, we aim to identify groups of variables such that no outer Pearson correlation strength exceeds 0.7, a threshold considered to indicate high correlation.

To achieve this, we computed the correlation distance from the matrix (i.e., $1 - |\text{correlation}|$) and applied Hierarchical Agglomerative Clustering (HAC) [27] with single linkage. We then cut the dendrogram at a height of 0.3, ensuring that no pair of variables with $|\text{correlation}| > 0.7$ are placed in different clusters. The resulting dendrogram is shown in Figure 1.11.

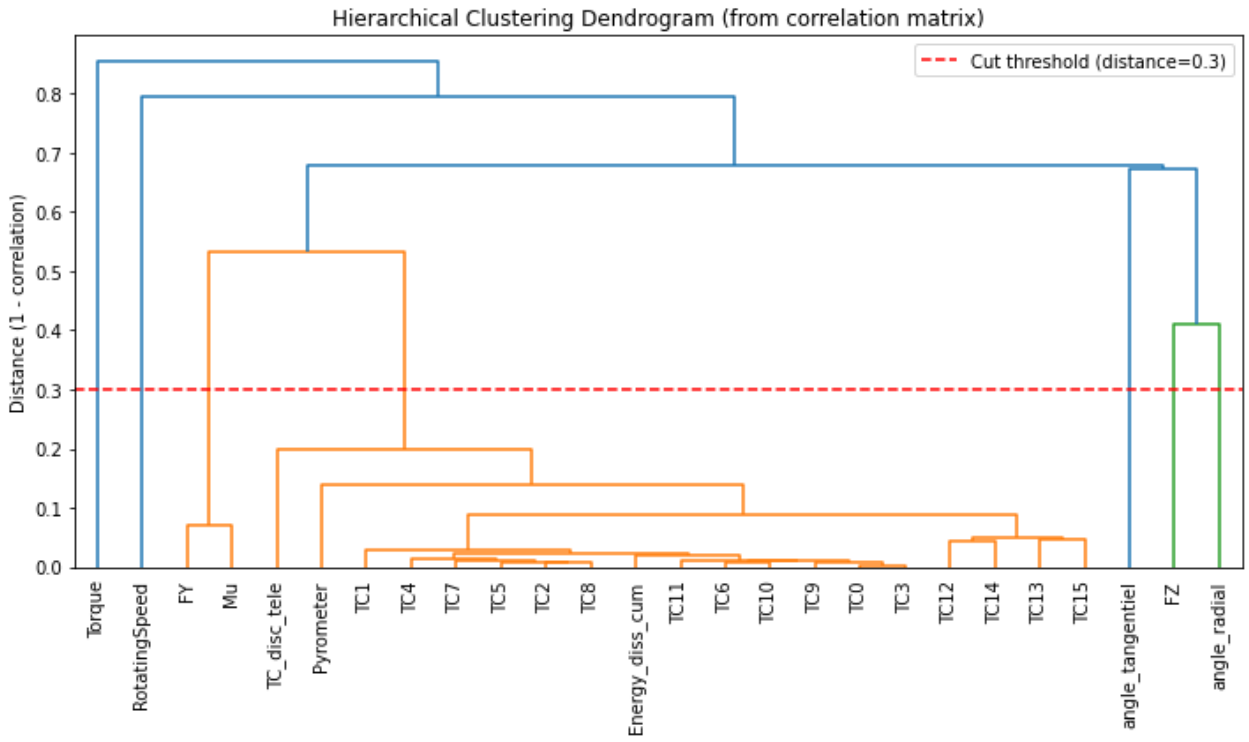


Figure 1.11: Hierarchical Clustering Dendrogram of Mechanical and Thermal Variables using single linkage and correlation distance

As expected, all temperature-related variables are grouped together. Some other interesting groupings emerge, such as the **Cumulative Dissipated Energy** being clustered with the temperature variables, or the **Friction Coefficient** and **Tangential Force**, which intuitively makes sense.

These groupings will be used in future chapters to guide feature selection and model development. They are valuable structural information that will help inform the development of more interpretable models and enhance the overall explainability of our work.

1.4.4 Summary of Exploratory Findings

In this section, we have explored the complexity and variability within the experimental dataset. First, we observed significant variability in the measurements, even under nearly identical test parameters, indicating that the system exhibits highly dynamic and nonlinear behavior. This variability underscores the need for sophisticated modeling techniques to account for these differences.

Secondly, we found that most of the recorded variables, both mechanical and thermal, exhibit multi-modal distributions. This is especially true for variables that are influenced by the test parameters. The multi-modality suggests that different operational regimes are present, and these distinct regimes contribute to the observed variations in the data.

Finally, we performed Hierarchical Agglomerative Clustering (HAC) to group variables based on their correlation strengths. This setup will guide feature selection and model development in the subsequent chapters, ensuring that we reduce redundancy and improve model interpretability.

1.5 Conclusion

This chapter has provided a comprehensive overview of the experimental setup and measurement protocol used to generate the dataset for this thesis. The tribological test bench, equipped with multi-modal instrumentation, allowed for the synchronized collection of mechanical, thermal, acoustic, and particle emission data, offering a rich, high-dimensional dataset that serves as the foundation for the analysis in subsequent chapters.

We also examined the variability within the data, observing significant fluctuations in measurements, even under nearly identical test conditions. This variability highlights the complex and nonlinear nature of the system and emphasizes the need for advanced modeling techniques that can account for these fluctuations and predict system behavior more effectively.

Through exploratory data analysis (EDA), we identified multi-modal distributions in both the mechanical and thermal variables. These findings suggest that the system operates in multiple distinct regimes, each contributing to the overall system behavior. Moreover, the Hierarchical Agglomerative Clustering (HAC) analysis revealed groups of variables that are highly correlated, providing valuable insights into the underlying relationships between the different measurement types. The HAC groupings will play a crucial role in future chapters, guiding feature selection, reducing redundancy, and improving model explainability.

Overall, the insights gained from this chapter not only enhance our understanding of the dataset but also lay the groundwork for the development of more sophisticated models and explainability methods. In the next chapters, we will build on these findings to develop predictive models that account for the complexity of the system and can effectively generalize across varying test conditions.

Chapter 2

Machine Learning-Based Prediction of Braking Emissions

This chapter explores the use of machine learning to predict pollution generated during braking, including both **particulate emissions** (measured via Engine Exhaust Particule Size (EEPS) and Optical Particule Size (OPS) instruments) and **acoustic emissions** (captured by a high-frequency microphone between 0 HZ to 25 kHz). These signals are complex, high-frequency, and often sparse — shaped by nonlinear interactions at the contact interface and difficult to capture with traditional analytical models. Moreover, beyond the modeling challenge, the physical system itself introduces intrinsic limitations: tribological contacts are notoriously non-reproducible due to multiscale surface inhomogeneities, third-body dynamics, and progressive wear, which means that two tests performed under nominally identical conditions may yield significantly different outcomes. This lack of reproducibility, combined with measurement uncertainties (sensor noise, calibration drift, environmental variability), further complicates both the interpretation of experimental data and the validation of predictive models.

We begin by investigating **time series forecasting** of acoustic emissions. This approach is motivated by earlier results on temperature signals from the same experimental setup, where forecasting models achieved promising accuracy. Here, the goal is to evaluate whether similar models can learn to anticipate sound emissions based on their past evolution and auxiliary signals like mechanical and Temperature related signals.

In the second part of the chapter, we shift focus to **real-time prediction**, where emission levels are estimated from physical inputs observed up to the prediction timestep. This setup is applied to all three types of pollution signals — EEPS, OPS, and Sound — and serves as the basis for testing different learning formulations:

- **Regression**, which predicts continuous values;
- **Classification**, which discretizes emissions into categories;
- **Distributional modeling**, which predicts a probability distribution (specifically GB2) to capture uncertainty in the emission levels.

Each approach is applied independently to each signal type. We also examine the relative contribution of different input variables using permutation-based feature importance.

The goal of this chapter is to quantify the achievable accuracy of instantaneous emission prediction from physical measurements.

2.1 Time Series Forecasting of Sound Emission

The objective of this section is to evaluate whether it is possible to forecast acoustic emissions from braking experiments using time series models. Forecasting refers to predicting future values of a signal based on its past and current observations. This direction was inspired by prior work conducted in the lab [44], where temperature signals were successfully forecasted with good precision. That study used both temperature data and the recorded normal force, collected from the same test bench employed in the present work.
More formally, we are trying to model a function f such that :

$$\hat{y}_t = f(y_{i\{i<t\}}, x_{i\{i<t\}}; \theta) + \epsilon$$

Where :

- $\hat{y}_t$ is our prediction of the sound value for a given timestep.
- $y_{i\{i<t\}}$ are past values of the sound signal.
- $x_{i\{i<t\}}$ are past values of other signals.
- θ are model parameters that are trained
- ϵ is an approximation error that we want to minimize.

2.1.1 Method

In this subsection, we outline the methods and techniques we employed to try to forecast sound pollution signals. Our approach consists of three main stages: data preparation, model architecture and training procedures.

Data Preparation : We first transformed the raw sound signals into spectrograms using the methodology described in A.4. Next, the auxiliary signals, if used, were interpolated to match the timestamps of the spectrograms. To prevent any chance of data leakage and maintain the forecasting nature of the method, we applied the LOCF interpolation method (A.5). Finally, the spectrograms were scaled using an AbsMax scaler fitted across all channels, while the auxiliary inputs were scaled using MinMax scalers (see A.6 for definitions). Although spectrogram values are natively non-negative when expressed as raw magnitudes, they can take on negative values once converted to the decibel scale. Their dynamic range can also vary strongly across frequency bins and test conditions, with occasional high-intensity peaks (e.g., squeal events). AbsMax scaling ensures that all spectrogram channels are normalized consistently with respect to their maximum observed magnitude, thereby preventing a few high-energy events from dominating the optimization. Conversely, the auxiliary mechanical and thermal variables are bounded within narrower physical ranges, making MinMax scaling more appropriate to preserve their relative scale and improve training stability.

Model Architectures : Given the temporal nature and complexity of the data, we naturally leaned toward neural network models for this task, althoug other methods were tested.
We explored two primary types of prediction methods: single time-step prediction and sequence-to-sequence (seq2seq) prediction. In the single time-step prediction approach, the model forecasts the value of the signal at the next timestep based on the previous ones. In

contrast, the seq2seq methods aim to predict multiple future values at once. See A.7 for more details.
For both single time-step and sequence-to-sequence prediction, we experimented with two input configurations: providing a fixed time window of past values, and supplying the full input history from the start of the test.
At the layer level, we experimented with several types of encoders (/decoders for seq2seq models) to capture the temporal dynamics in the data. Specifically, we tested architectures based on Long Short-Term Memory (LSTM)[24], Gated Recurrent Units (GRU)[12], Temporal Convolutional Networks (TCN)[5] and Transformer[64] layers.

Training Procedure : To train our models, we used the Adam[32] optimizer. This choice is particularly suitable for our setup, since the dataset is of medium size, the input signals are heterogeneous, and the loss landscape is potentially noisy and non-convex. Compared to simpler optimizers such as vanilla SGD, Adam generally achieves faster convergence and greater robustness to hyperparameter settings, which makes it a standard and reliable choice in time series deep learning applications. To prevent overfitting and improve generalization, we employed weight decay, dropout, and jittering of the inputs. In terms of loss functions, we experimented with various options including Mean Squared Error (MSE), Mean Absolute Error (MAE), Huber loss, and quantile loss. Regardless of the loss function used during training, we always evaluated model performance using MSE to maintain consistency in assessing prediction accuracy.
Finally, to achieve realistic test evaluation, we used a 6-fold cross-validation split detailed in A.3.1 using the method described in A.2. For each split, we used it once as the test split, selected one as the validation split, and used the remaining splits for training.

2.1.2 Results

Despite extensive hyperparameter tuning and architecture experimentation, most models collapsed to constant predictions, indicating a failure to extract meaningful predictive patterns. This behavior persisted even when auxiliary signals — such as mechanical and temperature-related inputs — were included.

We also experimented with forecasting a thresholded version of the sound spectrograms — intended to highlight high-intensity events and suppress background information — but this did not lead to improved results. Models still failed to learn meaningful temporal dynamics, even with this simplified target.

One partial exception was observed in the sequence-to-sequence setup. In this framework, auxiliary signals can be incorporated in two main ways: either by asking the model to predict them jointly with the sound signal, or by providing their true values jointly with the forecasting process. We tested the latter — supplying the model with the true normal force (FZ) — which led to noticeably better performance, particularly during high-intensity events.

While this initially appeared promising, closer inspection revealed a strong overlap between high-energy patterns in the Normal Force and Sound spectrograms as illustrated in Figure2.1:

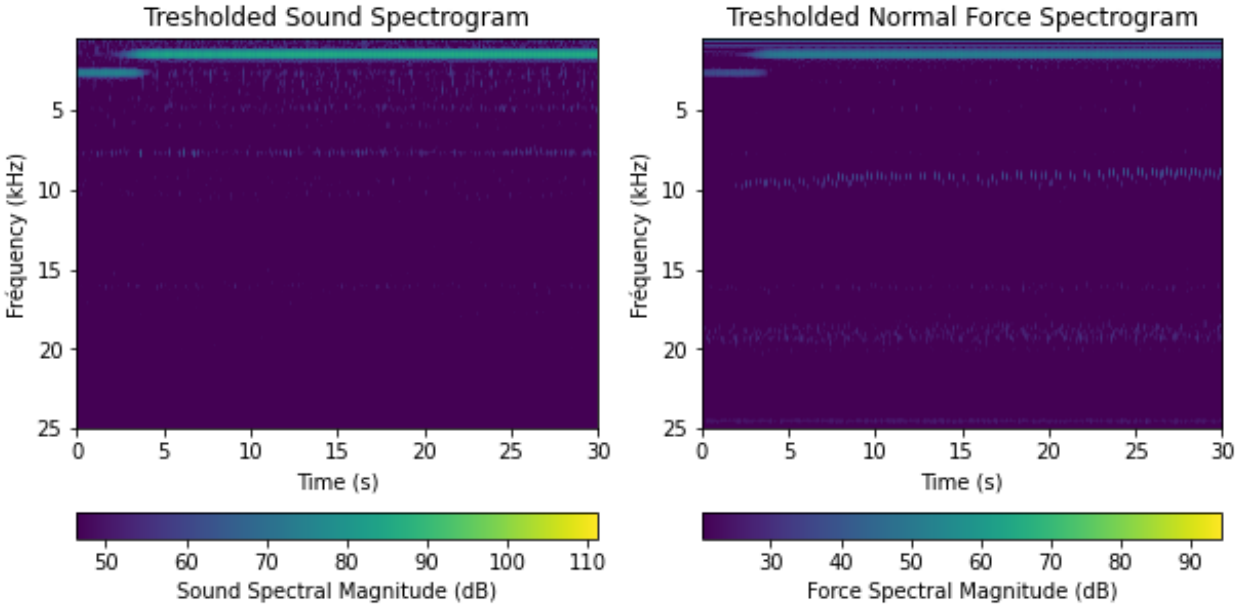


Figure 2.1: High-intensity behaviors in example sound and Normal Force spectrograms. Both spectrograms are converted to dB and thresholded using the 0.9 quantile of the distribution over the entire dataset.

This similarity is expected, as the physical mechanisms responsible for strong acoustic emissions also manifest in the force signals at similar frequencies. However, it suggests that the model was not truly *forecasting* the sound signal, but rather relying on access to the true normal force — which undermines the nature of the forecasting task.

In conclusion, while previous work on temperature prediction in tribological systems [44] achieved satisfactory forecasting results, forecasting of sound data did not succeed. While the inclusion of auxiliary variables such as the Normal Force improved apparent performance, this improvement did not reflect genuine forecasting capability. The model effectively performed a regression from the true Normal Force to the corresponding acoustic response, rather than learning to predict future sound behavior from past observations. As a result, we conclude that this forecasting setup does not provide meaningful predictions. Consequently, in the next section we shift our focus toward a regression formulation, where the goal is to estimate the emission level directly from concurrent mechanical and thermal inputs.

2.2 Real Time Pollution Prediction

In light of the challenges encountered in the forecasting study, this section focuses on **real-time pollution prediction**, where pollution encompasses both noise and particle emissions. Instead of a forecasting framework, we adopt a **time-series regression** approach, in which the aim is to estimate target variables from the past and current values of other measured signals.
Formally, we are trying to model a function f such that :

$$\hat{y}_t = f(x_{i\{i\leq t\}}; \theta) + \epsilon$$

Where :

- $\hat{y}_t$ is our prediction of the predicted variable for a given timestep.

- $x_{i\{i \leq t\}}$ are past and present values of the input variables.
- θ are model parameters that are trained
- ϵ is an approximation error that we want to minimize.

This approach differs significantly from the previous section's, as it does not take the predicted variable as input, and the model now sees the input variables up to the prediction timestep.

2.2.1 Single-Target Pollution Prediction

In this section, we focus on modeling each pollution signal, EEPS particles, OPS particles and Sound, independently, using only the available mechanical and thermal input variables. This approach provides a controlled framework for evaluating the performance of various modeling strategies on each emission type in isolation. It also allows us to identify the specific challenges and characteristic behaviors associated with predicting each signal individually.

2.2.1.1 Method

Our approach consists of five main stages: data preparation, target definitions, model architecture, training procedures and prediction method.

Data Preparation : We began by transforming the raw sound signals into spectrograms, following the methodology outlined in A.4. Subsequently, we interpolated all signals to uniform 0.1-second spaced timesteps. For the interpolation process, we employed two distinct methods: for the input variables, we used Last Observation Carried Forward (LOCF) as detailed in A.5, while for the output variables, we applied linear interpolation. This approach ensures there is no data leakage in the input, while also providing appropriate targets for our model's learning process. Finally, all input features were scaled using MinMax scalers, and all output variables were scaled using MaxAbs scalers over all channels at once (see A.6 for definitions).

Target definitions. EEPS and OPS targets are *per-size-bin* instantaneous concentrations at time t (32 and 16 channels, respectively; native sampling rates 10 Hz and 1 Hz).Sound targets are *per-frequency-bin* spectrogram magnitudes (dB) at time t with STFT parameters defined in A.4. Reported errors are mapped back to physical units for tables and plots.

Model Architecture : While various architectures could have been explored, we opted for a single, simple Transformer encoder model. Transformer models have been shown to perform strongly across time-series tasks, due to their ability to capture long-range dependencies and flexible feature interactions (see [66]). The architecture is illustrated in Figure 2.2.

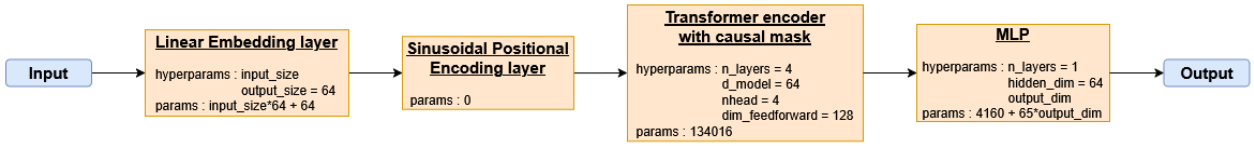


Figure 2.2: Transformer encoder model for time series regression

This architecture aligns well with our requirements, as it enables predictions at each time step using both past and present input values, thanks to the use of a causal mask within the transformer encoder.

For the applications presented in this section, the model has between 140,944 and 400,570 parameters. While this is quite low by modern deep learning standards, our testing has shown that larger models tend to overfit, likely due to the relatively limited size and diversity of our dataset. This aligns with known challenges when applying deep learning to small datasets.

Alternative transformer-based models—such as patch transformers—were also evaluated. However, they did not lead to improvements in our evaluation metrics. Therefore, we chose to retain this simpler, more efficient model.

Finally, while encoder-decoder models could potentially eliminate the need for interpolation, reducing approximation errors, their significantly larger parameter sizes would likely lead to overfitting, especially given the modest size of our dataset. For this reason, we prioritized model simplicity and robustness by using only an encoder.

Training Procedures : To train our models, we used the Adam [32] optimizer. To prevent overfitting and improve generalization, we employed weight decay, dropout, and jittering of the inputs.

To handle the varying lengths of the time series, a bucketing strategy was used to group sequences of similar lengths into the same batch. This approach reduces padding and improves training efficiency.

Due to the relatively small number of tests, we observed high variability in training performance, primarily caused by random model weight initialization. To mitigate this, we implemented a restart strategy: if the validation performance did not surpass that of a simple "dummy model" baseline—which always predicts the median—after a predefined number of epochs, the training was reset (with a maximum number of resets). This approach helped us escape poor initializations where effective learning failed to occur.

To ensure that each test contributed equally during training—and to prevent longer time series from disproportionately influencing the loss—we first computed the mean loss along the time axis, followed by averaging across all tests.

Finally, to achieve realistic test evaluation, we used one of the 6-fold cross-validation detailed in (A.3.1, A.3.2, A.3.3) depending on the objective. The cross-validation splits were determined using the method described in A.2. For each split, we used it once as the test split, selected one as the validation split, and used the remaining splits for training. Models were trained for 4000 epochs and we selected the best model by looking at the best validation performance.

Prediction Methods: Finally, regarding prediction methods, we implemented three different approaches:

- First, we implemented a standard regression model, where the output is a direct prediction of the target value. This approach is optimized using Huber loss.

- Second, we reformulated the problem as a classification task. Instead of predicting precise values, we discretized the target variable into intervals and trained the model to predict the corresponding interval for each observation. This approach is optimized using cross-entropy loss. For details on how the classes were created, see A.8. It should be noted that these methods were applied to the entire dataset, rather than just the training splits. While this inevitably leads to some data leakage, it allows us to compare models trained on different splits, as they then share the same classes.
- Third, we performed Distributional Regression by fitting the observations to a GB2 distribution, using negative log-likelihood (NLL) as the training loss. For the rationale behind selecting the GB2 distribution and further technical details, see A.9.

2.2.1.2 Results

This section presents the results based on different combinations of input and output variables, as well as the prediction method used. The mechanical and thermal related signals are consistently used as inputs when predicting EEPS, OPS, and Sound. In some cases, sound is also included as an additional input when predicting particle emissions.

We begin by examining the general results:

	Standard Regression (MAE)	Emission level classification (Accuracy)	GB2 distributional regression (MAE)
M&T =>EEPS	3620.6 ± 13352.3	0.746	2626 ± 13646.9
M&T&S =>EEPS	3512.1 ± 12980.9	0.75	2437.8 ± 13782.2
M&T =>OPS	979.7 ± 5857.2	0.675	375.6 ± 2161.9
M&T&S =>OPS	979.6 ± 5857.2	0.662	422.1 ± 3461.1
M&T =>Sound	9.3 ± 38.9	0.658	3.4 ± 14.3

Table 2.1: Single Target pollution prediction global evaluation metrics. Mean and Std for MAE evaluated methods and accuracy for classification based methods. M stands for Mechanical-related variables, T stands for Temperature-related variables and S stands for Sound.

Training and evaluation of the models required approximately 70 hours in total using a 3-way MIG setup. The permutation feature importance (PFI) analysis discussed later added an additional 14 hours of computation. See Appendix A.1 for details on the software, hardware, and computational resources used.

From the results, we observe that including sound as an additional input improves the prediction of EEPS across all prediction methods. However, for OPS, adding sound does not lead to any significant improvement.

Additionally, the Standard Regression model is consistently outperformed by the GB2 Distributional Regression across all predictions. This suggests that the distributional approach is better suited for this task.

It is important to note that comparing Classification-based methods with GB2 Distributional Regression directly is not feasible from this table alone, as they are evaluated using different metrics. To allow for a more precise comparison and to highlight specific challenges faced by each method, we now present results per objective.

2.2.1.2.1 EEPS When modeling EEPS particles, an analysis of the previous table reveals that incorporating sound as an input consistently improves performance across all metrics. Although the improvements are relatively small and could partly be attributed to training variability, the persistence of the trend is notable.

A similar connection has been highlighted in the literature: peaks in fine particle emissions (here measured with EEPS) often coincide with squeal phases or strong acoustic vibrations during braking. In [19], operational parameters such as braking pressure, speed, and temperature—conditions amplified during abrupt braking—are reported to influence not only nanometric particle emissions but also the likelihood of vibratory instabilities (squeal). Although a direct causal link is not always established, this suggests that in our dataset, the observed benefit of including sound as an input likely stems from its ability to capture information specific to abrupt braking events, where both particle and acoustic emissions are simultaneously elevated.

For these reasons, and based on both empirical trends and physical plausibility, we include sound as an input feature in the remainder of this study.

Evaluation plots

We begin by examining the evaluation plots corresponding to each modeling setup. For the regression approaches, the predicted-versus-true and residual-versus-true relationships are presented in Figures 2.3 and 2.4.

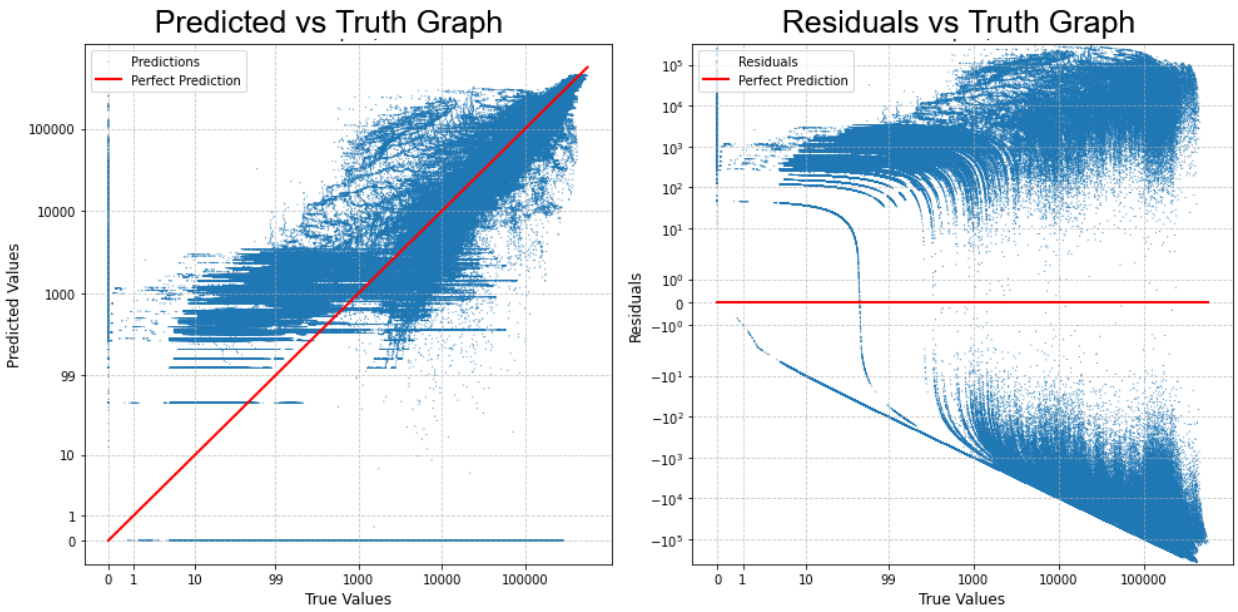


Figure 2.3: EEPS Standard Regression evaluation graphs

The prediction-versus-truth plots show that both models perform comparably for high emission values. However, the GB2 distributional approach clearly outperforms the standard regression model for medium to low emission levels. This behavior is consistent with the known tendency of conventional regression losses to prioritize large targets, since the gradients associated with smaller values are inherently weaker. Although target scaling (e.g., log or log–log transformations) is often used to mitigate this issue, none of the strategies tested here improved convergence. Combined with its lower mean absolute error (MAE), these re-

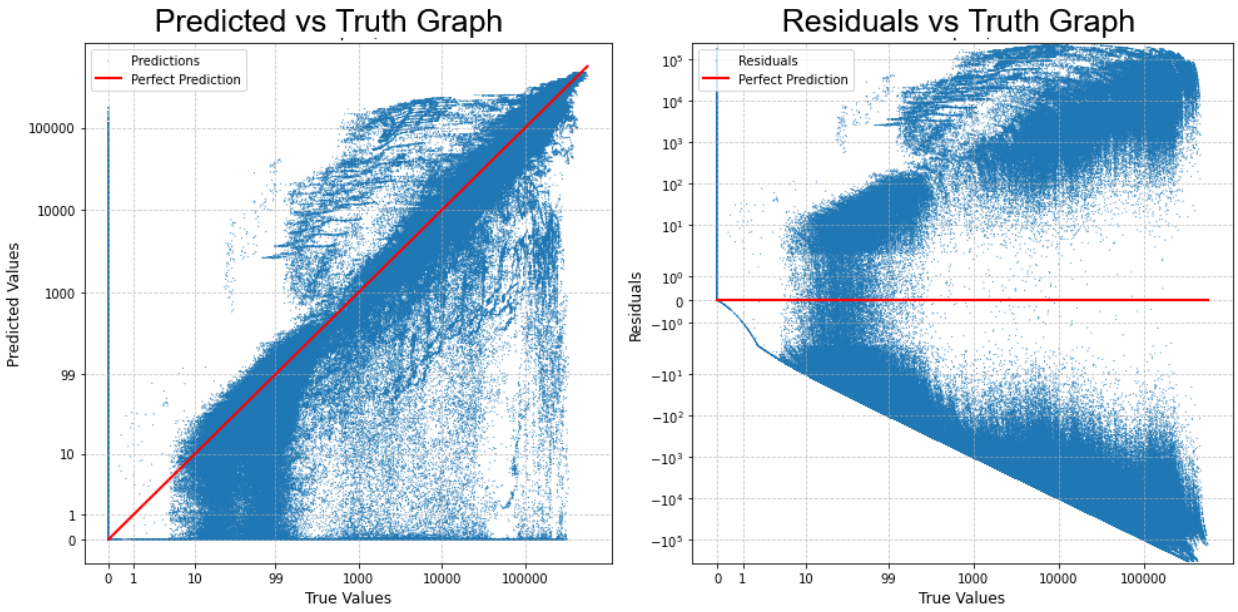


Figure 2.4: EEPS GB2 distributional Regression evaluation graphs

sults make GB2 Distributional Regression the preferred method for this task. Consequently, we focus on this approach in the remainder of the study.

We now turn to the Emission Level Classification model evaluation. The corresponding confusion matrix is presented in Figure 2.5.

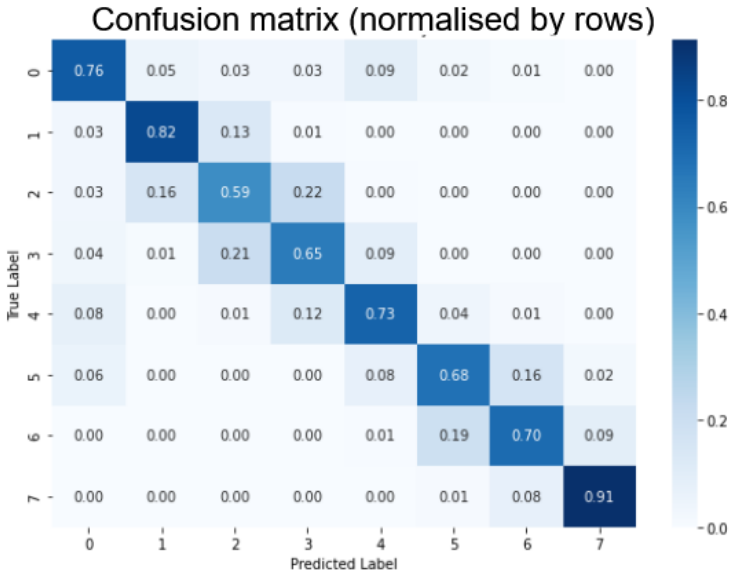


Figure 2.5: EEPS Emission level classification evaluation graphs

The resulting confusion matrix exhibits a predominantly tridiagonal structure, with most predictions falling within one class of the true label. The most frequent error occurs for class 0, which the model occasionally misclassifies—either by falsely predicting zero emissions or by failing to identify them. This behavior is expected, as class 0 represents the absence of emissions and thus corresponds to the underlying "emission or no-emission" subproblem, which is inherently more ambiguous.

For the remainder of this chapter, we do not report additional classical metrics such as ROC or precision–recall curves. These metrics rely on precision and recall scores, which are not particularly representative of model performance when dealing with **ordinal** data. Although some extensions of these metrics have been proposed for ordinal classification, they typically require defining a custom scoring function—something difficult to justify in our setup, where class 0 plays a qualitatively distinct role from the other classes.

Qualitative comparison

Comparing the Emission Level Classification with the GB2 Distributional Regression is challenging from a quantitative standpoint. One could, for instance, approximate the classification model's MAE by assigning predictions to interval midpoints, or conversely, assess the GB2 model's accuracy by measuring how often its predictions fall within the classification boundaries. However, neither approach is strictly consistent with the models' respective training objectives. Consequently, we adopt a qualitative comparison, presenting predictions from each model for a set of representativ test cases.

An analysis of the test results reveals four distinct categories of EEPS emissions, each predicted with varying degrees of accuracy by the different methods. Below, we present examples of both successful and unsuccessful predictions for each category.

First category: Little to No Emission

This category is characterized by the absence of detectable emissions or by very small, localized emission events. Figure 2.6 shows two representative tests and the corresponding model predictions:

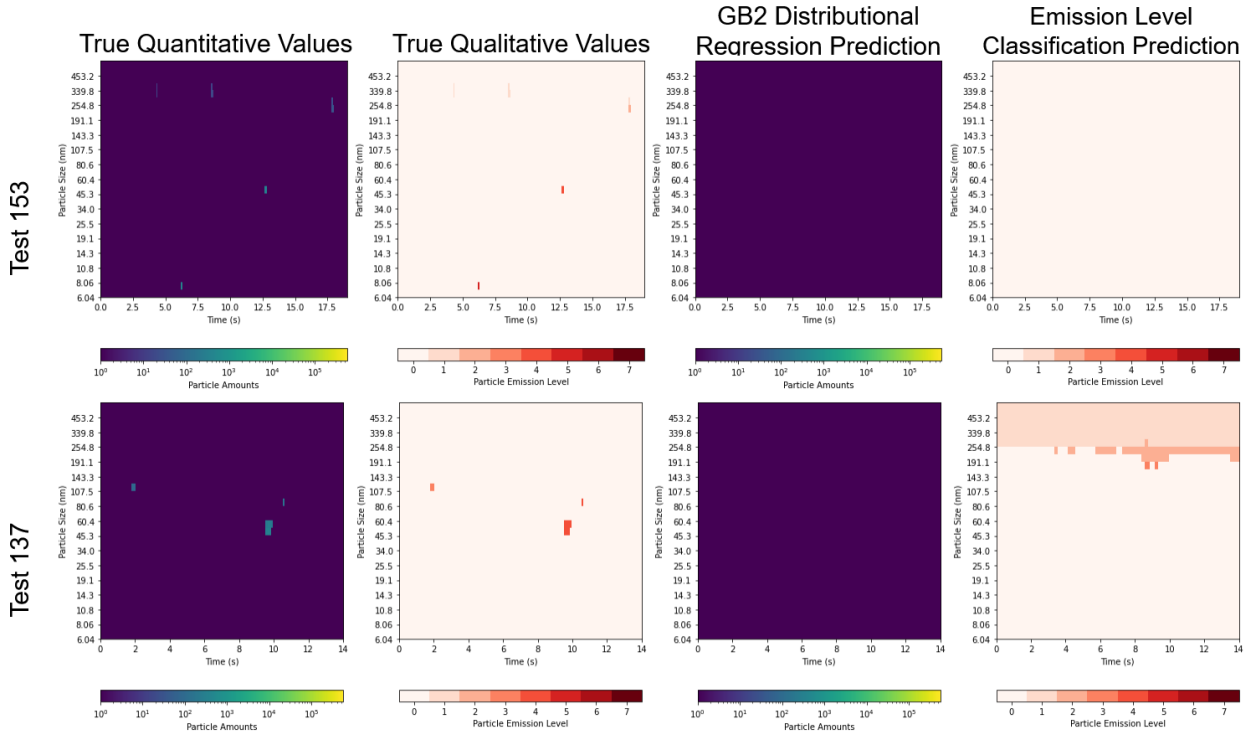


Figure 2.6: Examples of EEPS prediction in the "Little to No Emission" category.

Most examples of this categories are predicted like test number 153 : Emission-level classification and GB2 distributional regression models almost always predict zero emissions throughout the test, suggesting they handle this category well, even though they typically ignore the small, localized emissions sometimes present in the data. However, as seen in test number 137, these models can occasionally "hallucinate" emissions—especially the classification models—by predicting positive emissions where none actually occurred.

Second category: Sparse, Low-Intensity Emissions Across All Sizes

This emission category is characterized by low-intensity emissions that are sparse both in particle size and over time. Figure 2.7 shows two representative tests and the corresponding model predictions:

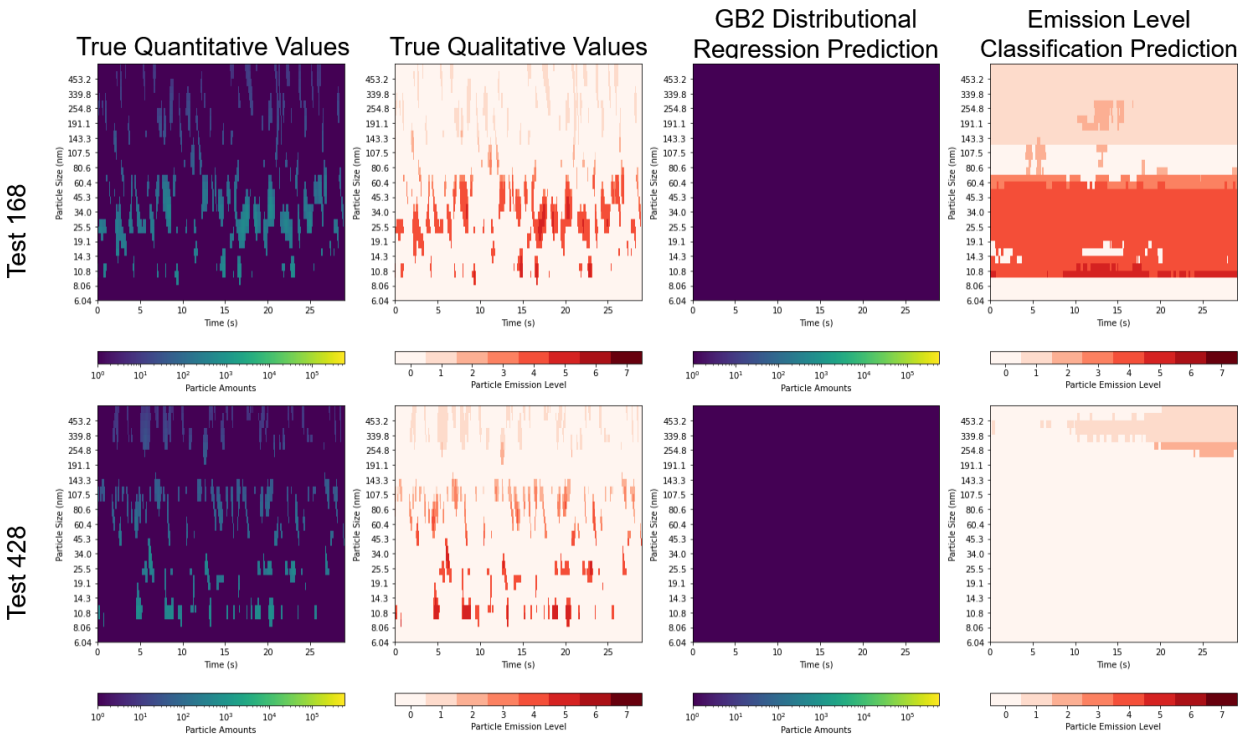


Figure 2.7: Examples of EEPS prediction in the "Sparse, Low-Intensity Emissions Across All Sizes" category.

Most tests in this category are predicted similarly to test number 168: GB2 distributional regression typically outputs a constant prediction equal to zero, indicating a lack of fine-grained pattern recognition. The emission-level classification model, on the other hand, appears to capture the overall emission pattern, including typical intensity and particle size range. However, it fails to capture the temporal and size sparsity, typically predicting dense emission bands at the correct sizes, despite the true emission being more sporadic.
While this behavior is consistent in most examples, the classification model sometimes still fails to detect this emission type entirely, leading to zero predictions, as illustrated in test 428.

Third category: Medium-Intensity, Dense Emissions Concentrated in Large Particle Sizes

This category typically shows dense medium particle emission, concentrated in the larger particle sizes. Figure 2.8 shows two representative tests and the corresponding model predictions:

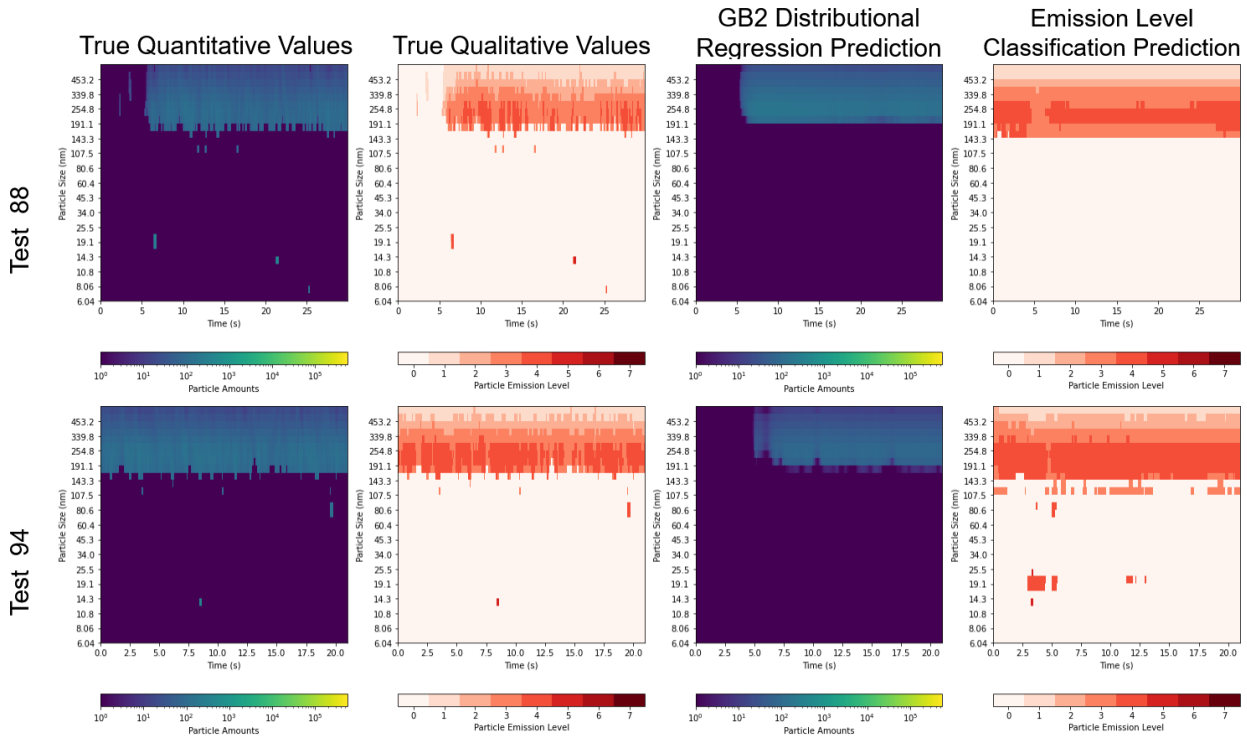
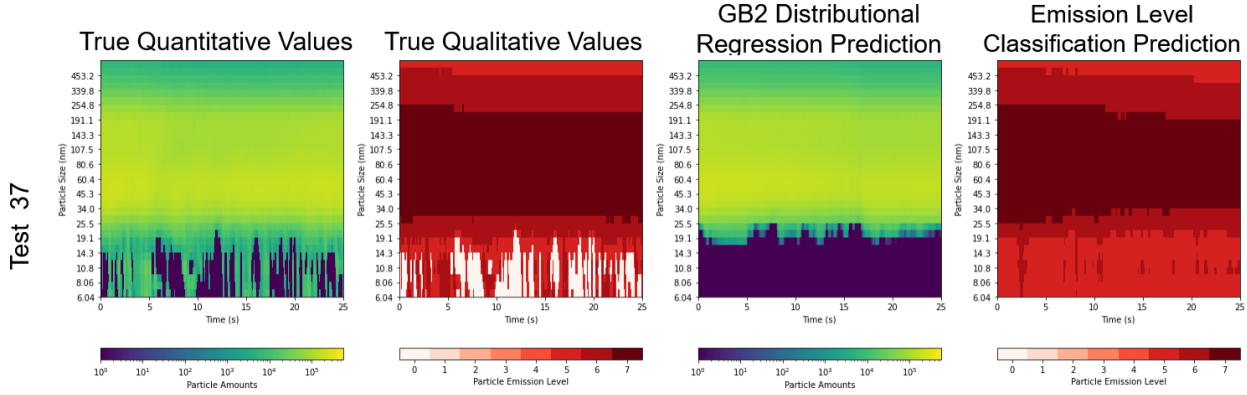


Figure 2.8: Examples of EEPS prediction in the "Medium-Intensity, Dense Emissions Concentrated in Large Particle Sizes" category.

Unlike earlier cases, we have not observed major prediction failures from the Emission-Level Classification and GB2 Distributional Regression models. While this emission category is less frequent than others, we have seen that in all examples, both models successfully detect the emission size and intensity, predicting dense bands that approximate well the true emissions. The main limitation, as illustrated in the two plots, is that they sometimes fail to capture the correct temporality of the emission bands—that is, the precise onset and duration of the emissions.

Fourth category: High-Intensity, Dense Emissions Spanning Most Particle Sizes

This category is characterized by dense and elevated emission levels spanning nearly all particle sizes. Figure 2.9 shows three representative tests and the corresponding model predictions:



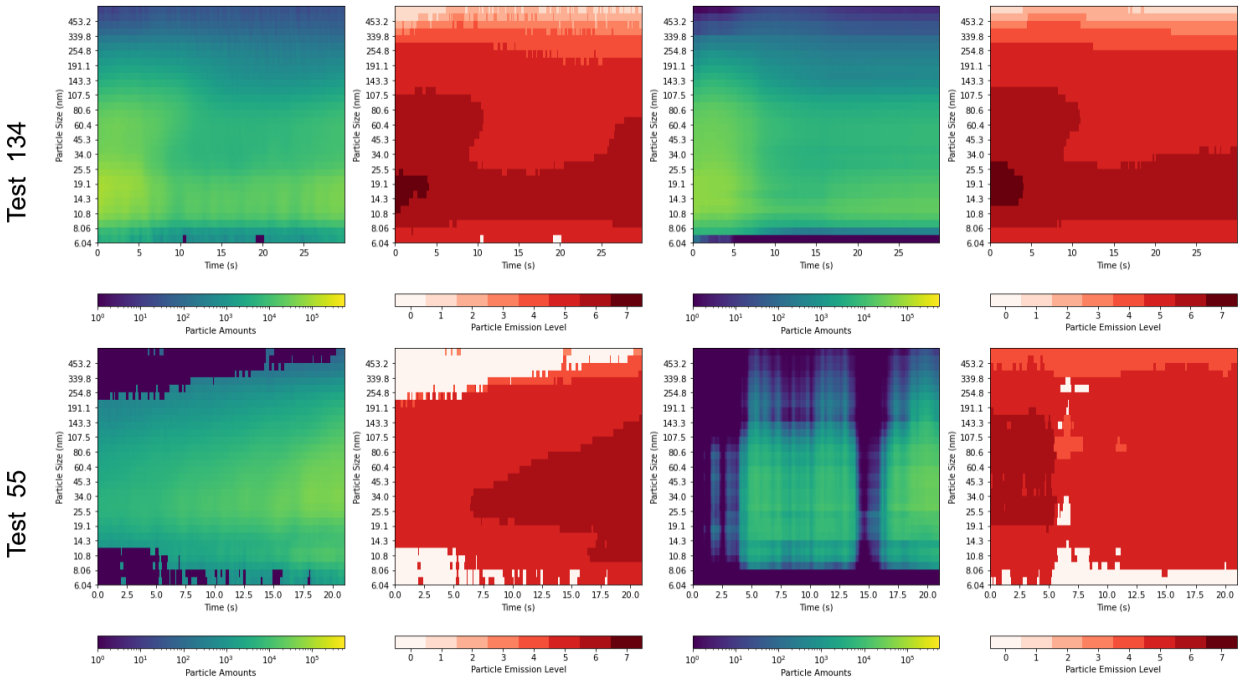


Figure 2.9: Examples of EEPS prediction in the "High-Intensity, Dense Emissions Spanning Most Particle Sizes" category.

The Emission Level Classification and GB2 Distributional Regression models both generally perform well in this category. This holds true not only for "simple" emission patterns, as seen in test 37, but also for more complex ones, such as ones seen in test 134. Still, prediction errors occur, as illustrated in in test 55.

Note: Multi-Emission Category Tests

Some tests, typically longer in duration, have been observed to produce particle emissions spanning multiple categories simultaneously .Figure 2.10 shows an example of such a case:

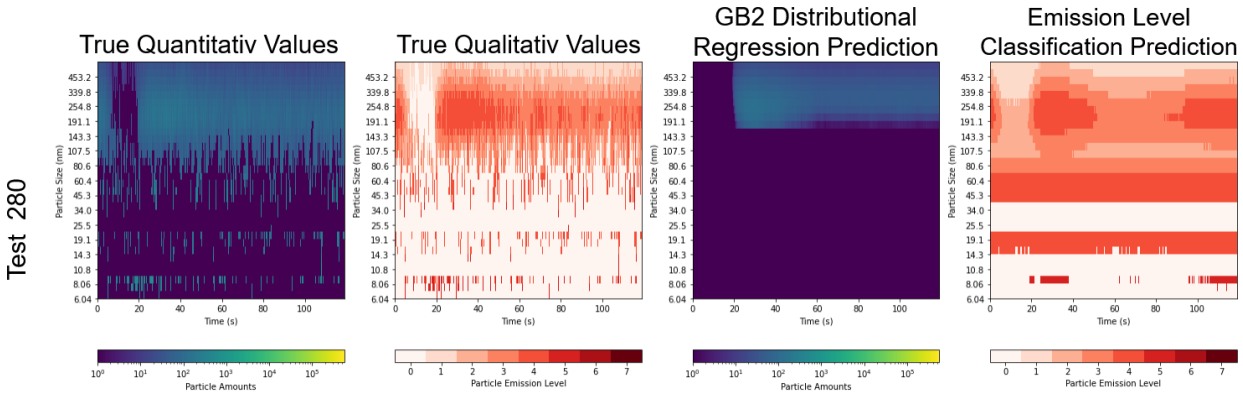


Figure 2.10: Example of EEPS prediction showing multiple emission categories simultaneously

In such cases, as seen in the example, the previous observations still apply: we see a dense emission band concentrated in the larger particle size range, accompanied by sparse, low-intensity emissions across the entire size spectrum. Consistent with our earlier findings,

both the emission-level classification and GB2 distributional regression models successfully predict the medium-density emission band, while only the classification model detects the sparse, low-intensity emissions.

Feature Importances

As a final note on the EEPS modeling results, we assessed feature importance using Permutation Feature Importance (PFI) (see A.10) computed on the MAE score for the GB2 distribution regression and on the accuracy for the Emission Level Classification. To ensure the reliability of the results, input variables were grouped as defined in Section 1.4.3, reducing the impact of correlation on the PFI computation. The PFI are shown in Figure 2.11.

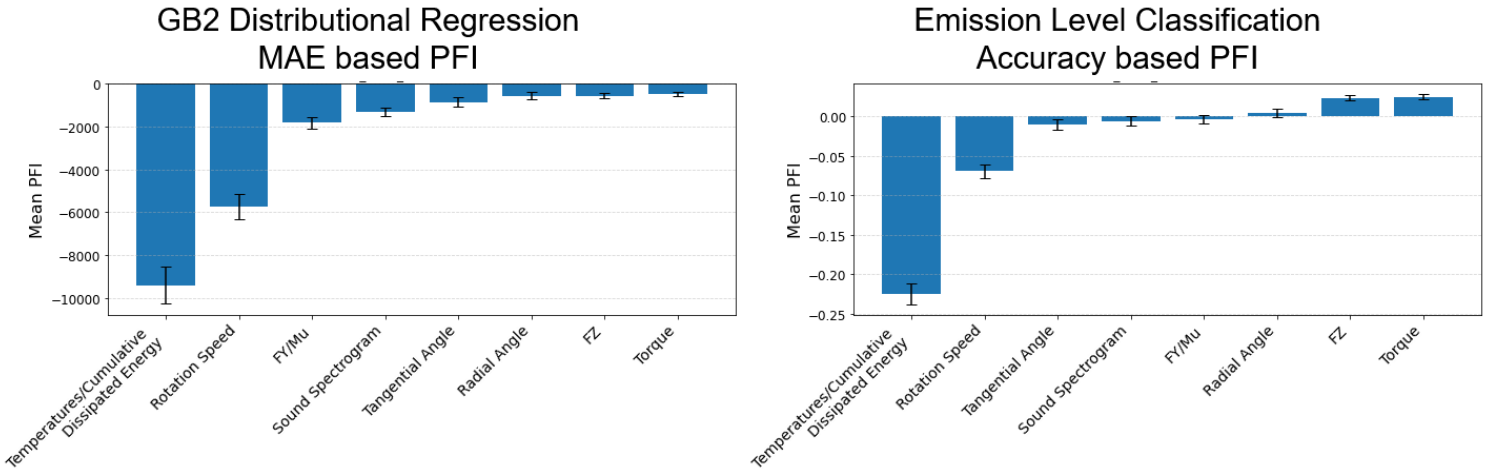


Figure 2.11: EEPS Permutation feature importance per grouped variables

The most important variable groups identified by both models are the same: *Temperatures/Cumulative Dissipated Energy* and *Rotation Speed*. Permuting either group results in a substantial drop in the overall score, confirming their central role in both emission regression and classification. The magnitude of this decrease further reinforces our confidence in the grouping, as it deviates from the typical behavior of PFI on correlated variable groups, which would normally yield only minor decreases. Moreover, the agreement between the two models supports the consistency of these findings and underscores the relevance of these variables.

We initially considered that the MAE-based PFI might disproportionately emphasize high target values due to the long-tailed nature of the emission data—a known limitation of MAE, which tends to assign greater weight to large errors. Classification accuracy, in turn, is also imperfect for ordinal outcomes, since it penalizes all misclassifications equally (e.g., being off by several classes is treated the same as being off by one). Despite these limitations, both metrics independently highlight the same top two variable groups. This strengthens our confidence that these features are genuinely impactful, regardless of the evaluation criterion.

The situation differs for the remaining variable groups, some of which even show slightly positive PFI scores. While this could suggest that these groups contribute little information or potentially introduce noise, it is noteworthy that the *Sound Spectrogram* is among them—despite our earlier observation that including sound data improved overall performance.

Whether the relatively low importance of these groups reflects suboptimal grouping or genuinely lower informativeness remains uncertain. Nonetheless, this analysis confirms that the two leading variable groups—*Temperatures/Cumulative Dissipated Energy* and *Rotation Speed*—are robustly and consistently important across both regression and classification tasks. As for the sound data, we may have overestimated its contribution, and its apparent benefit may arise more from training variability than from a direct predictive effect.

Remarks:

In this study, we evaluated multiple modeling strategies to predict particle emissions measured by EEPS, using mechanical and thermal-related variables, as well as sound, as inputs. Various prediction methods were employed, and their effectiveness was assessed.

Among the regression approaches, GB2 distributional regression proved to be the most reliable, particularly for small to medium emission levels where standard regression methods tended to struggle.

Emission-level classification showed promising results, especially in detecting sparse or low-intensity emissions—areas where GB2 often underpredicted. The classification model frequently predicted the correct class or higher in those cases, while occasionally producing false positives in non-emitting tests.

A category-based analysis of test examples revealed consistent patterns in model performance. Both models handled medium to high-intensity emissions effectively. For smaller emissions, GB2 generally underpredicted, whereas classification frequently overpredicted. Although this observation is qualitative, it provides a sense of an *upper bound/lower bound* relationship between the two models for small emissions.

While the inclusion of sound as an input yielded slight improvements across metrics, permutation feature importance analysis indicated that its impact was less significant than initially anticipated. This raises the possibility that the observed benefit may be due more to training variability than to true predictive power.

Overall, the combination of GB2 distributional regression and emission-level classification—supported by consistently important features such as rotation speed and temperature—provides a robust framework for modeling EEPS emissions. Future work may focus on improving the detection of sparse emissions, enhancing the utilization of secondary inputs, and deepening our understanding of their roles.

2.2.1.2.2 OPS When considering OPS emissions, in contrast to EEPS, the inclusion of sound consistently worsened or failed to improve model performance. This outcome remains consistent with the literature, as OPS particles are micrometric in scale, whereas previously reported associations with squeal and acoustic phenomena pertain primarily to nanometric particles. Consequently, sound was excluded as an input in the subsequent analyses.

Evaluation plots

As for the EEPS case, We begin by examining the evaluation plots corresponding to each

modeling setup. For the regression approaches, the predicted-versus-true and residual-versus-true relationships are presented in Figures 2.12 and 2.13:

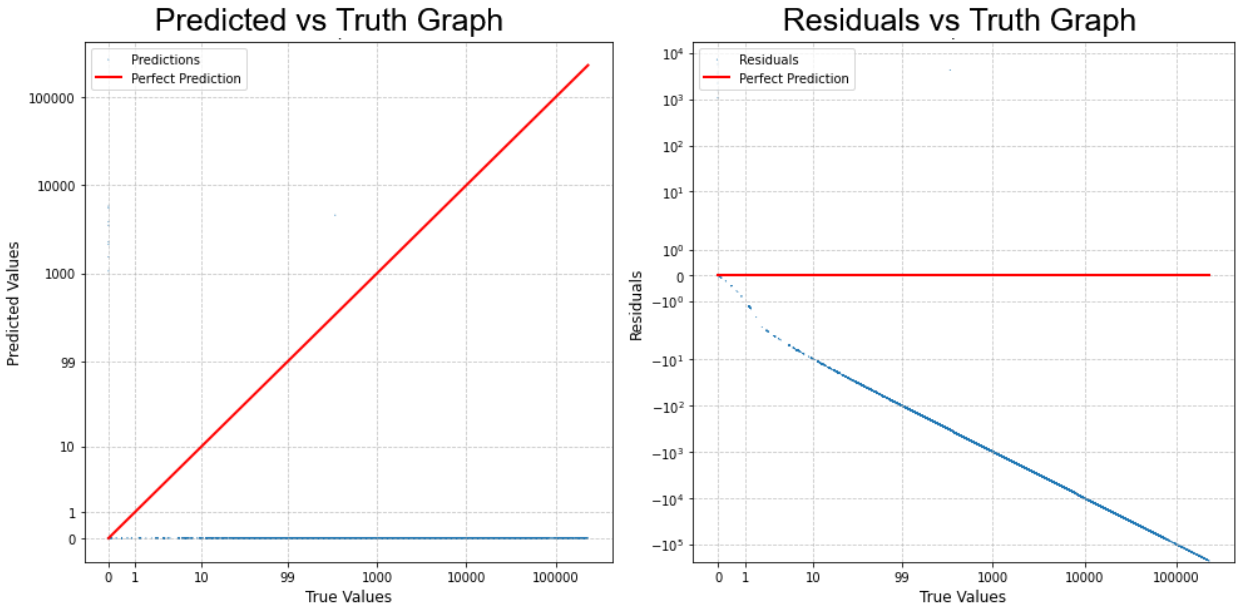


Figure 2.12: OPS Standard Regression evaluation graphs

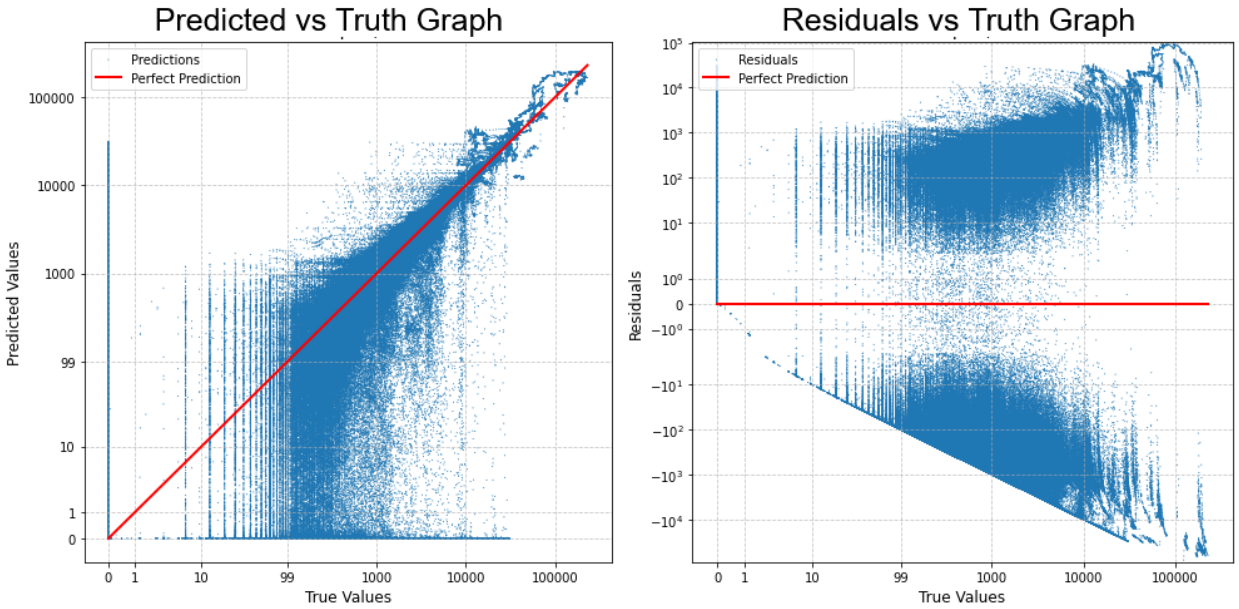


Figure 2.13: OPS GB2 distributional Regression evaluation graphs

In contrast to the findings on EEPS emissions, the Standard Regression model for this dataset collapses entirely, outputting zeros regardless of the true value. This indicates that it failed to learn any meaningful mapping from the data. As before, none of the target-scaling strategies explored (e.g., log or log–log transformations) improved model performance.

The GB2 Distributional Regression model, on the other hand, converged well for high emission values—as evidenced by its correct identification of nearly all large-emission cases—performed moderately for medium emissions, and struggled with smaller ones.

Given these results, and consistent with our previous conclusions, we once again focus on GB2 Distributional Regression for the remainder of this study.

We now turn to the Emission Level Classification model evaluation. The corresponding confusion matrix is presented in Figure 2.14.

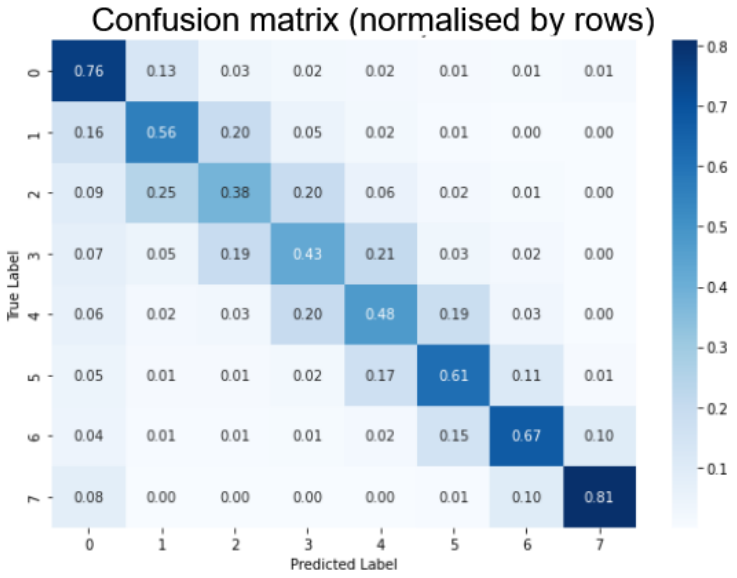


Figure 2.14: OPS Emission level classification evaluation graphs

The resulting confusion matrix is notably less tridiagonal than the one observed with EEPS emissions. While most predictions fall within one class of the true label, misclassifications outside this range are noticeably more frequent than in EEPS, particularly in mid-level emission classes. The model also continues to struggle with the "is there emission or not" sub-problem, with most global errors arising from predicting zero emissions when emissions are present, or vice versa.

Qualitative comparison

Comparing the two models remains challenging for the same reasons discussed previously. Because of this, we once again resort to qualitative observation of the predictions of a few test examples. OPS emissions exhibit significantly simpler patterns than EEPS emissions. Whether this is due to the lower sampling rate or intrinsic properties of the emissions themselves remains unclear. Since we did not observe clear emission categories as we did with the EEPS data, we instead present examples of both successful and unsuccessful predictions across different emission intensity levels.

First emission intensity: Little to No Emission

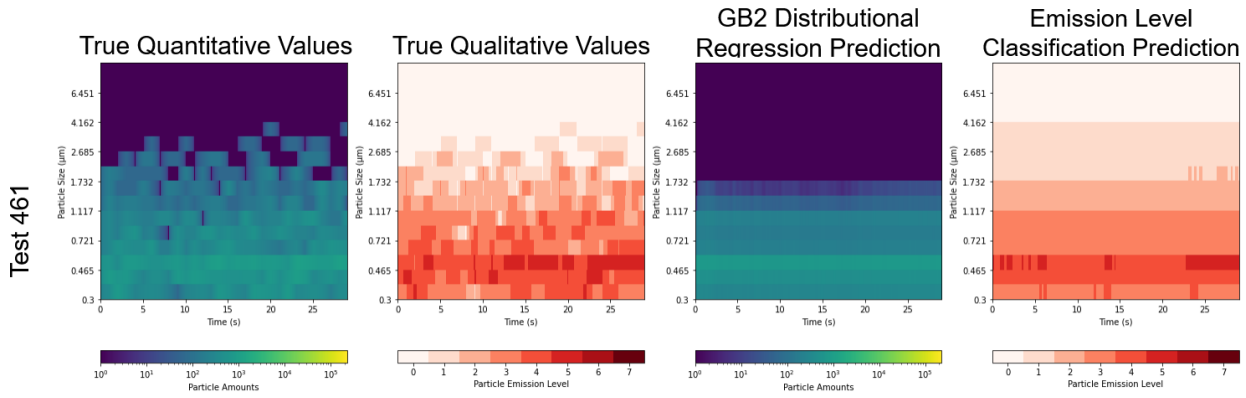
This category is characterized by the absence of detectable emissions or by very small, localized emission events. Figure 2.15 shows two representative tests and the corresponding model predictions:

Figure 2.15: Examples of OPS prediction in the "Little to No Emission" category.

Both models perform well for cases with zero or low emissions, typically predicting zero values. Although they overlook small, localized emission events, it still reflects a correct understanding that emissions are largely absent. Nonetheless, the number of incorrect predictions is not negligible. As shown in test example 71, the models—particularly the classification-based ones—can occasionally "hallucinate" emissions, predicting positive values where none actually occurred.

Second emission intensity: Medium Emissions

We now turn to predictions for test examples with medium-level emissions. Figure 2.16 shows two representative tests and the corresponding model predictions:



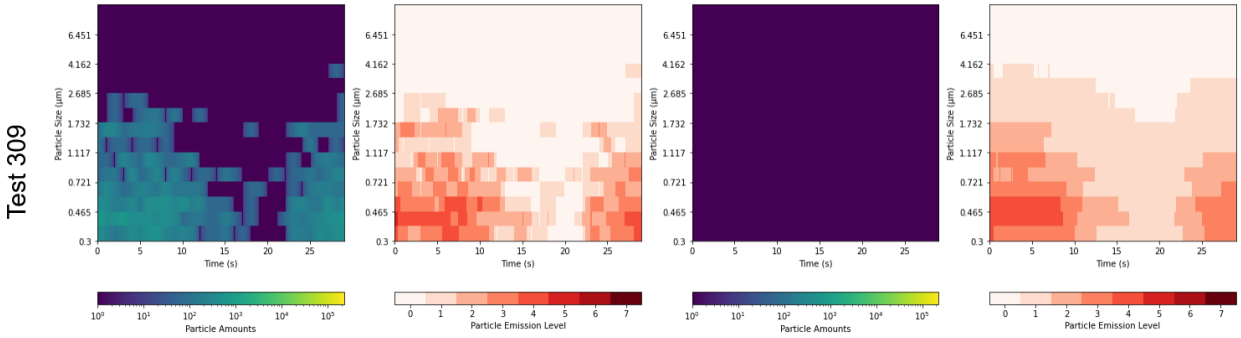


Figure 2.16: Examples of OPS prediction in the "Medium emissions" category.

For medium emissions, the emission-level classification model performs well, capturing emission patterns in most cases, as seen in tests 461 and 309. In contrast, the GB2 Distributional Regression model occasionally approximates the general emission pattern (e.g., test 461), but more often fails by predicting zero emissions throughout the test (e.g., test 309).

Third emission intensity: High Emissions

Finally, we examine predictions for test examples with high emission levels. Figure 2.17 shows three representative tests and the corresponding model predictions:

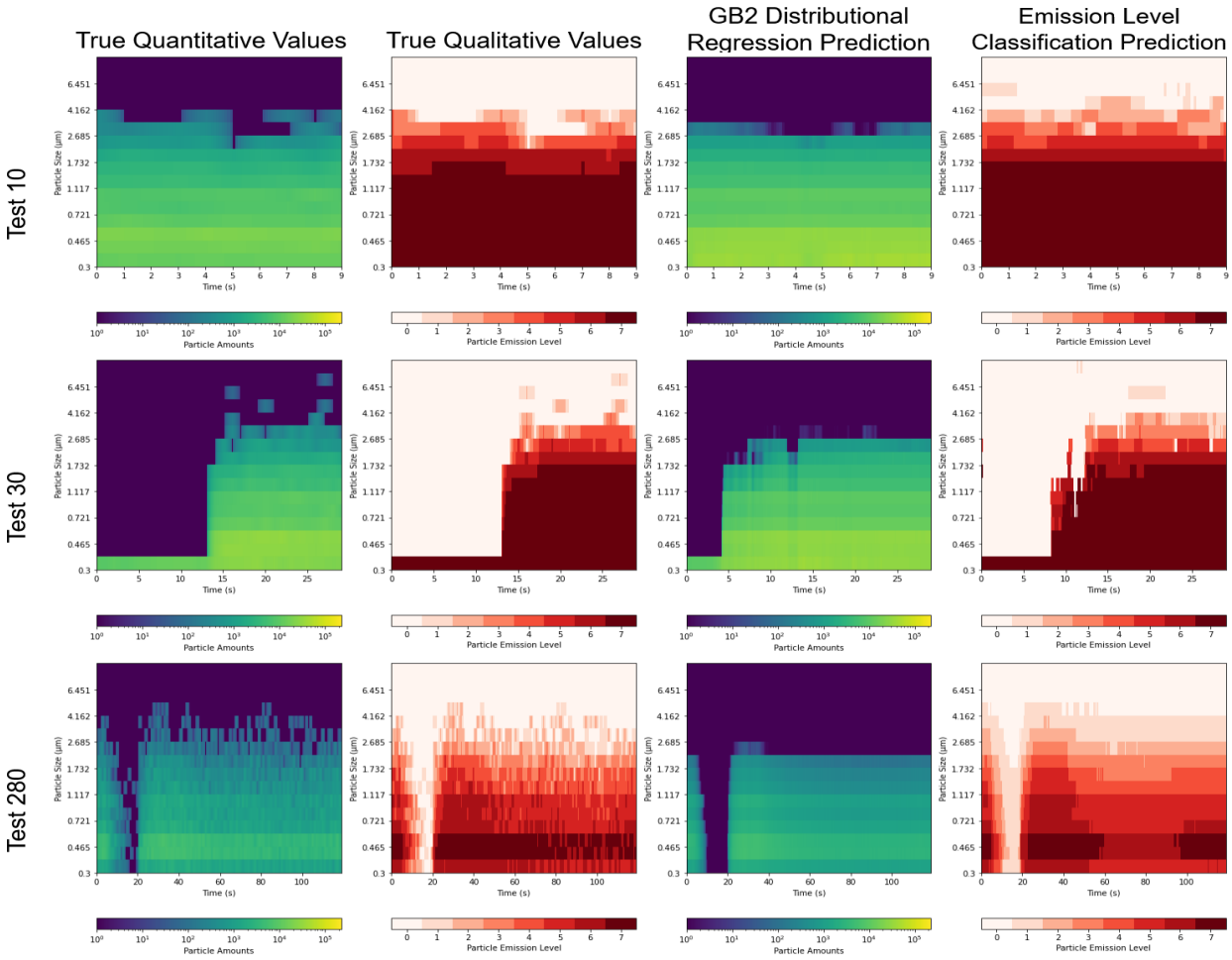


Figure 2.17: Examples of OPS prediction in the "High emissions" category.

Both models generally perform well when emission levels are high. This holds true for relatively simple, near-constant emissions (e.g., test 10), as well as for more complex patterns (e.g., tests 280 and 30 — though test 30 shows significant errors). We did not observe any cases of catastrophic prediction failure, though such cases may exist, and the definition of such failures is highly subjective.

Feature Importances

Next, we examine the feature importance results. Similar to previously, we computed Permutation Feature Importance (PFI) (see Section A.10), based on the MAE score for the GB2 distributional regression and accuracy for the Emission Level Classification. The input variables were grouped as defined in Section 1.4.3 to mitigate the effect of correlation on the PFI computation. The resulting PFIs are presented in Figure 2.18.

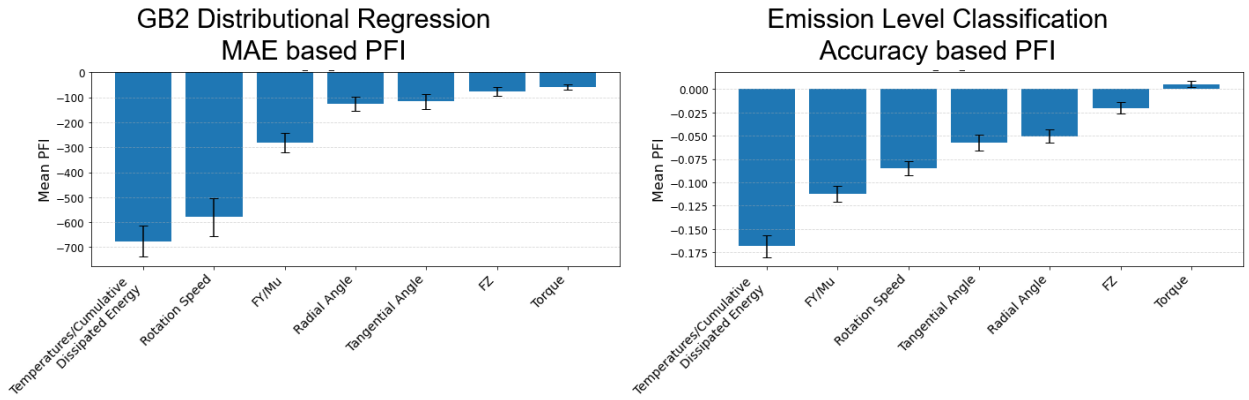


Figure 2.18: OPS Permutation feature importance per grouped variables

Both models identify the "Temperatures / Cumulative Dissipated Energy" group as the most important feature set. Additionally, the two models share the same top three important groups, with "Rotation Speed" and "FY/Mu" following closely behind. Beyond these, the importance values tend to diminish significantly.

As before, it remains difficult to determine whether the lower importance of some groups is due to suboptimal grouping or their genuinely low informativeness for the predictions. However, the consistent appearance of the same top three groups across both models strengthens our confidence that these variables groups are indeed impactful for the task at hand.

Remarks :

In this study, we evaluated multiple modeling strategies to predict particle emissions measured by OPS, using Thermal and mechanical related signals as inputs. Several prediction methods were tested, and their performances were assessed.

Among the regression techniques, the GB2 distributional regression outperformed the standard regression, which only predicted zero values due to bad optimization.

The emission level classification performed well, particularly in detecting medium emission levels, where GB2 distributional regression tended to underpredict.

As with EEPS, prediction performance improves with emission magnitude: the higher the emission amount, the better the models perform. However, OPS emissions exhibit lower predictive accuracy—especially for medium and low intensities. Whether this performance gap stems from the lower sampling rate or intrinsic variability in OPS emissions remains unclear.

Overall, combining both models—supported by consistently important features like Temperatures, Rotation Speed, and FY/Mu—provides a robust framework for predicting high emissions. Future work should focus on improving predictions for medium and low emissions, potentially through enhanced feature engineering to better leverage the less influential variables.

2.2.1.2.3 Sound Finally, we present the results of the sound emission modeling. Compared to particulate measurements, the sound channel is acquired at $\sim 50{,}000\,\text{Hz}$ and then converted to spectrograms with an effective frame rate of $\approx 50\,\text{Hz}$ (given the chosen STFT hop). This much higher raw sampling rate captures rapid, narrow-band transients (e.g., squeal) as well as broader, low-intensity structures, resulting in more intricate and highly variable temporal patterns. Consequently, the sound data are inherently more complex—and potentially harder for the models to learn—than the more slowly varying particulate emissions.

Evaluation plots

As with the previous studies, we begin by examining the evaluation plots corresponding to each modeling setup. For the regression approaches, the predicted-versus-true and residual-versus-true relationships are presented in Figures 2.19 and 2.20.

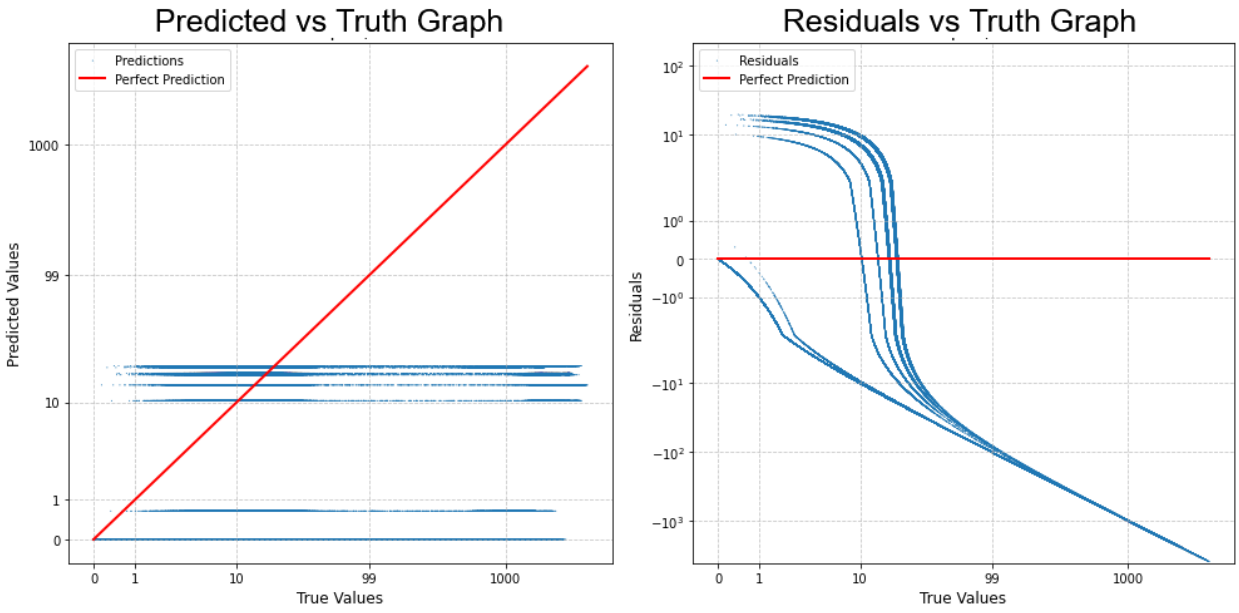


Figure 2.19: Sound Standard Regression evaluation graphs

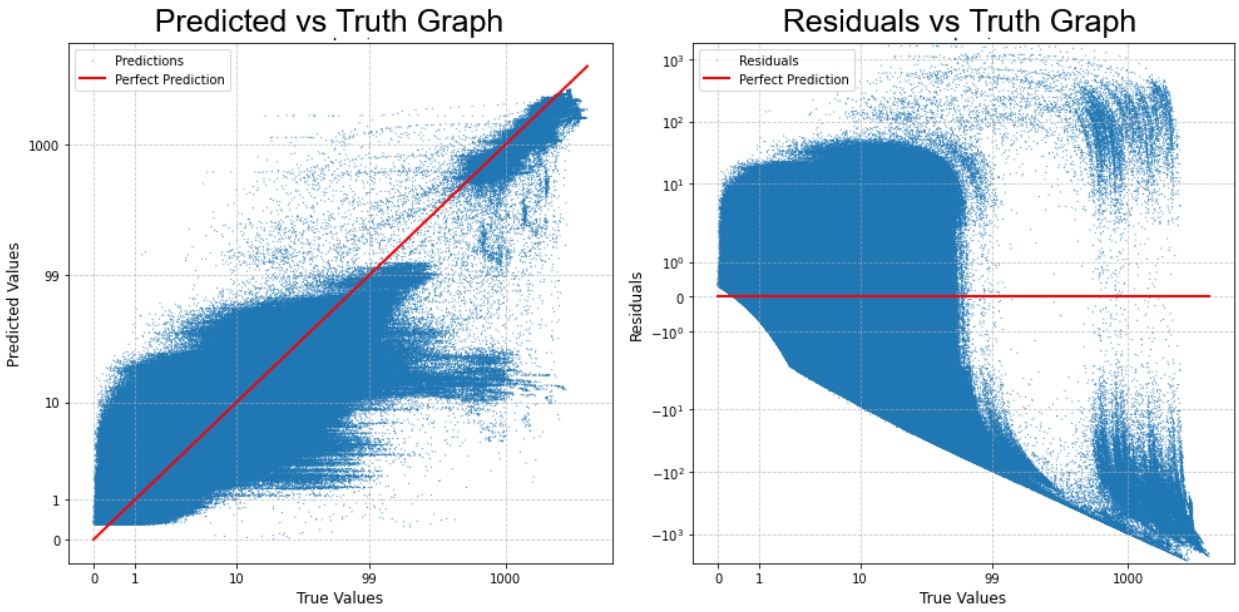


Figure 2.20: Sound GB2 distributional Regression evaluation graphs

Once again, GB2 distributional regression outperforms Standard Regression, with standard regression collapsing entirely, outputting a few different modes regardless of the true value—a clear failure to learn meaningful mapping (even when using target scaling strategies such as log and log-log). As far as GB2 distributional regression goes, we see a different patterns : when sound amplitude is around 1000 (90 dB), the predictions are relativly close to the truth. For the rest of the small and medium values, the predictions are less accurate. Given these results, we once again focus on GB2 Distributional Regression for the remainder of this study.

We now turn to the Emission Level Classification model evaluation. The corresponding confusion matrix is presented in Figure 2.21.

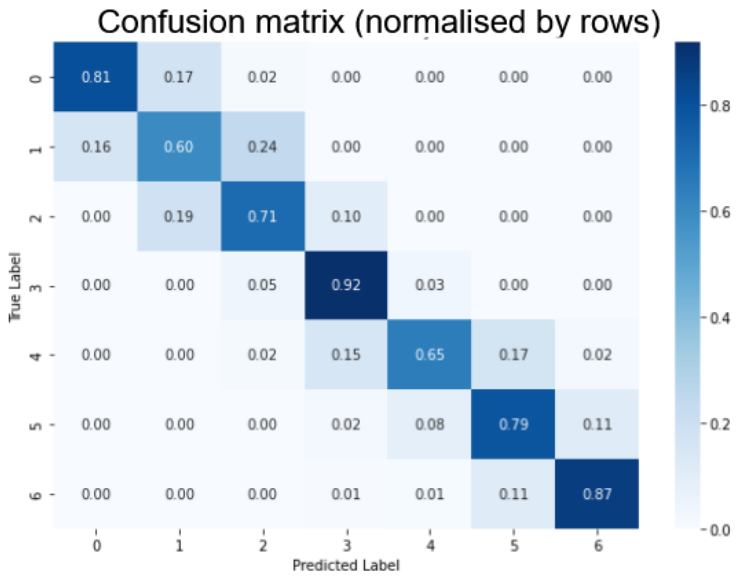


Figure 2.21: Sound Emission level classification evaluation graphs

The confusion matrix is once again largely tridiagonal. In this configuration, the classes are

defined without a "zero-emission" category, removing the "emission vs. no-emission" sub-problem observed in earlier models.

Qualitative comparison

As in the previous sections, comparing the two models remains difficult for the same underlying reasons. We therefore again rely on qualitative inspection of a few representative test examples. Due to the experimental setup—including the test bench and the materials used for the disk and pin—the acoustic behaviors captured in the dataset are relatively simple. As a result, we focus on two key aspects of the spectrogram: low-intensity patterns and squeal.

Low-Intensity Patterns:

The low-intensity components observed in the spectrograms are primarily attributed to background noise and mechanical noise emitted by the bench. However, differences between tests suggest that these patterns are influenced by the specific experimental conditions. This makes them a meaningful target for evaluation, as we can assess whether the models are capable of accurately predicting these faint but potentially informative signals. Figure 2.22 shows two representative tests and the corresponding model predictions:

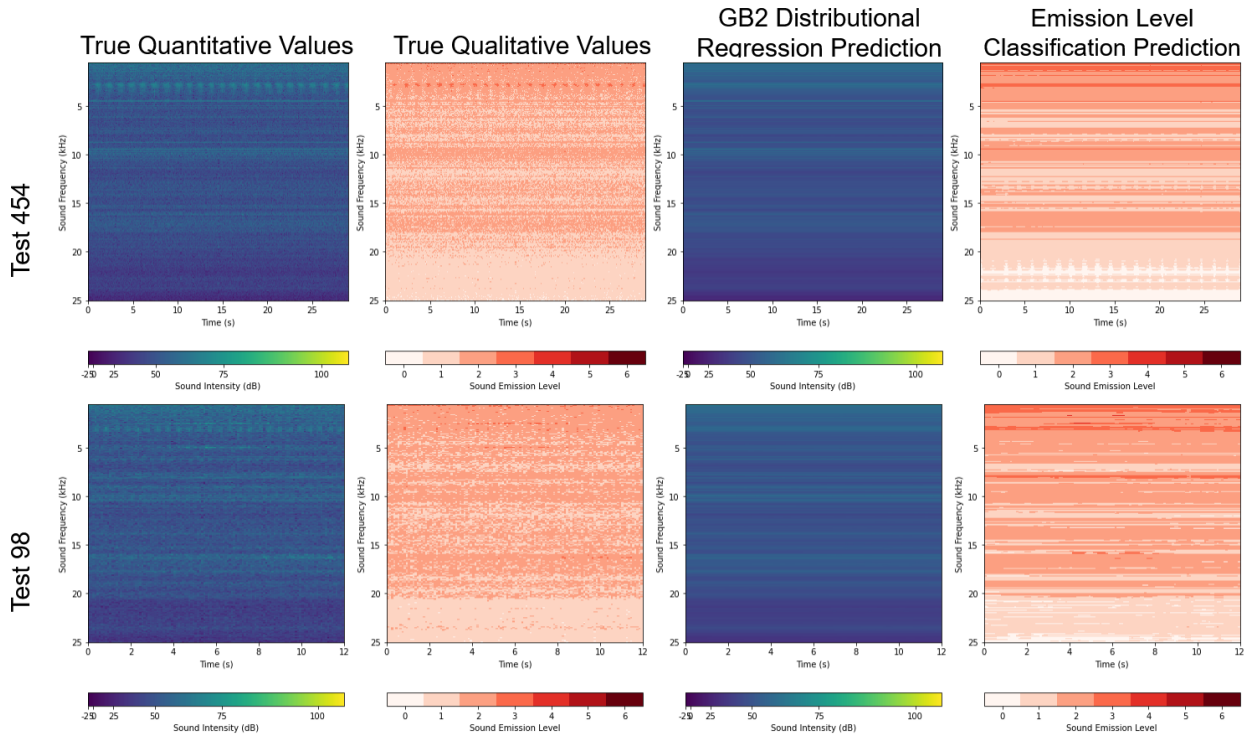


Figure 2.22: Examples of Sound prediction in the "Low-Intensity Patters" category.

Neither test 98 nor test 454 exhibits high-intensity behaviors. However, they display distinct low-intensity patterns. Test 454 shows steady emissions around 4.5 kHz, a cyclic pattern near 2.5 kHz, and broad emission bands around 10 kHz and 17.5 kHz. In contrast, test 98 also exhibits a cyclic pattern at 2.5 kHz, but lacks the same wide emission bands. Instead, it presents a narrower band around 10 kHz and a more pronounced one at 17.5 kHz.

A few observations can be made regarding the model predictions:

- Both models detect the cyclic patterns but tend to interpret them as constant emissions rather than capturing their periodic nature.

- Both predictions resemble the true spectrograms reasonably well, though in a notably smoothed form.
- Visual inspection suggests that GB2 Distributional Regression performs better overall. It approximates the actual emissions more accurately and avoids the tendency to hallucinate high values—an issue observed in the Emission Level Classification model, particularly in test 98. One should however note that this is highly subjective.

Squeal:

Squeal refers to high-intensity, high-frequency acoustic events that typically appear as narrow, well-defined bands in the spectrogram. These events are often linked to friction-induced instabilities and are of particular interest due to their distinct spectral signatures and practical relevance in diagnostics. Unlike low-intensity patterns, squeals are much easier to identify visually. Figure 2.23 shows three representative tests in which squeal occured and the corresponding model predictions:

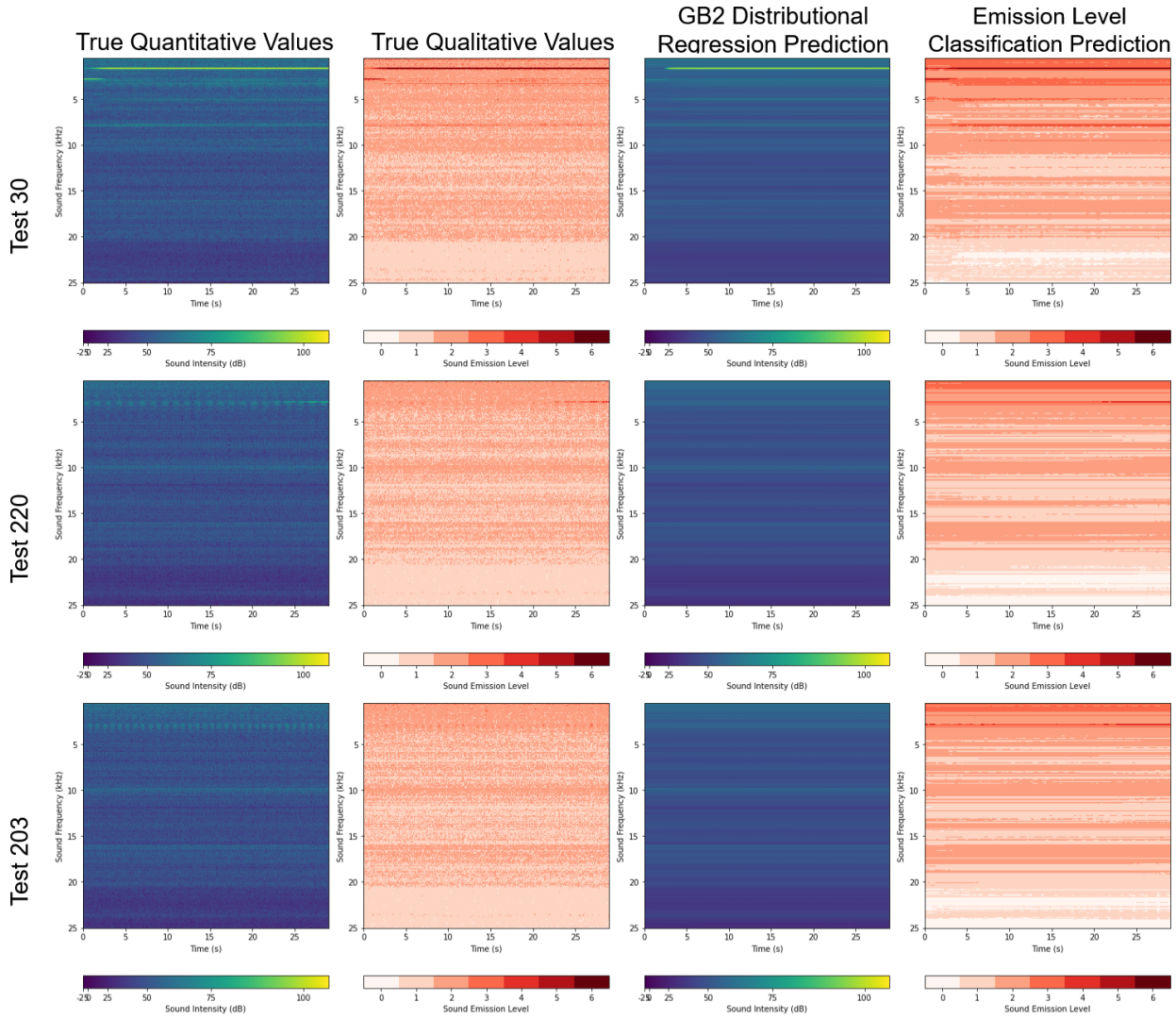


Figure 2.23: Examples of Sound prediction in the "Squeal" category.

For clear, well-defined squeal events—such as the 1.5 kHz squeal observed in test 30—both models predict the event reasonably accurately, although their predictions tend to begin slightly late. For smaller or less prominent events, like the 3 kHz squeals in tests 30 and 220,

the GB2 Distributional Regression model often fails to detect them, defaulting to background-level predictions. In contrast, the Emission Level Classification model occasionally captures these events. However, the classification model is also prone to false positives, sometimes hallucinating squeal events, such as in test 203 around 3 kHz.

Feature Importances

We can now examine feature importance results. As we did before, we computed Permutation Feature Importance (PFI) (see Section A.10) based on the MAE score for the GB2 distributional regression and accuracy for the Emission Level Classification. The input variables were grouped as defined in section 1.4.3 to mitigate to effect of correlation on the PFI computation. The resulting PFIs are presented in Figure 2.24:

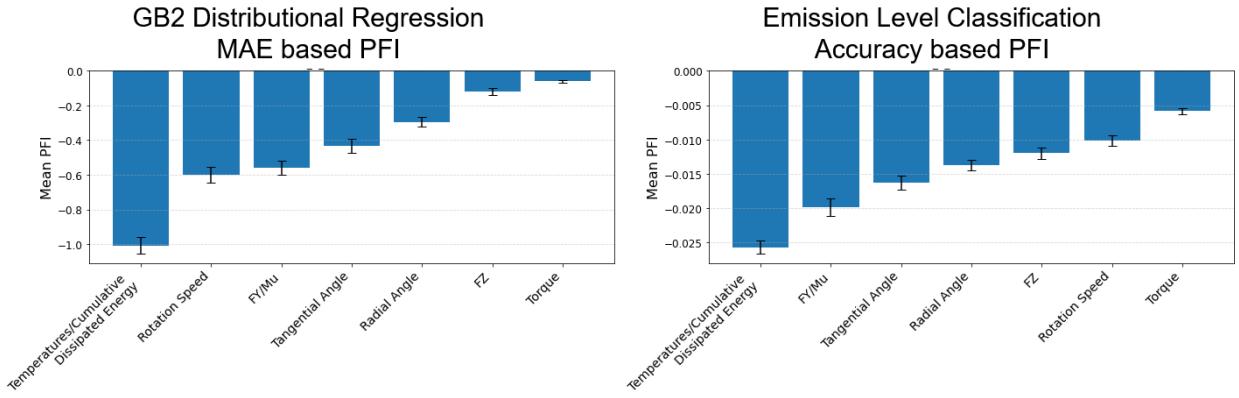


Figure 2.24: Sound Permutation feature importance per grouped variables

Unlike previous results, where both model types agreed closely on feature importance rankings, this is no longer the case. While both models still identify the Temperatures / Cumulative Dissipated Energy group as the most influential, their treatment of the remaining variable groups diverges significantly.

GB2 Distributional Regression exhibits a drop in importance values followed by a sharp decline, ending with FZ and Torque appearing nearly irrelevant. In contrast, the Emission Level Classification model shows a more gradual decline in feature importance, with no input group appearing entirely dispensable.

This discrepancy may be partially attributed to the nature of the evaluation metrics: MAE is known to be sensitive to target distribution bias, which can cause models to prioritize only the most dominant signals. Classification, by contrast, builds decision boundaries, potentially incorporating weaker but still informative features that help distinguish between classes. However, because the task involves ordinal classes, standard classification accuracy does not fully capture prediction quality — it treats all errors equally, even though misclassifying by several classes is more severe than being off by one.

It is also important to acknowledge that these feature importances reflect how the current models use the input variables—not necessarily the true strength of the relationships between those inputs and the target. In other words, they indicate learned dependencies, which may still be limited by model imperfections or underfitting.

Remarks :

The sound modeling task reveals both parallels and key differences compared to particle emission modeling. As before, GB2 Distributional Regression outperforms Standard Regression, which fails to learn meaningful mappings. Both GB2 and Emission Level Classification capture background noise and cyclic acoustic patterns reasonably well, although often in a smoothed or simplified form. They also perform well on clear, well-defined high intensity events -such as squeal- accurately identifying their frequency and duration. However, for faint but high-intensity events, classification tends to be more sensitive and often detects them more reliably. GB2, in contrast, frequently underpredicts these localized peaks, defaulting to background-level values.

In contrast to the consistency observed in emission modeling, the feature importance profiles diverge more significantly in the sound prediction task. While both models rank Temperatures / Cumulative Dissipated Energy as the most informative input group, GB2 shows a sharp drop-off in importance beyond this, effectively ignoring groups like FZ and Torque. The classification model distributes importance more gradually across inputs, suggesting a broader, less selective use of features. This likely reflects both the models' objectives and their loss functions—MAE favoring dominant global patterns, while classification can leverage multiple weak signals to construct class boundaries. Importantly, these feature importances reflect the models' internal usage patterns, not necessarily the true underlying causal relationships.

2.2.1.3 Conclusion

In this subsection, we assessed the effectiveness of different modeling strategies—standard regression, GB2 distributional regression, and emission-level classification—for predicting each pollution signal independently: EEPS, OPS and Sound.

Across all targets, standard regression consistently failed to learn meaningful mappings, often collapsing to constant outputs. This is likely due to the highly skewed nature of our target distributions; however, no form of target scaling (e.g., log or log-log) enabled successful training. Standard regression only showed partial convergence for the EEPS dataset, which had the most data available (32 channels at 10 Hz), and completely failed for OPS (16 channels at 1 Hz) and Sound (1 channel at 50,000 Hz). In contrast, GB2 distributional regression converged across all targets, suggesting it may be more sample efficient—better able to extract useful patterns even from more smaller or lower-resolution input datasets.
Due to this, GB2 distributional regression emerged as the most reliable regression-based approach, especially for high-emission or high-intensity events. However, it often underpredicted low and medium values, particularly in the OPS and Sound cases, where the signal-to-noise ratio or complexity may have limited its sensitivity.

Emission-level classification showed complementary strengths. It performed better on medium emissions and faint acoustic events but suffered from occasional overpredictions, including hallucinations in low- or no-emission scenarios. This trade-off between conservativeness and sensitivity was consistent across pollution types.

Feature importance analysis revealed both consistent patterns and task-specific differences. Across all targets, Temperatures / Cumulative Dissipated Energy consistently emerged as the most influential input group, highlighting its central role in pollution modeling. For particle emissions (EEPS and OPS), both GB2 Distributional Regression and Emission-Level Classification largely agreed on the ranking of key input groups, suggesting that the relationships between inputs and emissions are relatively stable and similarly exploited by both models. In contrast, for sound prediction, the two models diverged more noticeably in the way they assigned importance to input features. The GB2 model concentrated on a narrow set of dominant variables, whereas the classification model distributed its attention more evenly across a broader range. This lack of agreement likely indicates that sound is more difficult to predict from the available input signals, either because those signals are only weakly related to acoustic behavior or because of the inherent noise present in the sound data itself.

It's worth noting that all results in this section were obtained using optimized cross-validation splits designed to balance experimental settings. Similar experiments using random or unbalanced splits led to significantly degraded performance. This reinforces the importance of thoughtful experimental protocol—particularly the need for diverse, representative training data—in pollution modeling tasks.

Taken together, these results suggest that using GB2 and classification models in tandem can help compensate for their respective weaknesses—primarily because neither model alone is yet sufficiently reliable across all emission types and intensities. Even when using both, overall performance remains far from optimal. Medium and low-intensity events, in particular, remain challenging to capture consistently. Improvements could likely come from increased data diversity, better feature representations, or more advanced architectures. To try and improve upon these results, the next section explores a multi-target prediction approach, where all pollution signals are modeled jointly rather than separately.

2.2.2 Multi-Target Pollution Prediction

In this section, we extend our analysis by jointly modeling multiple pollution signals—EEPS, OPS and Sound—using the available physical and mechanical input variables. Unlike single-target approaches, multi-target prediction allows the model to learn a more generalized encoder representation. These generalized representations can enhance predictive accuracy and enable the encoder to be effectively applied to a range of related tasks without requiring full re-training.

2.2.2.1 Method

This approach differs from the one described in Section 2.2.1.1 primarily in the model architecture and the loss function. The revised setup is illustrated in Figure 2.25.

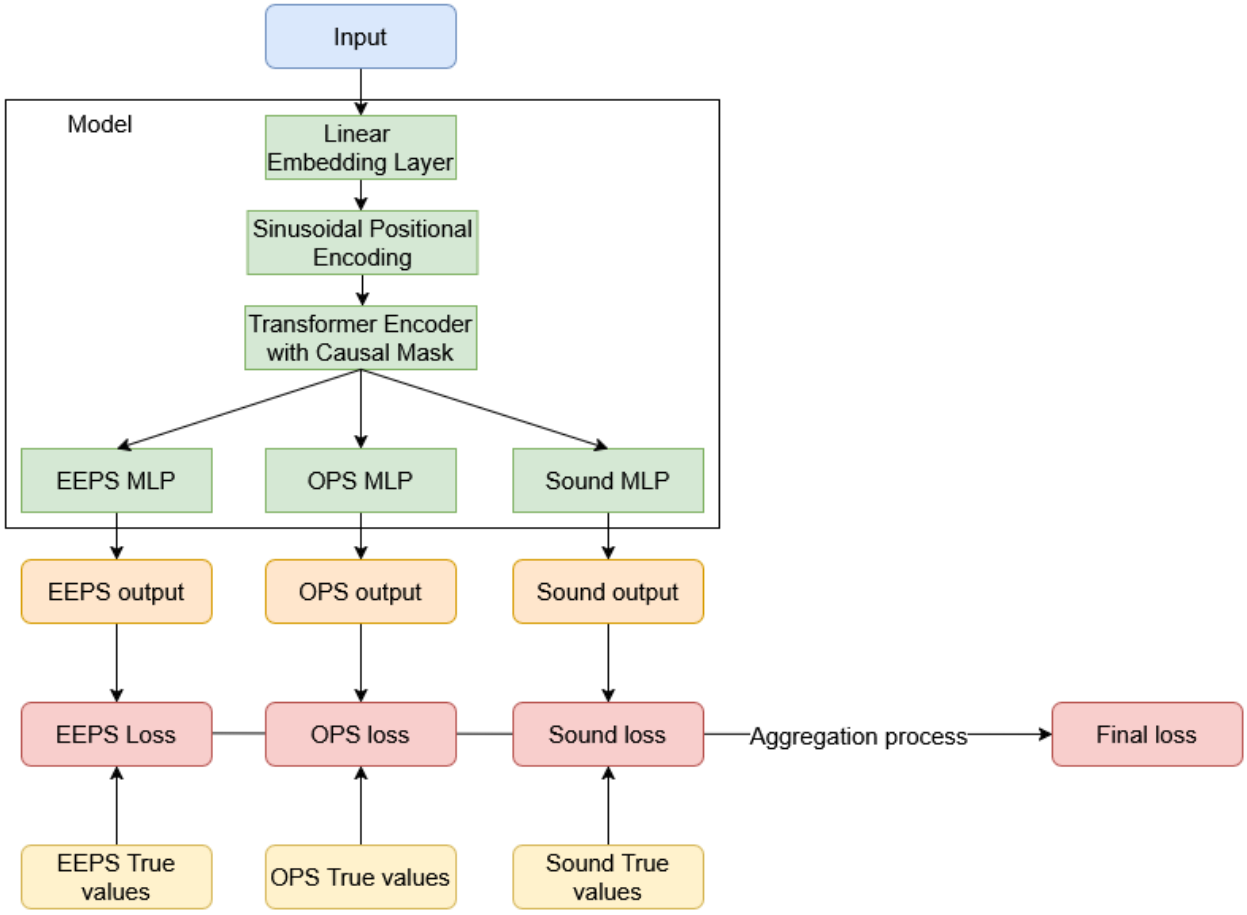


Figure 2.25: Multi-target model architecture and optimization process.

The loss aggregation process is designed to assign equal importance to each time series prediction across all targets. For each target, the individual losses are first averaged along the time dimension. The resulting loss tensors are then flattened and concatenated and a final mean is computed. This strategy mitigates the risk of longer time series dominating the overall loss.
Finally, to accommodate the increased complexity of training a multi-target model, we extended the number of training epochs to 8000.

2.2.2.2 Results

Using mechanical and thermal related variables as inputs, we obtained the results summarized in Table 2.2.

	Standard Regression (MAE)	Emission level classification (Accuracy)	GB2 distributional regression (MAE)
EEPS	3609.1 ± 18793.2	0.753	2541.9 ± 12953.4
OPS	855.1 ± 4767.2	0.694	362.3 ± 2203.1
Sound	8.7 ± 22.8	0.653	3.8 ± 30.0

Table 2.2: Multi-Target prediction overall evaluation metrics

Training and evaluation of the models required approximately 150 hours in total using a 3-way MIG setup. See Appendix A.1 for details on the software, hardware, and computational

resources used.

It is clear that none of the multi-target models drastically improve on the single-target baselines, that is to say improve on all target at once. Whether this is due to limited data or suboptimal training hyperparameters remains to be determined.

The fact that the global model performs comparably to baseline results, leaves us hopeful that with more data, the models could converge more effectively—yielding improved performance and more generalizable encoder representations that could be used for other tasks.
Possible avenues for improvement include using a larger model, applying multi-task loss weighting strategies such as uncertainty-based weighting [31], incorporating pretraining through input reconstruction in an encoder–decoder framework and many more.

2.3 Conclusion

This chapter explored the application of machine learning techniques for predicting braking emissions, with a particular focus on both forecasting and real-time prediction tasks. Our aim was to evaluate the effectiveness of various modeling strategies in capturing the complex dynamics of emissions, while also providing valuable theoretical insights through explainability methods.

The forecasting task was inspired by previous work conducted in our lab, where temperature signals were successfully forecasted with good precision. This motivated us to apply similar time series models to forecast sound emissions. However, despite experimenting with different architectures, including LSTM, GRU, and Transformer-based models, we encountered significant challenges. The forecasting models struggled to capture meaningful temporal dynamics in sound emissions, unless true Normal Force values were provided as inputs. When the true normal force values were incorporated, the models showed partial progress, but further analysis revealed that the model was heavily reliant on these normal force values, which prompted a shift toward real-time prediction.

In the context of real-time pollution prediction, we focused on predicting particulate and sound emissions independently, using mechanical and temperature-related variables. We explored three primary modeling approaches: regression, classification, and distributional modeling (GB2). For EEPS emissions, incorporating sound as an additional input improved model performance across all methods. This is consistent with theoretical reasoning, as both fine particles and sound emissions arise from similar underlying mechanisms, such as surface instabilities or sustained frictional contact during braking. In contrast, for OPS emissions, including sound did not improve model performance, leading us to exclude it from further analysis. This decision is theoretically justified, as coarser particles (measured by OPS) are typically linked to more abrupt mechanical events, such as chunk detachment or large-scale wear phenomena, which are less systematically captured by acoustic emissions.

When evaluating prediction methods, GB2 distributional regression consistently outperformed standard regression. The comparison between classification and GB2 distributional regression was more complex, as these models are evaluated using different metrics. We found that the two methods are comparable for high emissions, but for medium emissions, classification outperformed GB2. However, GB2 distributional regression offers the advantage of providing continuous predictions, which are more precise than the discrete intervals produced

by classification. Thus, the choice between classification and GB2 depends on whether the priority is accuracy in discrete predictions or continuous uncertainty modeling. In practice, both models could complement each other when used jointly.

Feature Importance (PFI) analysis revealed that for both EEPS and OPS, the most influential variables were temperature-related variables, along with Cumulative Dissipated Energy and Rotation Speed. This is further reinforced by the fact that both classification and GB2 distributional regression models showed the same feature importance pattern, reassuring us in our interpretation. However, for sound emissions, the feature importance results diverged between the two models. While both identified temperature related variables and Cumulative Dissipated Energy as significant, the divergence in feature importance for sound emissions suggests that the models may not have fully converged. This raises concerns about the robustness of the sound prediction models, suggesting that further refinement is needed.

In addition to modeling each emission type independently, we also explored a multi-target approach, jointly predicting EEPS, OPS, and Sound emissions. However, this strategy did not outperform the single-target baselines, suggesting that—given the current architecture and dataset—multi-target modeling does not yet provide a clear advantage. We hypothesize that with larger and more diverse datasets, multi-target architectures could become more effective. In particular, encoder–decoder designs—feasible only with substantially more data—might help eliminate the need for interpolation and enable more robust joint learning across emission types.

The next chapter moves beyond raw prediction to focus on **understanding acoustic emissions**. Instead of forecasting emission levels directly, we will cluster acoustic responses into recurring patterns and then explain how these patterns arise from auxiliary conditions. This shift places emphasis on *interpretability*: by combining clustering, classification, and explainability tools, the goal is not only to recognize different emission states, but also to shed light on the physical mechanisms that generate them.

Chapter 3

From Prediction to Mechanical understanding: Insights into Acoustic Emissions

The previous chapter evaluated machine learning methods for predicting braking emissions. While certain models achieved partial success, overall performance remained limited, and our efforts to interpret the predictions yielded inconsistent results: although PFI highlighted relevant thermal and mechanical variables, the results varied across models and suggested that the underlying mechanisms were not being consistently captured. These outcomes pointed to a central limitation of prediction-focused approaches—useful for correlation, but not necessarily for explanation.

In this chapter, we adopt a different perspective. Rather than attempting to predict emission values directly, we aim to **understand acoustic emissions by organizing them into interpretable patterns**. Acoustic signals are particularly well suited for such an approach: their spectrograms exhibit structured frequency bands and recurrent motifs that likely correspond to distinct physical processes during braking. By clustering these acoustic responses, we can uncover latent "acoustic states" that group together similar spectral signatures.

To connect these states back to the system, we then train classifiers that predict cluster identity from auxiliary variables such as forces, rotation speed, and thermal signals. Interpreting these classifiers with explainability tools like Integrated Gradients [62] allows us to highlight which variables and time segments are most influential, offering insights into the mechanisms behind each emission pattern. In this way, the chapter reframes the problem from "predicting emission values" to "explaining emission states."

This exploratory approach also opens perspectives for future work. By examining threshold effects (e.g., 55 dB vs. 75 dB), we can separate artefacts from genuine phenomena, and by framing emissions as combinations of frequency-band sources, we set the stage for more principled methods such as Non-negative Matrix Factorisation (NMF).

In short, the goal of this chapter is not higher prediction accuracy, but deeper understanding: uncovering structured links between acoustic emissions, auxiliary signals and the physical mechanisms that generate them.

3.1 Motivation

The motivation for this method originated from preliminary exploratory analyses conducted upon receiving the dataset. Our initial goal was to discover hidden structures within the large volume of data by clustering the sound spectrograms of various experiments.

We treated each spectrogram as a grayscale image. To address variability in time duration, we padded all spectrograms along the time axis to match the 95^{th} percentile length found in the dataset. We then applied max pooling to downsample the spectrograms to a uniform size of 256×256, preserving prominent features more effectively than average pooling. Afterward, we flattened the resulting images and applied Principal Component Analysis (PCA) for dimensionality reduction, followed by KMeans clustering. Figure 3.1 shows examples of grouped spectrograms.

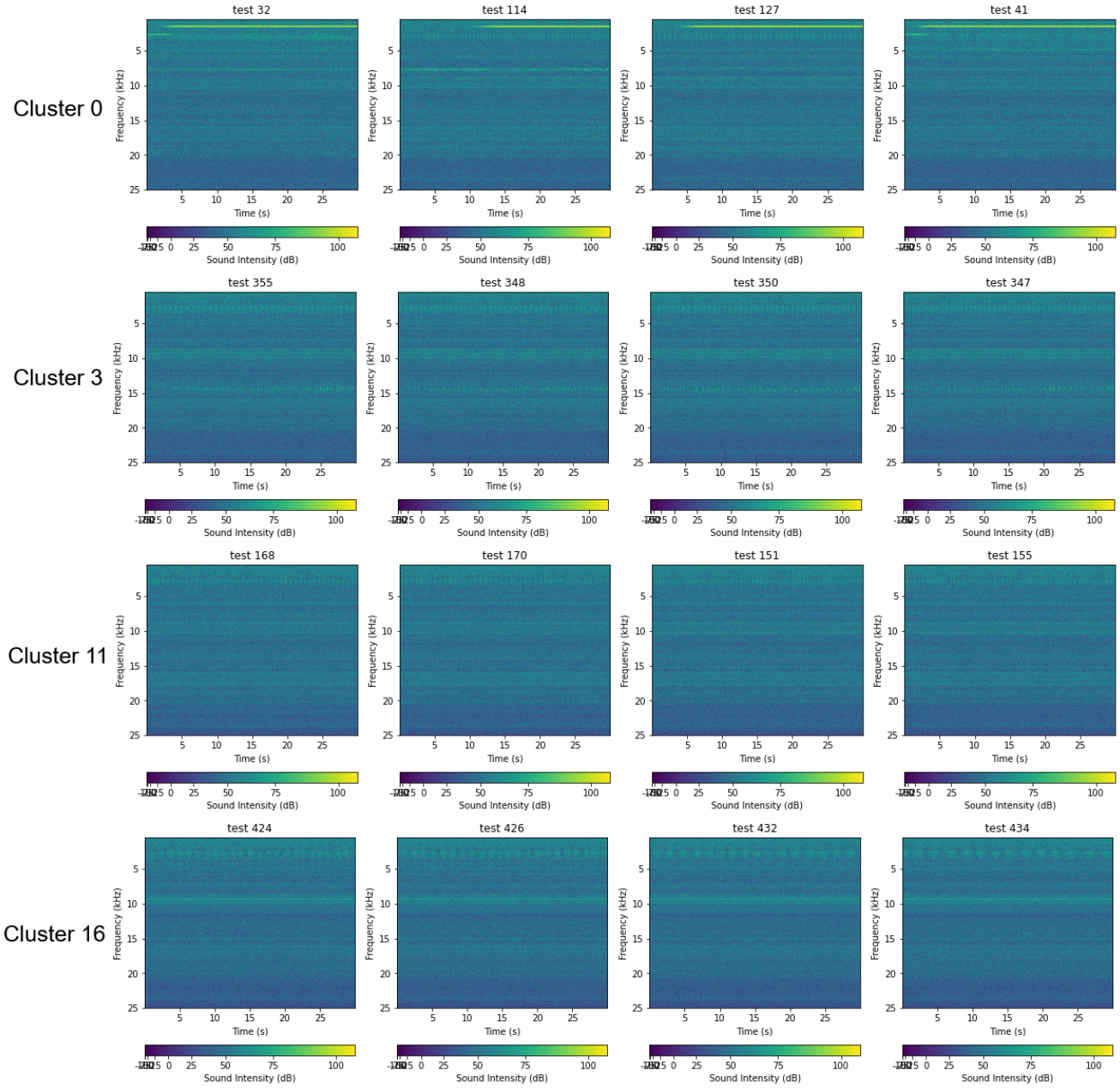


Figure 3.1: Examples of clustered Spectrograms. Clusters are not shown in full.

Clear visual differences emerge between clusters:

- **Cluster 0** is characterized by the presence of intermittent high-intensity events exceeding 100 dB, concentrated primarily around 2 kHz. These appear as narrow, bright horizontal streaks, indicative of short, high-energy squealing phenomena.
- **Cluster 3** displays a more structured background, with a distinct emission band around 9 kHz and a noisier, less coherent feature emerging around 14 kHz. This may suggest overlapping mechanisms or variable noise sources at higher frequencies.
- **Cluster 11** lacks both the 9 kHz and 14 kHz features found in Cluster 3. Instead, it shows a broader emission band spanning approximately 8–11 kHz, possibly indicating a wider-band but less intense acoustic process.
- **Cluster 16** resembles Cluster 3 in exhibiting a prominent band around 9 kHz, but lacks the messy 14 kHz emission. Its consistency suggests a stable, recurring mechanism producing tonal emissions at that frequency.

Interestingly, we also observed a temporal coherence within the clusters: most clusters tend to group tests that were conducted in close chronological proximity. This suggests that background noise patterns—which dominate in most clusters—may be influenced by slowly varying experimental conditions, such as pin surface state. An exception is Cluster 0, which aggregates isolated high-intensity squeal events regardless of when they occurred.

While this method is not without limitations—particularly the padding of spectrograms to uniform length, which may obscure some temporal dynamics—it provided valuable initial insights. More advanced distance measures, such as Dynamic Time Warping (DTW), could further enhance the clustering by allowing better alignment of events that occur at different times.

Nevertheless, clustering full spectrograms revealed recurring patterns across multiple experiments. These repeated structures suggest that common underlying mechanisms are producing similar acoustic signatures, even under varying conditions. The objective of this chapter is to investigate these distinct acoustic “states” and understand the mechanisms responsible for them. By identifying and explaining these patterns, we aim to gain deeper insights into the emission processes—insights that may ultimately prove more informative than direct predictive modeling alone.

3.2 Method

Rather than focusing on direct prediction of emission values, this approach centers on understanding the structure of acoustic emissions. Specifically, we aim to identify recurring acoustic patterns and then interpret the underlying causes of each pattern through auxiliary sensor data.

The workflow can be summarized in three main steps:

1. **Clustering:** We begin by clustering spectrogram segments to discover recurrent acoustic signatures across experiments. This serves to define meaningful ”acoustic states” without prior labeling.

2. **Classification:** Next, we train a model to predict the cluster label from the associated auxiliary signals, learning how the operating conditions or machine state might give rise to specific acoustic patterns.

3. **Explainability:** Finally, we analyze the trained classifier with Integrated Gradients[62] to obtain *per-timestep* attributions, identifying which variables and time segments most strongly drive each acoustic state, thereby shedding light on the emission mechanisms themselves.

This approach shifts the problem from prediction to interpretation. Rather than asking "what will the emission be?", we ask "what kind of emission pattern is this, and why might it occur?". This reframing allows us to extract structured insight from noisy, high-dimensional data.

3.2.1 Building the Clustering Database

To prepare the data, we first linearly interpolated the auxiliary signals to match the timestamps of the microphone signal. There is no problem with data leakage here as we are not trying to maintain causality. We then transformed the raw sound signals into spectrograms, following the methodology described in A.4, only deviating by cutting frequencies under 650 Hz instead of 500 as some testing proved it more efficient for this application. Finally, from each spectrogram, we sampled one hundred equally spaced FFT results (a time slice of width 1), along with the corresponding time segments of the auxiliary time series. The sampling process is illustrated in Figure 3.2 :

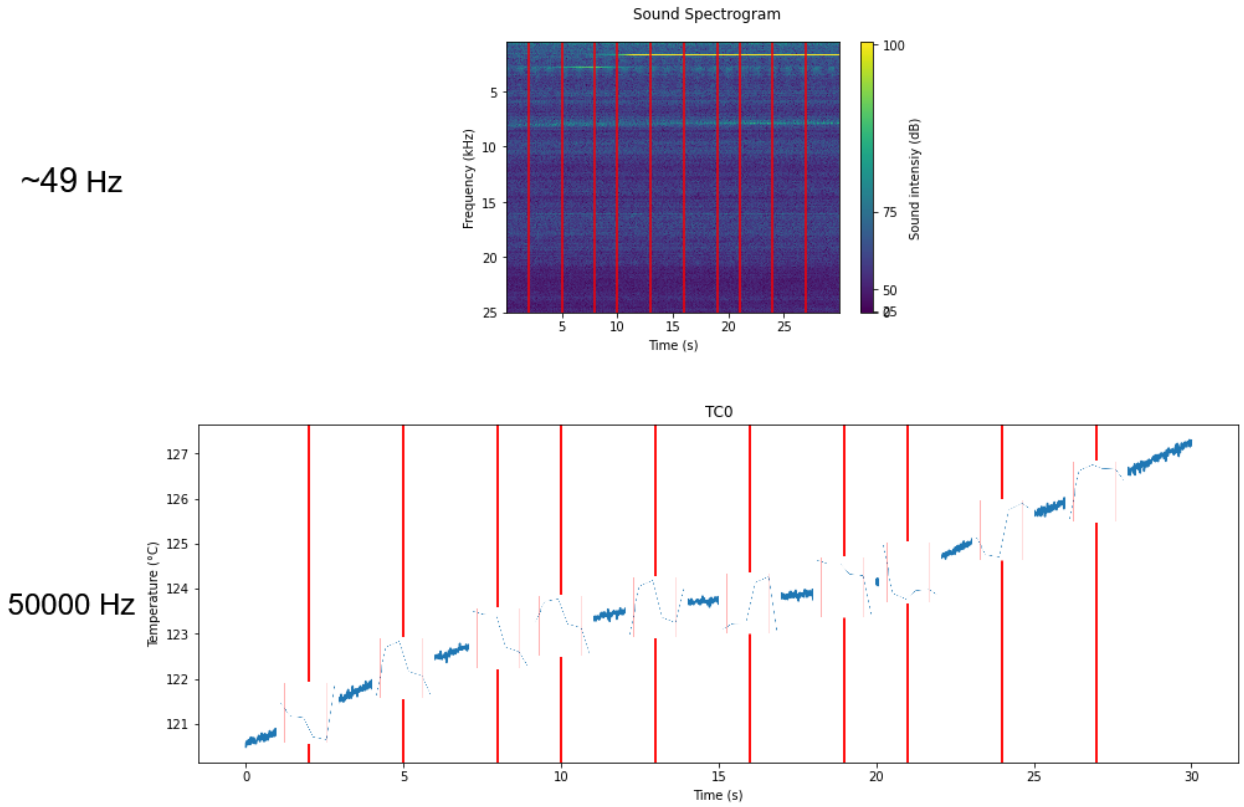


Figure 3.2: Illustration of the clustering Sampling Process. Red vertical lines signalize the sampled data.

This results in a set of 502 long frequency vectors for the spectrogram data, and a set of 1024 long time series data for each auxiliary variable. Accounting for errors and the lengths of spectrograms, that gives us a total of 50392 samples.

3.2.2 Clustering Method

One of the major challenges in clustering is determining an appropriate number of clusters—an issue that becomes even more pronounced in noisy datasets like ours. Since our data doesn't have a simple physics based number, we chose a clustering pipeline which circumvents the need for such a choice.

Dimensionality Reduction : Our clustering approach focuses on the fact that our dataset of FFTs is both noisy and high-dimensional. To tackle these challenges, we use the Uniform Manifold Approximation and Projection (UMAP [43]) algorithm. UMAP is effective not only for visualization but also for general non-linear dimensionality reduction. It captures the global manifold structure while preserving local relationships within the data, enabling it to reduce dimensionality while retaining more semantic information than traditional methods like PCA[25]. Moreover, it is significantly faster than other manifold learning techniques such as Isomap and t-SNE, especially with high-dimensional latent spaces. Figure 3.3 illustrates UMAP's ability to understand manifolds.

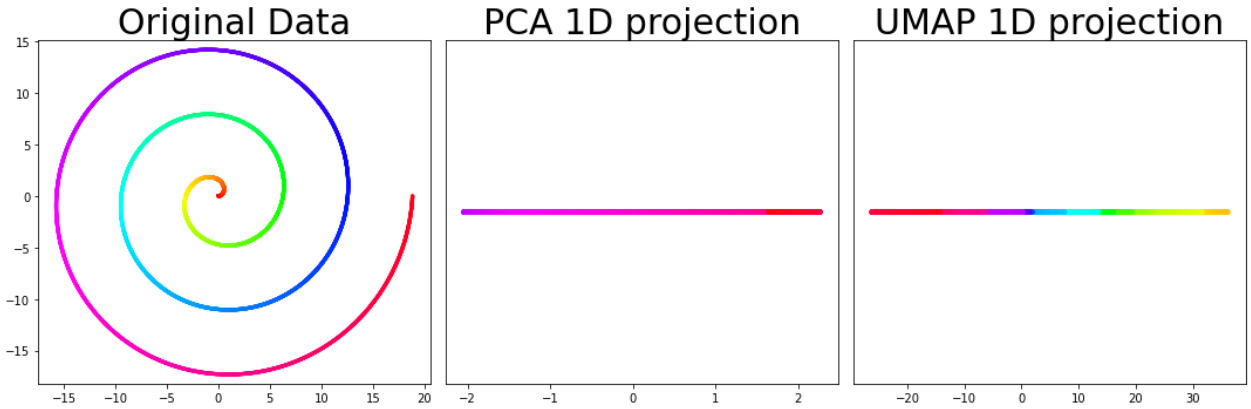


Figure 3.3: Illustration of UMAP's manifold understanding: UMAP unrolls the spiral, preserving structure, whereas PCA flattens it, mixing the data.

Clustering : After projecting the data into a lower-dimensional space, we apply the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN [42]) algorithm. Unlike methods such as K-means[39], which assume spherical clusters, HDBSCAN can handle the irregular cluster shapes produced by UMAP and is even recommanded by UMAP's original authors. It also provides robust outlier detection, which is particularly valuable for our highly noisy data. Figure 3.4 illustrates HDBSCAN's superiority over simpler methods like K-means:

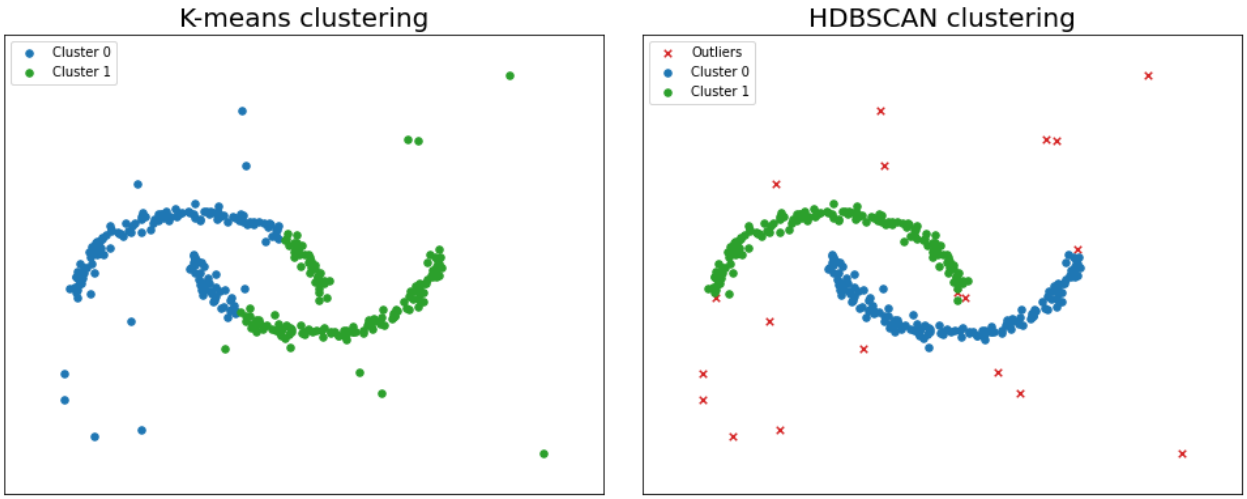


Figure 3.4: Illustration of HDBSCAN's performance: it successfully separates well-defined clusters and detects randomly added outliers, where K-means fails at both.

Implementation : To keep the pipeline scalable, we use the GPU-accelerated implementations of UMAP and HDBSCAN from RAPIDS cuML [49]. This substantially reduces runtime on large, high-dimensional datasets like ours and makes it feasible to evaluate many hyperparameter configurations during optimization.

Hyperparameter Tuning : Finally, both UMAP and HDBSCAN are sensitive to hyperparameter choices and random initialization. To ensure the robustness and consistency of our approach, we use the Optuna[2] library to optimize these hyperparameters. Our objective is to balance two competing criteria: (1) maximizing clustering stability across multiple runs (measured by the adjusted mutual information score), (2) maximising cluster separability (measured by the Silouhette score from the original tresholded space). This optimization process aims to identify a Pareto front of optimal hyperparameter sets, from which we will hand pick one. The process is illustrated in Figure 3.5.

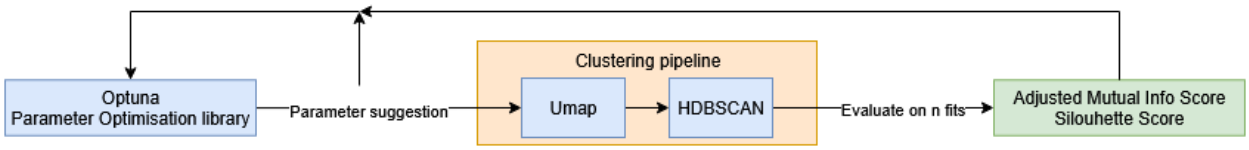


Figure 3.5: Overview of the Optuna-driven optimization loop for the UMAP + HDBSCAN clustering pipeline. Optuna[2] suggests parameters, the pipeline is evaluated using clustering metrics, and feedback is used to guide further suggestions.

Hyperparameter names and descriptions:
Both algorithms have a variety of useful hyperparameters to be tuned. We selected the following subset, which we found most relevant to clustering performance:

- **Scalers**
  - **Initial scaler**: scaling method applied to the raw FFT features before UMAP.
  - **Embedding scaler**: optional rescaling applied to the UMAP embedding before HDBSCAN.
- **UMAP (nonlinear dimensionality reduction)**

  - **Neighborhood size**: controls the balance between local and global structure in UMAP's graph.
  - **Minimum distance**: controls how tightly points can pack in the embedding; lower values yield tighter clusters.
  - **Embedding dimension**: target dimensionality of the UMAP latent space.
  - **Distance metric (input space)**: defines how similarity between FFTs is measured before embedding.
  - **Spread**: sets the overall scale of the embedding; interacts with the minimum-distance setting.
  - **Local connectivity**: guarantees at least a given number of neighbors are connected, aiding variable-density manifolds.
  - **Learning rate**: step size for the stochastic optimization that fits the embedding.
  - **Training epochs**: number of optimization passes used to learn the embedding.

- **HDBSCAN (density-based clustering)**
  - **Minimum cluster size**: smallest group considered a valid cluster; larger values favor coarser groupings.
  - **Minimum samples**: density threshold for core points; higher values are more conservative (more points labeled as noise).
  - **Cluster selection method**: rule used to extract flat clusters from the hierarchy (e.g., EOM or Leaf).
  - **Distance metric (embedded space)**: metric used by HDBSCAN on the UMAP embedding.
  - **Alpha**: controls how single-linkage is softened; affects cluster stability and resolution.
  - **Cluster-selection epsilon**: extra separation required when selecting clusters, helping split close groups.

Although additional hyperparameters could be explored, we restricted our search to these, which were sufficient to achieve strong performance on our data while keeping the search space manageable.

Summary : In summary, this pipeline—combining UMAP for dimensionality reduction with HDBSCAN for clustering and outlier detection—is tailored to uncover semantically meaningful groups within the data. Crucially, it circumvents the issue of determining an optimal number of clusters, which, as we have discussed, is particularly problematic in noisy, high-dimensional datasets.

While there are metrics that can be used to compare clustering quality, we have not performed studies to compare this pipeline to others as we have not found a metric that evaluated clustering quality well for this application.

3.2.3 Classification of Resulting Clusters

Having identified distinct clusters of acoustic responses, the next step is to develop a classification model capable of recognizing these clusters from their corresponding auxiliary time

series. Two constraints shape our choice of classifier: (i) inputs are sequential, and (ii) we must obtain reliable *per-timestep* IG [62] attributions to link auxiliary signals to acoustic states. Due to the sequential nature of these inputs, neural network architectures are a natural choice for this task.

A wide range of neural architectures has been proposed for time series classification. However, we require model that supports per-timestep attributions via Integrated Gradients. This necessitates an architecture that not only handles sequential data effectively but also allows for attribution of importance back to individual timesteps in a meaningful and interpretable way.

One-dimensional convolutional neural networks (1D CNNs) have shown strong performance in time series classification, but their reliance on pooling operations tends to dilute the attribution signal across time, making it difficult to interpret contributions of individual timesteps. Recurrent neural networks (RNNs), such as LSTMs or GRUs, do not suffer from the same attribution dilution, but they often exhibit a bias toward more recent timesteps—placing undue importance on the tail of the sequence.

In contrast, Transformer encoder architectures offer a compelling alternative. By treating each timestep as an independent token and employing self-attention mechanisms, Transformers allow for rich modeling of temporal dependencies without the sequential biases of RNNs or the attribution smearing caused by pooling layers in CNNs.

For these reasons, we adopt a Transformer-based classifier to model the clustered acoustic responses. Figure 3.6 illustrates the final architecture :

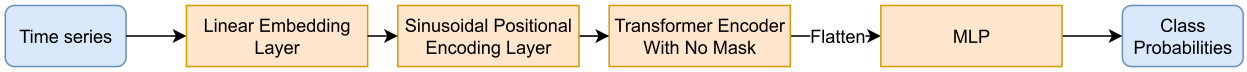


Figure 3.6: Overview of Classification Transformer model used

We were unable to identify a single set of hyperparameters that performed optimally across the tasks considered in this work; consequently, the exact architecture and number of parameters vary depending on the task. From a technical standpoint, the proposed model differs from those presented in the previous chapter in two key aspects:

- *No causal masking in the encoder :* Since causality is not a concern in this task, we do not apply a causal mask to the Transformer encoder. This allows each timestep to attend to every other timestep when computing its embedding, enabling the model to produce individually interpretable IG values for each timestep.
- *Flattened transformer embeddings before classification:* Instead of making predictions at each timestep, we flatten the Transformer embeddings and pass them to a multilayer perceptron (MLP) to produce a single prediction for the entire multivariate time series. Although this design choice can cause a rapid increase in the number of parameters due to the long sequence length, it also improves the ability to attribute contributions from individual timesteps to the final prediction.

To train our models, we used the Adam[32] optimizer. To prevent overfitting and improve generalization, we employed weight decay, dropout, and jittering of the inputs. The inputs

were normalised using a MinMax Scaler.

We observed high variability in training performance, primarily caused by random model weight initialization. To mitigate this, we implemented a restart strategy: if the validation performance did not surpass a threshold after a predefined number of epochs, the training was reset (up to a maximum number of resets). Unlike in the previous chapter, however, we could not identify a single, universally effective threshold that consistently encouraged efficient learning. Instead, the threshold was set manually for each application by observing the training process.

Finally, to achieve realistic test evaluation, we used the 6-fold cross-validation split detailed in (A.3.1). The split was determined using the method described in A.2. For each split, we used it once as the test split, selected one as the validation split, and used the remaining splits for training. This means that samples from the same test are never split across training and test sets, preventing information leakage since acoustic responses within a test are strongly correlated.

To enable meaningful interpretation of input attributions, we carefully engineered features to minimize correlation between variables. This was done in two ways:

- First, based on the correlation-based variable clustering shown in Figure 1.11, we selected a set of uncorrelated Mechanical-related variables: Rotation Speed, FY, tangential angle, and FZ.
- Second, for the temperature-related variables, we summarized the surface information by computing the barycenter of the shallow thermocouples (by assuming uniform depth). This gave us two features—Bx and By—representing the x and y coordinates of the barycenter, which evolve in time and are not strongly correlated.

Using these decorrelated features helps reduce the overlapping attribution effect that often occurs when computing attributions for correlated variables, making the results more interpretable.

3.2.4 Integrated gradient calibration

Once our models are trained, we compute input attributions using the Integrated Gradients (IG[62]) method—a widely used technique for interpreting deep learning models (see A.11 for some high level detail). A critical aspect of IG is the choice of a baseline input, which serves as a neutral reference point for attribution. While computer vision tasks often use a black (zero) or white (maximum) image, defining a "neutral" input for time series is less straightforward. For instance, a zero-valued time series can carry semantic meaning in certain domains.

To address this, we adopt a data-driven approach for selecting an appropriate baseline. We evaluate a range of candidate baselines—both conventional and unconventional—based on how well their resulting IG attributions align with Permutation Feature Importance (PFI) scores.

PFI (A.10) is a model-agnostic method that estimates feature importance by measuring the drop in model performance when each feature is randomly permuted. It does not rely on a

baseline and is less sensitive to implementation choices. Since our input features are decorrelated, typical issues like grouped importance are minimized.

IG provides fine-grained, per-timestep attributions, that we can aggregate over time on *successfull predictions* to produce a global importance score for each feature. We then compare this IG-based ranking with that from PFI using pearson correlation. The baseline that yields the closest agreement is selected.

While this calibration process offers a practical and interpretable strategy for mitigating IG's baseline sensitivity in time series applications, it comes with no theoretical guarantees and may not always yield optimal results. Nevertheless, aligning IG attributions with a model-agnostic reference like PFI provides a reasonable heuristic for selecting meaningful baselines.

3.2.5 Integrated gradient baselines

The different Integrated Gradients baselines we will evaluate are:

- **Time-dependent mean:** The mean signal across the dataset, computed for each timestep.
- **Constant mean:** A static baseline using the mean value across all timesteps and samples, repeated uniformly across time.
- **Time-dependent median:** The median signal across the dataset, computed for each timestep.
- **Constant median:** A static baseline using the median value across all timesteps and samples, applied uniformly across time.
- **Zero baseline:** A baseline consisting entirely of zeros. Since the input features are MinMax-scaled, this corresponds to the minimum observed value for each feature and represents a "minimum input" baseline.
- **Ones baseline:** A baseline filled entirely with ones. Under MinMax scaling, this represents the maximum observed value for each feature and serves as a "maximum input" baseline.
- **Midpoint baseline:** A constant baseline with all values set to 0.5. In MinMax-scaled inputs, this approximates the midpoint between minimum and maximum values, and serves as a generic "neutral" baseline.
- **Physics-driven baseline 1 and 2:** Two baseline designed using physical insight about the system (see below for details).

Physics-driven baselines 1 and 2 : In both baselines, the normal load F_Z and the rotation speed are set to the sample means (consistent with their role as controlled test inputs). The lateral force F_Y is fixed to its sample mean as a neutral, non-informed choice because of a lack of theoritical modeling. The tangential angle is obtained in both cases from the quasi-static component of the semi-analytical model of Magnier et al. [40]. The baselines differ only in how the contact barycenter is determined. *Baseline 1* assumes a uniform temperature field in the pad—in the style of Newcomb's thermal model—yielding a barycenter at the

geometric center of the pad–disc interface. *Baseline 2* also leverages Newcomb's framework but accounts for the slight depth offsets of each thermocouple: temperatures are predicted at their respective depths and the barycenter is then computed from this depth-resolved temperature distribution (see 1.2 for thermocouple depths).

Beyond their construction, these physics-driven baselines differ fundamentally from the conventional, data-driven alternatives. Rather than being derived from statistical aggregates of the dataset, they represent idealized reference states grounded in simplified physical models. Since such models are inevitably approximate and often idealistic, using them as baselines provides a meaningful way to measure how strongly the experimental signals and their attributions deviate from theoretical expectations. In this sense, the physics-driven baselines not only offer a new class of Integrated Gradients references, but also enable a direct comparison between data-driven explanations and physically motivated, "ideal" behavior.

3.3 Results

During our initial experiments, we observed an intriguing trade-off between classification accuracy and interpretability. When clustering raw spectrogram values without any preprocessing, the resulting clusters lacked clear semantic meaning. However, classifiers trained on these clusters achieved high accuracy. In contrast, applying a decibel threshold to retain only high-intensity acoustic responses yielded more interpretable clusters, yet led to a substantial drop in classification performance.

This suggested the presence of a background signal or noise component that, while aiding classification, might obscure meaningful structure in the data. One hypothesis is that this background component encodes information related to the experimental setup/protocol rather than the underlying behaviors of interest.

To investigate this hypothesis, we leveraged experimental knowledge: tests are performed in sets, and parameters within each set are typically similar. We reasoned that if this background signal varies consistently across test sets, a model should be able to exploit it to predict from which test set a given acoustic response originates.

While a full clustering-and-classification degradation analysis across a range of thresholds would be ideal, it is computationally intensive. As a more efficient alternative, we employed a simplified diagnostic approach. For a set of predefined decibel thresholds, we trained an XGBoost classifier to predict the test set identity directly from the thresholded acoustic responses. Evaluation was performed using a 5-way Grouped Cross-Validation (i.e., a cross-validation scheme ensuring that each test set is represented as fairly as possible across training and validation folds).

This task serves as a proxy to assess the amount of structured signal—possibly originating from background noise—preserved at each threshold. High predictive performance at a given threshold indicates the presence of such structure, whereas performance drops suggest information loss due to excessive filtering.

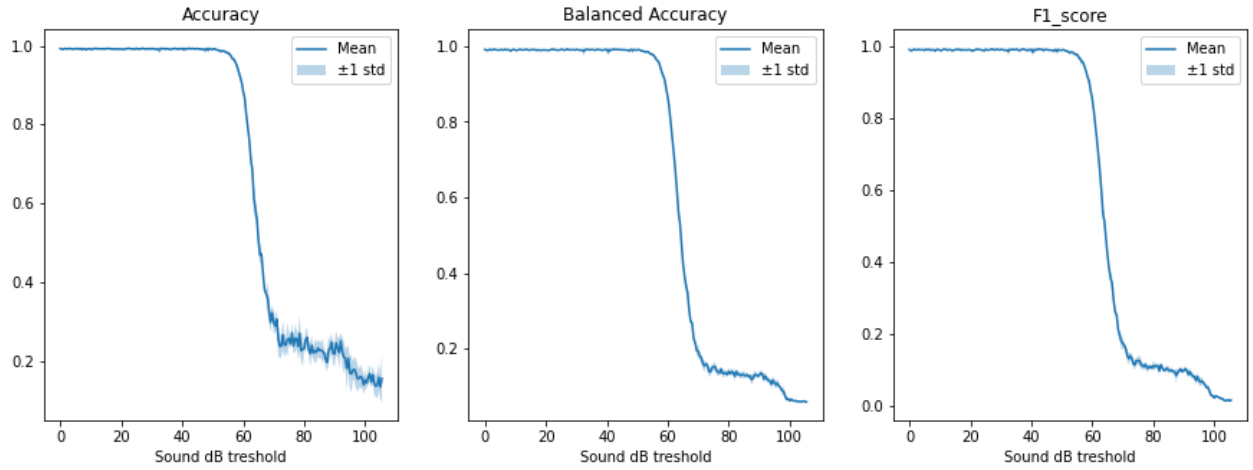


Figure 3.7: XGBoost performance metrics (mean ± 1 std) for predicting test set identity across different acoustic decibel thresholds.

As shown in Figure 3.7, all evaluation metrics—accuracy, balanced accuracy, and F1-score—remain high and stable up to approximately 55 dB. Beyond this threshold, performance drops sharply and eventually stabilizes at low values after 75 dB. This behavior supports our hypothesis that the background component contains set-identifying information and is gradually filtered out as the threshold increases.

Whether this background signal reflects unintended experimental artifacts or consistent similarities in acoustic responses within each test set is difficult to determine. Nonetheless, its impact on classification performance and interpretability is significant.

To further investigate this trade-off, we selected two representative thresholds for detailed analysis: 55 dB, where predictive information from the background signal is still present, and 75 dB, where this signal is largely removed. The subsequent sections will examine the clustering, classification, and interpretability pipeline at both thresholds.

3.3.1 Low Threshold (55 dB): Preserved Background Signal

In this subsection, we examine the acoustic-response results obtained with a 55 dB threshold, a setting in which residual background noise remains perceptible.

3.3.1.1 Clustering Results

We begin by evaluating the clustering outcomes, starting with the results generated from the Optuna study which are illustrated in Figure 3.8.

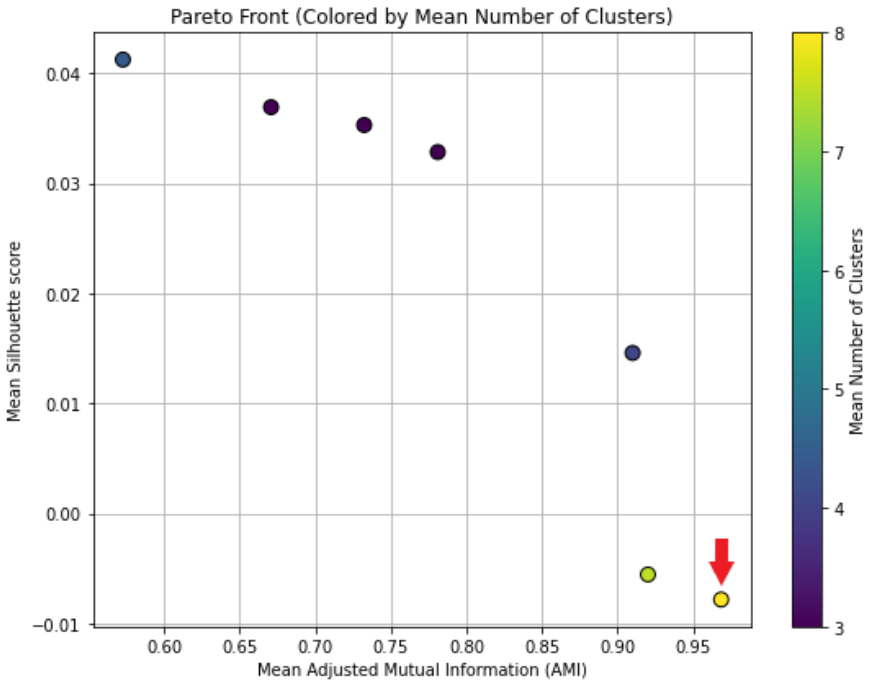


Figure 3.8: Resulting Pareto front from the Optuna trials for clustering of 55 dB thresholded acoustic responses. The selected point is indicated by the red arrow.

The selected point, shown above, was chosen because all trials on the Pareto front yielded poor silhouette scores (ranging from -0.01 to 0.04). Selecting 0.04 over -0.01 offers negligible improvement in clustering quality. Therefore, we opted for the most stable clustering configuration—represented by the rightmost point—which corresponds to the following hyperparameters:

Metrics					
Mean AMI	0.9685	Mean Silhouette	-0.0078	Mean Number of Clusters	8
Scalers					
Initial scaler	maxabs	Embedding scaler	none		
UMAP (nonlinear dimensionality reduction)					
Neighborhood size	34	Minimum distance	0.4884	Embedding dimension	4
Distance metric (input space)	correlation	Spread	1.5474	Local connectivity	2
Learning rate	0.3759	Training epochs	4090		
HDBSCAN (density-based clustering)					
Minimum cluster size	106	Minimum samples	78	Cluster selection method	eom
Distance metric (embedded space)	euclidean	Alpha	1.9843	Cluster-selection epsilon	0.1600

Table 3.1: Selected configuration for the clustering of acoustic responses at a 55 dB threshold

Selecting the setup with the highest mean AMI resulted in a configuration where the number of clusters remained constant across runs, which is desirable for consistency. Typically, this number fluctuates due to the inherent randomness of the algorithms—a variability further amplified by their GPU implementations. The embedded space, colored by the cluster IDs, is shown in Figure 3.9:

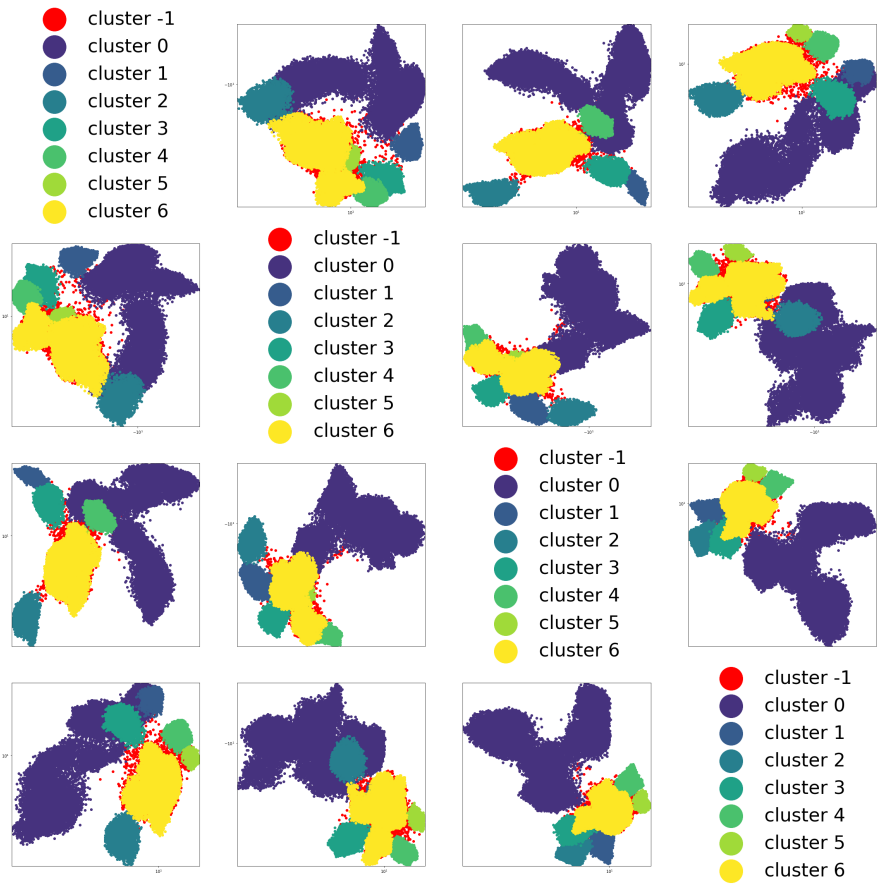
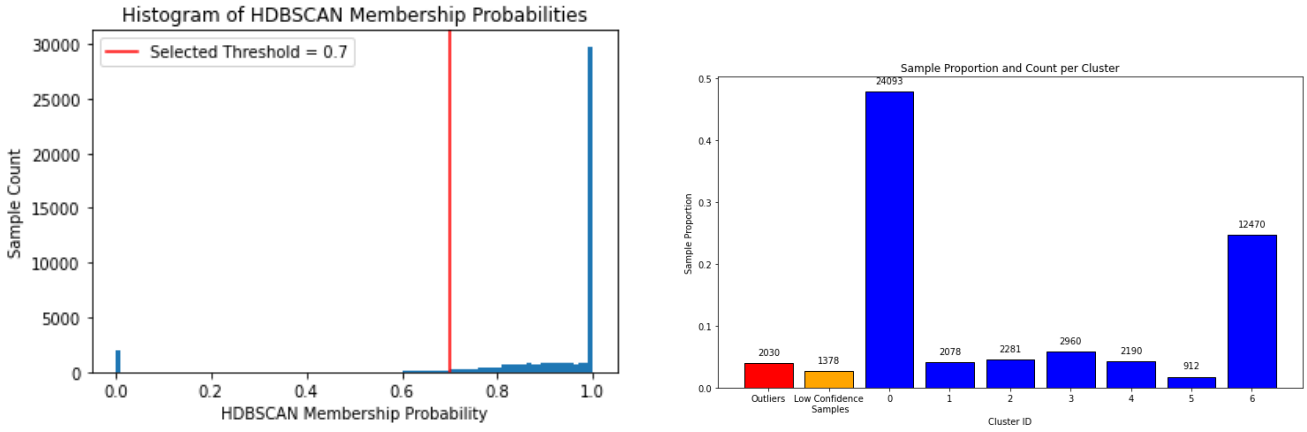


Figure 3.9: Embedded space colored by cluster ID. -1 cluster contains the outliers detected by HDBSCAN

The embedded space does not exhibit strong cluster separation, which, in our experience, appears to be a limitation of the GPU version of UMAP. Nevertheless, the separation is sufficient to produce meaningful clusters. That said, there is a reasonable argument that the CPU version could yield better clustering results, albeit at the cost of significantly longer computation times. Next, we examine cluster assignment and confidence, based on HDBSCAN membership probabilities and label distributions, as illustrated in Figure 3.10.



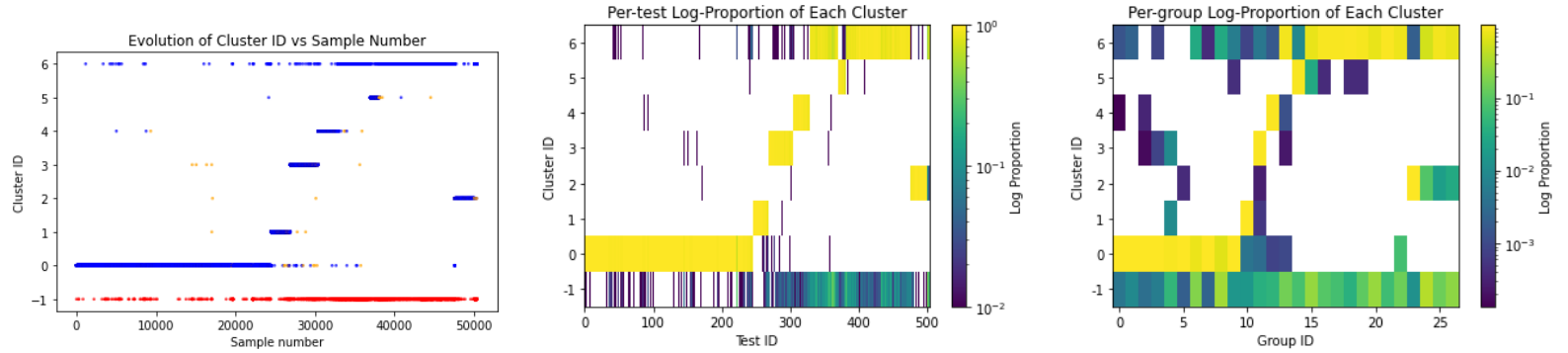
(a) HDBSCAN membership probabilities histogram.

(b) Sample proportion and count per cluster.

Figure 3.10: Clustering confidence and distribution.

The histogram of HDBSCAN membership probabilities reveals a large proportion of highly confident points, followed by a gradual decline down to approximately 0.7, beyond which it levels off with relatively few points. We selected a cutoff at 0.7 to remove less informative points that might otherwise skew the training process. The resulting histogram shows that, while clusters 0 and 6 occur far more frequently than others, nearly all remaining clusters still contain at least one thousand samples, which is encouraging for training stability. Furthermore, only a small fraction of points (outliers and low-confidence cases) are excluded from training—an acceptable reduction.

Finally, we analyse the evolution of cluster assignments across different ordering schemes to visualise potential patterns. Specifically, we examine their progression through ordered samples, ordered tests, and ordered test groups, as illustrated in Figure 3.11.



(a) Cluster ID through time-ordered samples.

(b) Histograms of cluster ID distribution per ordered test.

(c) Histograms of cluster ID distribution per ordered test group.

Figure 3.11: Evolution of clustering through time, tests, and test groups.

From graph (a), we observe that clusters are not uniformly distributed over time: cluster 0 comprises the vast majority of samples during the first half of the experimental protocol. After approximately the halfway point, other clusters appear more frequently—most notably cluster 6, which, together with cluster 0, forms the most common clusters overall. This suggests a shift in underlying behavior around the midpoint of the protocol, with cluster 0 dominating before and cluster 6 prevailing afterward, along with increased deviations to other clusters.
Graphs (b) and (c) display a similar overall pattern, but also reveal that clusters 1, 2, 3, 4, and 5 originate primarily from a small subset of tests or test groups. This supports our earlier observation that background noise specific to certain test groups may be embedded in the recordings when thresholding the acoustic responses at 55 dB or lower, suggesting that these clusters may, in fact, be capturing artefacts from differences in background noise rather than genuine variations in mechanical behavior.

3.3.1.2 Cluster analysis

Having identified the cluster structure at the 55 dB threshold, we now investigate the characteristics of each cluster to assess their distinctiveness and potential physical meaning. We begin by observing the acoustic responses grouped by cluster, illustrated in Figure 3.12:

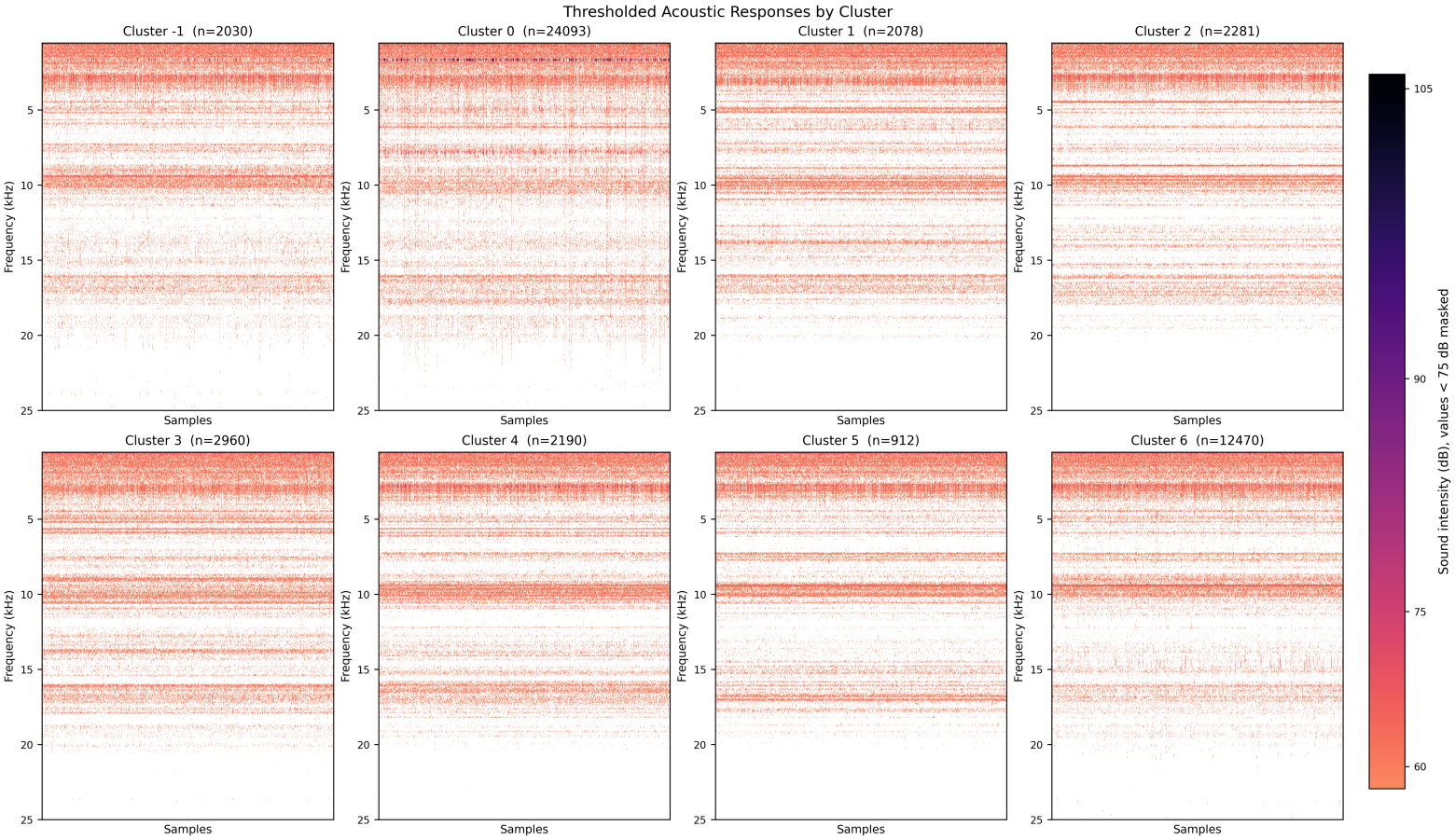


Figure 3.12: Thresholded acoustic responses grouped by cluster. Values below the 55 dB threshold are masked (white regions). Each subplot corresponds to one cluster, with samples along the horizontal axis and frequency (kHz) along the vertical axis.

Elements within each cluster display notable internal similarity, and clear differences can be observed across clusters. Cluster 0 is the only one showing clear high-intensity events around 1.5kHz, resembling the squeal events identified previously, whereas others—such as cluster 6—exhibit broader spectral content concentrated at higher frequencies. Some clusters share partially overlapping patterns, yet still maintain distinctive overall shape.

Interestingly, the noise cluster (-1), which should in principle consist only of outliers, also appears to contain a coherent structure, closely resembling that of cluster 6. While closer inspection reveals some dissimilar samples within this group, the majority show strong similarity to cluster 6. Whether this arises from imperfect UMAP embedding or from HDBSCAN's internal decision process is unclear; however, it does not undermine the downstream interpretation of the recognised clusters.

Overall, these observations suggest that the clustering captures recurring spectral signatures in the acoustic responses. In the following, we investigate whether these spectral differences are also reflected in the auxiliary signals.

We start with the Thermal-related variable histograms (Figure 3.13).

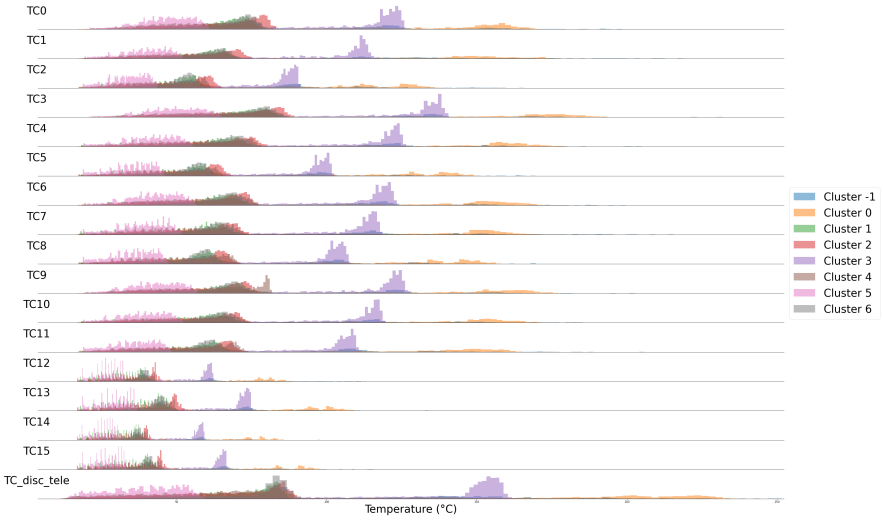


Figure 3.13: Thermal Variable Histograms colored by Acoustic Cluster.

The clusters exhibit clear thermal separation. Cluster 0 consistently records the highest temperatures, followed by cluster 3. At the opposite end, cluster 5 is associated with the lowest temperatures. The remaining clusters display overlapping distributions without a distinct ordering. This suggests that certain clusters occur predominantly under either high- or low-temperature conditions, indicating a potential link between acoustic behaviour and thermal state. We continue with the Mechanical-related variable histograms (Figure 3.14).

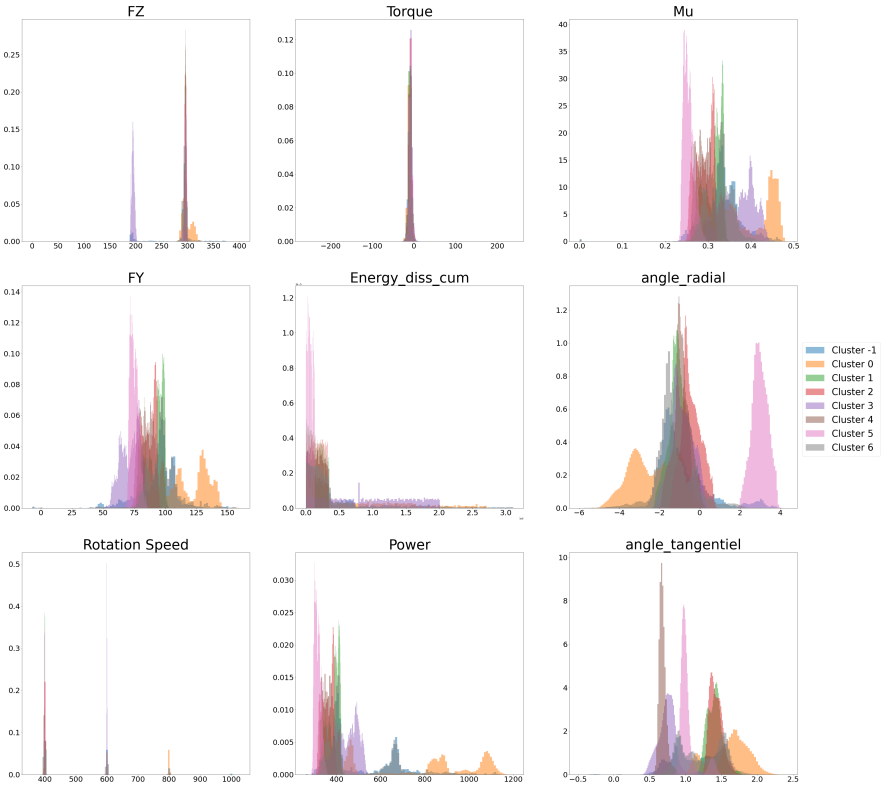
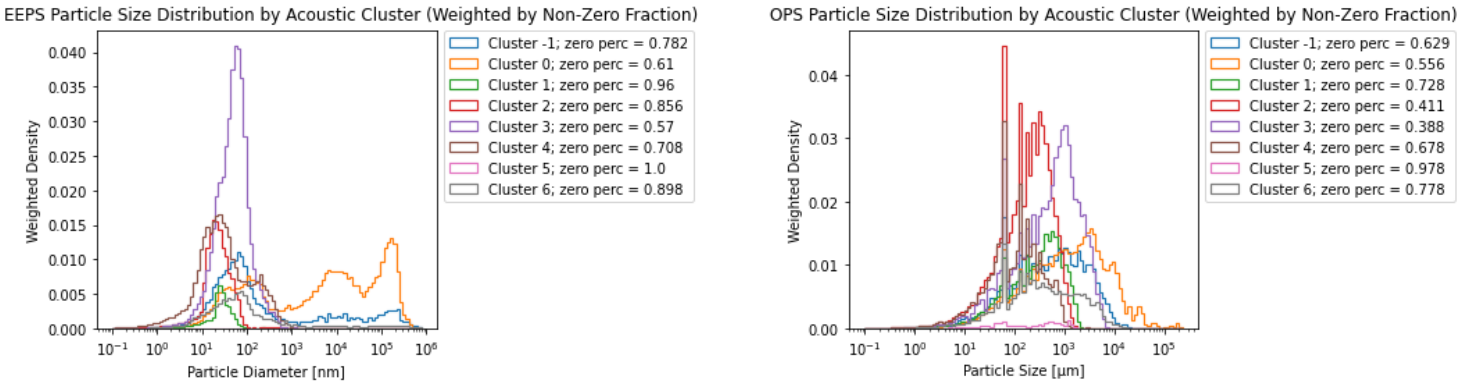


Figure 3.14: Mechanical Variable Histograms colored by Acoustic Cluster.

The clusters again reveal markedly different patterns. Cluster 0 is notable for its consistently extreme values across nearly all variables, reflecting operating conditions well outside the range of most other clusters. Cluster 5 is distinguished by a persistently high radial angle. Both the normal load (F_Z) and rotation speed—although nominally constant input parameters—are in practice recorded as time signals, where the nominal setpoint is perturbed by the braking process. These parameters emerge as discriminative variables for certain clusters; for example, Cluster 3 is clearly separated from the others in F_Z, while rotation speed also varies systematically across clusters. The fact that these input parameters are strongly discriminative supports the interpretation that the clustering may in part be capturing background noise patterns linked to specific test groups rather than purely differences in mechanical behaviour.

We now turn to particle emissions, whose distributions across acoustic clusters are shown in Figure 3.15.



(a) EEPS Distribution by cluster

(b) OPS Distribution by cluster

Figure 3.15: Flattened EEPS and OPS particle amount distributions by acoustic cluster. All size bins were combined into a single distribution.The Densities weighted by the proportion of non-zero readings to allow for better comparison.

Particle emissions exhibit less distinct clustering patterns. For both EEPS and OPS distributions, Cluster 0 accounts for nearly all instances of exceptionally high emissions, while the remaining clusters are generally centred around comparatively low emission levels. This once again seems theoretically backed as squeal events and particle emissions have been linked in the litterature.

A final observation concerns the contact location, estimated via a proxy measure. For each acoustic response, the temperature deltas from the shallow thermocouples are computed, and their barycentre is then determined. This provides a *naive* approximation of where the majority of contact occurred during the acoustic event. The resulting two-dimensional barycentre positions can then be visualised as histograms for each cluster in Figure 3.16.

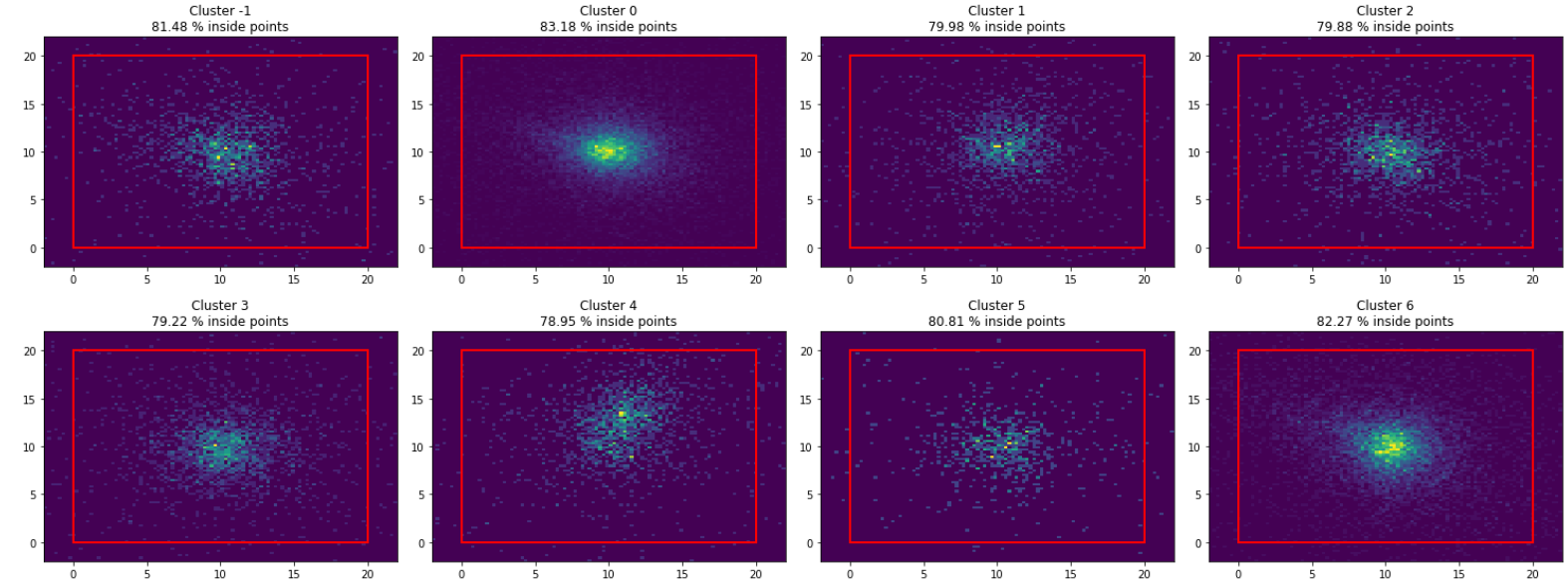


Figure 3.16: Surface temperature thermocouple time delta barycentres. Red rectangle represents the border of the pin. The percentage of inside point is the percentage of computed barycentres that is within the pin. The samples outside the pin can be explained by errors made during the capture of the data.

While some clusters seem to have a visibly different barycentre distribution like cluster 4 which is higher up than the others, it is hard to see the differences between the others. For this reason, we applied a method that approximates a 2D Kolmogorov–Smirnov test on non normal values via bootstrapping [46]. The test was conducted only on barycentres falling within the pin boundaries. This was done using the public code NDTEST [1]. The resulting p-values are show in Figure 3.17 :

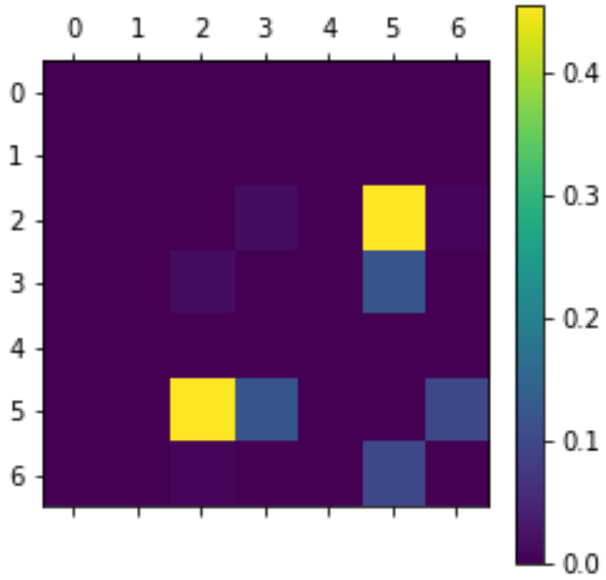


Figure 3.17: 2D Kolmogorov–Smirnov approximation test p-values by cluster ID.

[1] Written by Zhaozhou Li, `https://github.com/syrte/ndtest`

Most cluster pairs show a distinct difference in the thermal delta barycentre distribution across tests. At the 95% confidence level, only the pairs (5,2), (5,3), and (5,6) cannot be said to have different distributions. This may be due to Cluster 5 being the least populated, and therefore lacking sufficient data to reveal a clearly discriminative distribution, or it may simply represent a non-discriminative cluster whose contact location resembles that of others. In any case, these results indicate that our proxy for contact location differs between most clusters.

In summary, the cluster analysis reveals that the acoustic-based grouping is reflected across multiple domains: thermal behaviour, mechanical variables, particle emissions, and, in most cases, proxy contact location. This cross-domain consistency suggests that the clustering may capture meaningful mechanisms rather than purely random variation but it may also reflect test-group-specific artefacts. In the next section, we explore whether these clusters can be predicted directly from auxiliary signals to further try to explain them.

3.3.1.3 Classification

Building on the clustering results and analysis, we next examine the classification stage. Before discussing the performance metrics, we first provide a summary of the model architecture used and it's parameter counts (Figure 3.18).

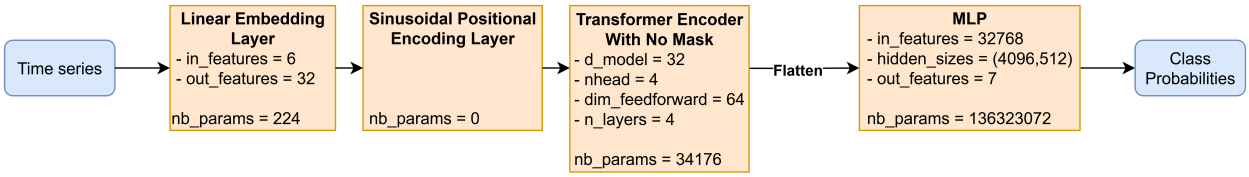
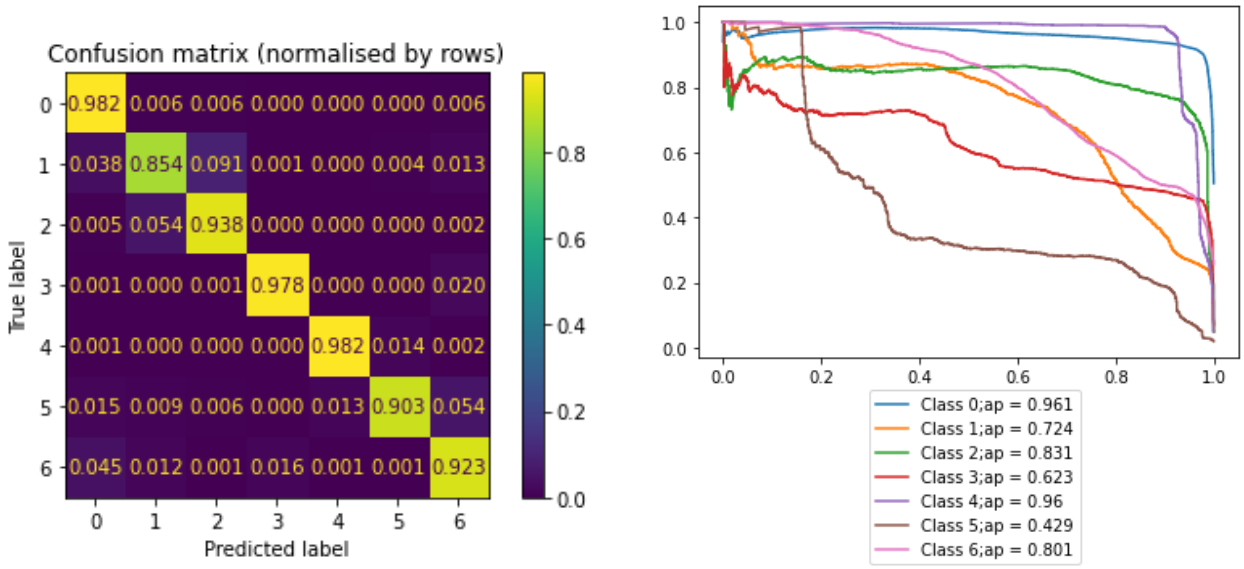


Figure 3.18: Overview of Classification Transformer model used for the 55 dB threshold

This architecture results in a total of 136,357,472 parameters, the overwhelming majority of which lie in the MLP component, as discussed earlier. While this number is not unusually large by modern deep learning standards, it is considerable given the modest size and limited diversity of the database. Moreover, the fact that no smaller configuration achieved comparable performance again suggests that the model is likely overfitting to the experimental protocol rather than capturing meaningful mechanical phenomena.

With this limitation in mind, we now turn to the evaluation of the classification stage. The results are presented in the form of the confusion matrix, one-vs-rest precision–recall curves (see Figure 3.19), and the full classification report (see Table 3.2):



(a) Classifier Confusion Matrix (normalised by True labels)

(b) OvR Precision Recall Curve with Average Precision

Figure 3.19: Classification results

Class	Precision	Recall	F1-score	Support
0	0.97	0.98	0.98	22698
1	0.83	0.85	0.84	2198
2	0.85	0.94	0.89	2186
3	0.94	0.98	0.96	3166
4	0.99	0.98	0.98	2233
5	0.94	0.90	0.92	890
6	0.97	0.92	0.95	11574
Accuracy			0.96	44945
Balanced accuracy			0.94	44945
Macro avg	0.93	0.94	0.93	44945
Weighted avg	0.96	0.96	0.96	44945

Table 3.2: Classification report

The confusion matrix reveals that the most challenging class to predict is cluster 1, with an accuracy of 85.4%. All other clusters exceed 90% accuracy, with several surpassing 95%. This pattern is consistent in the classification report, where cluster 1 also has the lowest F1-score (0.84), while the remaining clusters achieve substantially higher values, in some cases above 0.95.

The OvR Precision–Recall curves provide additional insight into model confidence. While the model attains high average precision (AP) for clusters 0, 3, and 4 (above 0.96), it exhibits markedly lower AP for clusters 1, 2, and 5, with cluster 5 showing the lowest AP (0.429). This discrepancy suggests that although the model often predicts these classes correctly, it does so with relatively low confidence, potentially reducing the sharpness of its predictions and impacting the quality of the resulting IG attributions.

Another observation is that the balanced accuracy (0.94) is close to the raw accuracy (0.96),

indicating that the classifier maintains strong performance even when accounting for class imbalance. Nevertheless, this metric may, in fact, be a more reliable indicator of real-world performance for this dataset, as it better reflects the model's ability to handle minority classes without being dominated by majority-class accuracy.

Taken together, these results indicate that the classification model performs well overall, with a weighted average F1-score of 0.96 and competitive balanced accuracy. However, specific clusters—particularly 1, 2, and 5—remain harder to distinguish with high certainty, likely due to overlapping feature distributions or similarities in background noise patterns. Addressing these weaknesses could involve targeted data augmentation, improved training strategies, or even modifications to the experimental protocol to ensure more frequent representation of the less prevalent classes.

3.3.1.4 IG Calibration

Now that the classifier has been trained and shown to perform reliably, we proceed with the explainability stage, focusing on Integrated Gradients (IG) baseline calibration. As outlined previously, this process begins with the computation of Permutation Feature Importance (PFI), which we evaluate here using balanced accuracy rather than raw accuracy, since it provides a more appropriate measure in our classification setting with highly imbalanced class distributions. The results are as shown in Figure 3.20 :

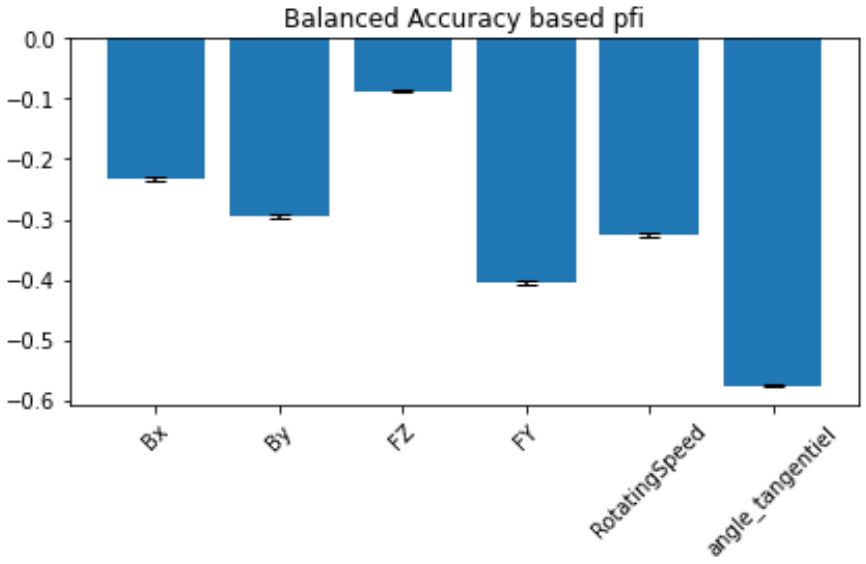


Figure 3.20: Balanced accuracy based PFI for the classification model at 55 dB threshold

Based on these scores, the next step is to identify the most suitable IG baseline. To do so, we calculate the absolute aggregated attributions for each variable under the different baseline candidates and then assess their alignment with the PFI scores. The results are reported in Table 3.3.

Method\Variable	B_x	B_y	F_z	F_y	Rotation speed	Tangential angle	Pearson r
PFI	-0.23 (5)	-0.30 (4)	-0.09 (6)	-0.40 (2)	-0.33 (3)	-0.58 (1)	
Time-dependent mean	3.36×10^{-3} (4)	3.33×10^{-3} (5)	9.86×10^{-4} (6)	4.92×10^{-3} (3)	7.20×10^{-3} (1)	5.20×10^{-3} (2)	-0.68
Constant mean	3.55×10^{-3} (5)	3.91×10^{-3} (3)	3.29×10^{-3} (6)	3.66×10^{-3} (4)	8.23×10^{-3} (1)	4.89×10^{-3} (2)	-0.28
Time-dependent median	3.73×10^{-3} (5)	4.36×10^{-3} (4)	1.06×10^{-3} (6)	6.05×10^{-3} (1)	5.57×10^{-3} (3)	6.00×10^{-3} (2)	**-0.89**
Constant median	4.43×10^{-3} (2)	3.08×10^{-3} (5)	2.21×10^{-3} (6)	3.22×10^{-3} (4)	8.51×10^{-3} (1)	4.09×10^{-3} (3)	-0.21
Zeros	9.10×10^{-3} (4)	1.00×10^{-2} (2)	8.81×10^{-3} (5)	1.00×10^{-2} (3)	5.18×10^{-3} (6)	2.00×10^{-2} (1)	-0.69
Ones	1.00×10^{-2} (1)	5.80×10^{-3} (5)	4.57×10^{-3} (6)	8.28×10^{-3} (4)	1.00×10^{-2} (2)	9.09×10^{-3} (3)	-0.31
Midpoint	3.04×10^{-3} (6)	5.51×10^{-3} (4)	5.15×10^{-3} (5)	5.56×10^{-3} (3)	7.47×10^{-3} (1)	6.92×10^{-3} (2)	-0.53
Physical baseline 1	1.00×10^{-2} (2)	2.60×10^{-3} (3)	6.34×10^{-5} (5)	2.85×10^{-4} (4)	2.78×10^{-5} (6)	2.00×10^{-2} (1)	-0.61
Physical baseline 2	8.13×10^{-3} (2)	4.35×10^{-3} (3)	7.06×10^{-5} (5)	2.95×10^{-4} (4)	3.01×10^{-5} (6)	2.00×10^{-2} (1)	-0.66

Table 3.3: PFI scores and absolute IG aggregated values per variable, along with the Pearson r between PFI and IG for each baseline attribution. Value in parantheses is the rank of each variable according to the method. Negative Pearson r values arise because absolute IG increases with feature importance, whereas PFI is defined as a performance drop and therefore decreases with importance.

According to the Pearson r correlation, the best-performing baseline is the time-dependent median. This choice is well supported in the context of time-series IG attribution, as it behaves similarly to the time-dependent mean while offering greater robustness. The global attributions obtained with PFI and time-dependant median IG are compared Figure 3.21.

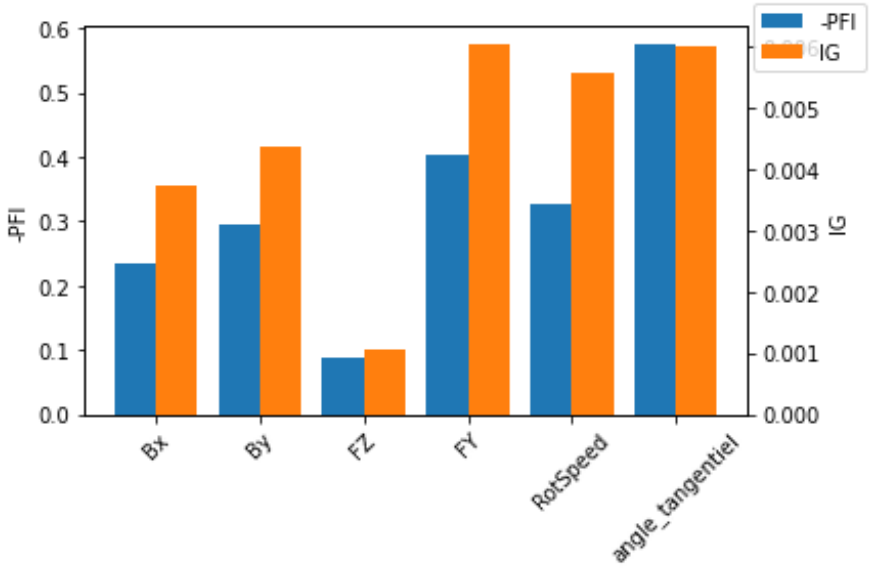
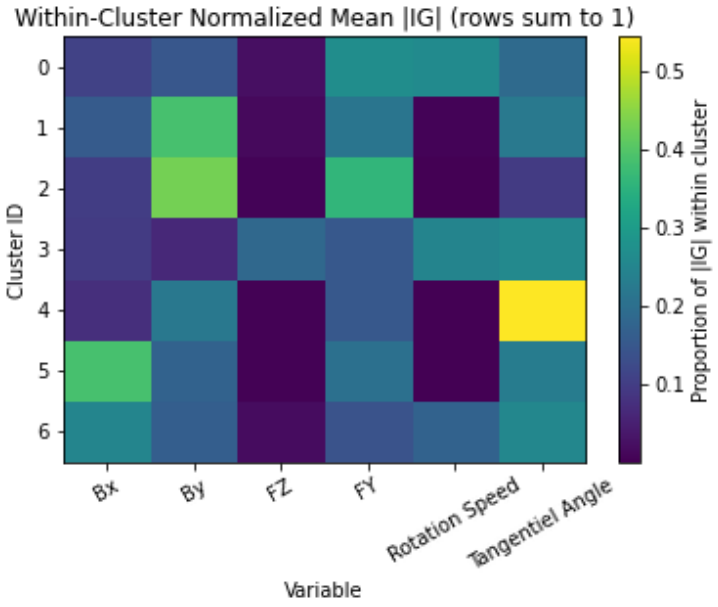


Figure 3.21: PFI vs Time-dependant median guided IG global scores per variables

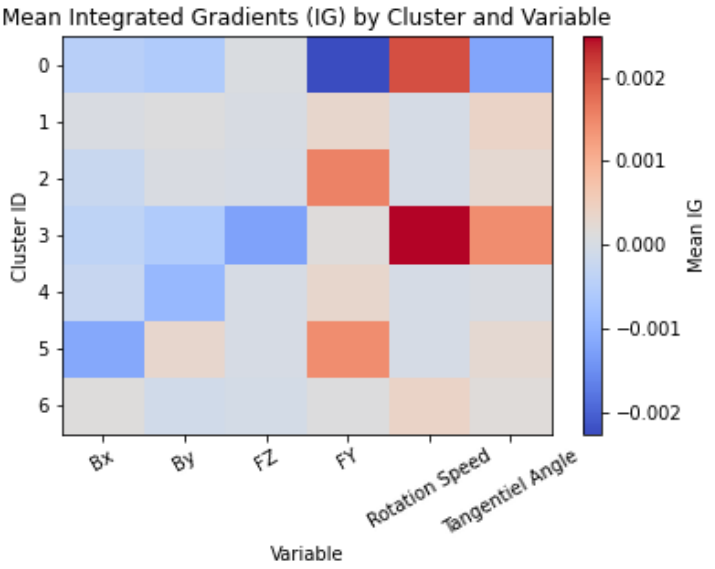
The comparison between PFI and IG attributions indicates a strong overall consistency: variables deemed important by PFI also receive large IG scores, while less influential variables consistently obtain lower values. Although the exact ranking differs across the two methods—as can be seen in 3.3—this broad agreement confirms that the chosen IG baseline produces attributions broadly aligned with the model-agnostic reference. At the same time, the discrepancies in variable ordering highlight a limitation of this calibration strategy, suggesting that while Pearson r provides a useful heuristic for baseline selection, there is ultimately no guarantee that it yields fully correct or faithful attributions.

3.3.1.5 Interpretation

With the baseline established, we can now leverage IG to interpret the model's decision process. To this end, we first examine both the absolute and signed mean attributions of each variable across clusters, as illustrated in Figure 3.22.



(a) Within-cluster normalized mean absolute IG values, reflecting the relative importance of variables inside each cluster.

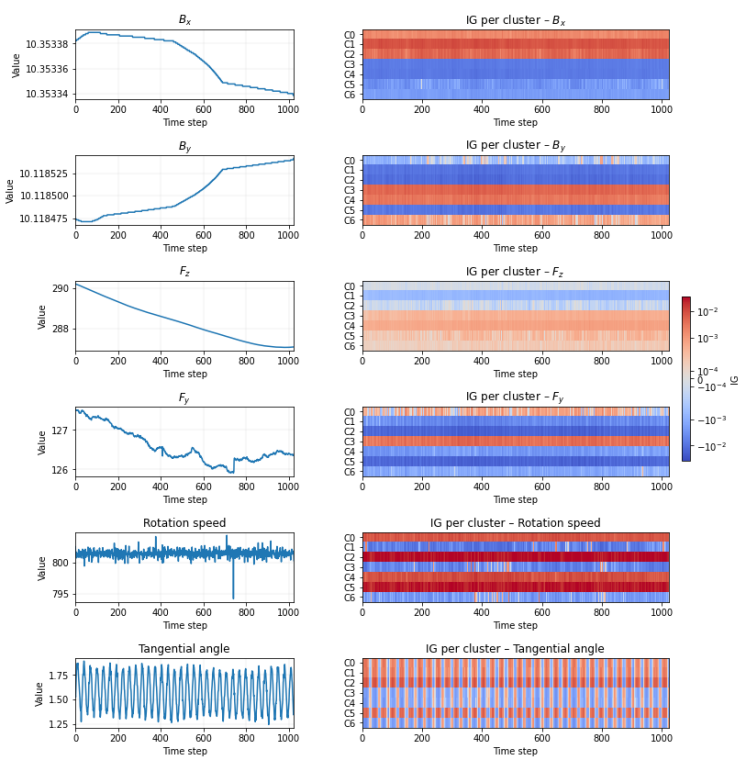


(b) Mean signed IG values, indicating whether variables push the prediction towards or away from each cluster in general.

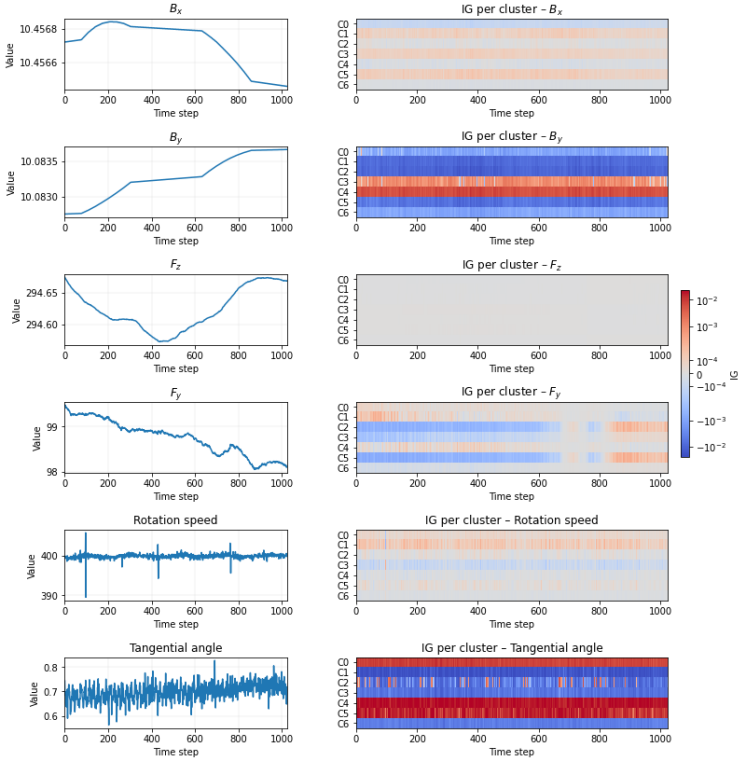
Figure 3.22: Cluster-level attribution analysis. (a) Relative importance of variables per cluster based on normalized mean absolute IG. (b) Directional effect of variables based on mean signed IG.

The two graphs serve complementary purposes and should be interpreted in conjunction. Graph (a) highlights which variables each cluster is predicted by most strongly. For example, in cluster 4 the most decisive variable is the tangential angle. Graph (b), on the other hand, provides the average direction of effect for each variable within each cluster. Continuing with the same example, the tangential angle for cluster 4 generally shows a negative attribution, meaning that larger values decrease the likelihood of this cluster being predicted. A similar pattern is observed for F_y in cluster 2: it emerges as the second most important variable overall in graph (a), while graph (b) reveals a clearly positive effect—higher values of F_y increase the probability of assigning samples to cluster 2. Finally, the fact that some high-importance cluster–variable pairs exhibit mean IG values close to zero suggests non-linear behavior in the model. In these cases, the variable strongly influences the prediction but does so in opposite directions depending on the local context, leading to positive and negative contributions that cancel out on average.

While the preceding analysis has focused on global and cluster-level attributions, IG can also be applied at the level of individual samples. This allows us to move from aggregate patterns to instance-specific explanations. To demonstrate this, we present two examples in Figure 3.23:



(a) Instance-level IG visualization for sample 961.



(b) Instance-level IG visualization for sample 6453.

Figure 3.23: Inputs (left of each panel) and signed IG per cluster over time (right of each panel) for two contrasting samples. A shared symmetric–log color scale is used across variables and clusters.

The second example shows that the prediction relies almost entirely on a small subset of variables. This indicates that the classifier can use features in a context-dependent manner, focusing on different variables for different samples rather than applying a uniform decision rule across the dataset.

In the first example, the attribution for the tangential angle alternates between positive and negative values, closely tracking oscillations in the raw input signal. In the second example, the attribution for F_y also changes sign midway through the trajectory, again in line with variations in the raw values. These patterns suggest that the classifier's attributions are often driven directly by raw magnitudes rather than by intricate temporal dynamics.

It is worth noting, however, that while these observations point towards reliance on relatively simple signals, simpler tabular models such as XGBoost trained on summary statistics of the time series (minimum, mean, maximum, etc.) did not achieve comparable performance. This suggests that the classifier might be capturing subtler dependencies that we have not managed to observe in the IG attributions, possibly because of limitations introduced by the chosen baseline.

Overall, this subsection provides useful intuition about global and instance level attributions, but the results should be interpreted with caution. Integrated Gradients itself is a principled attribution method, but our procedure for baseline selection is heuristic, and the quality of the explanations is therefore contingent on this choice.

3.3.1.6 Conclusion

To conclude the analysis at the 55 dB threshold, we summarize the main findings.
The Optuna study produced very stable results and yielded eight clusters that, while not strongly supported by silhouette scores, were nevertheless visibly differentiable. The background noise appears to influence the grouping, with clusters often aligning with test groups, as anticipated from prior analysis.

Despite this, cluster analysis revealed meaningfully different distributions of key properties across clusters, though it remains difficult to determine whether these differences arise from the experimental protocol, background noise, or genuine mechanical phenomena.

Classification on the clustered data performed reliably, with balanced accuracy close to raw accuracy, indicating that uneven cluster sizes did not undermine performance, even if some clusters were predicted less accurately than others.

PFI-based calibration identified the time-dependent median as the most suitable IG baseline, a classical and robust choice. With this baseline, the interpretation stage provided insights into the main factors driving each cluster and their directional influence, while instance-level examples suggested that predictions often relied on raw variable magnitudes rather than temporal patterns—raising the possibility of overfitting to background noise.

Taken together, these results show that the proposed pipeline can provide meaningful insights at the 55 dB threshold, while also highlighting limitations related to background noise and baseline dependence.
The whole pipeline took $\sim$ 35 Hours of compute time in total. See Appendix A.1 for details on the software, hardware, and computational resources used.

3.3.2 High Threshold (75 dB): Reduced Background Influence

After studying our method's effectiveness at the 55 dB threshold, we now turn to the 75 dB case in order to further assess robustness and reduce the influence of background noise.

3.3.2.1 Clustering Results

We once again begin by looking at the clustering results, starting from the Optuna optimisation Pareto front illustrated in Figure 3.24.

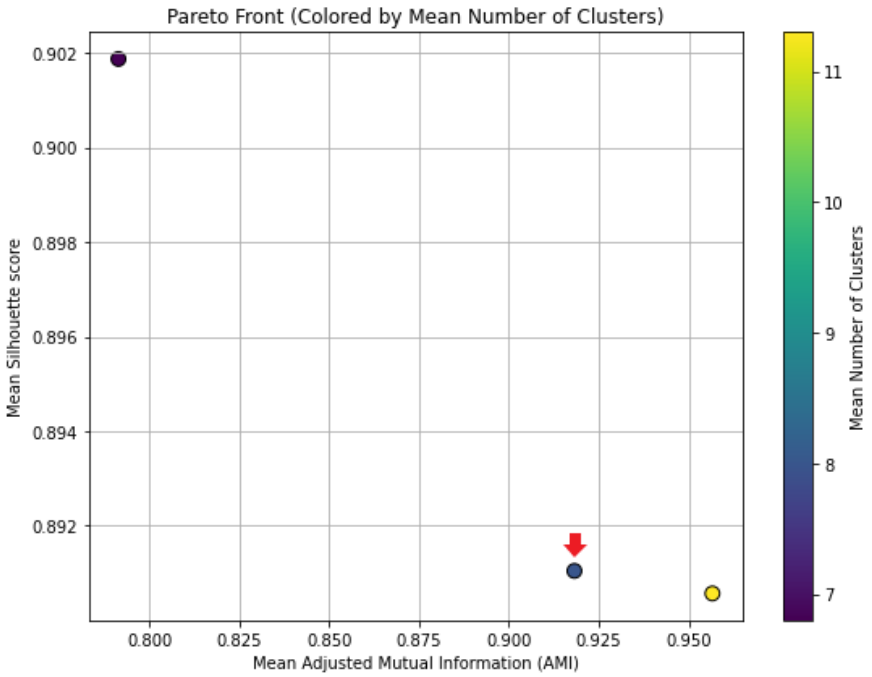


Figure 3.24: Resulting Pareto front from the Optuna trials for clustering of 75 dB thresholded acoustic responses. The selected point is indicated by the red arrow.

Our rationale for choosing the selected point differs from the 55 dB case. At 75 dB, both the Silhouette values and AMI scores are generally higher. The leftmost point, however, yields an unacceptably low AMI score and is therefore excluded. Between the two remaining candidates, a conventional interpretation would favour the rightmost point, since an increase of nearly 0.05 in AMI comes at the cost of less than 0.001 in Silhouette score. Nevertheless, our visual inspections revealed that the resulting clusters from this configuration were difficult to distinguish by eye. To prioritise interpretability and separability, we instead selected the middle point, which produces fewer but more distinct clusters. The corresponding hyperparameters are as follows:

Metrics					
Mean AMI	0.9182	Mean Silhouette	0.8910	Mean Number of Clusters	8
Scalers					
Initial scaler	minmax	Embedding scaler	none		
UMAP (nonlinear dimensionality reduction)					
Neighborhood size	47	Minimum distance	0.1204	Embedding dimension	7
Distance metric (input space)	chebyshev	Spread	1.0357	Local connectivity	6
Learning rate	0.6551	Training epochs	4540		
HDBSCAN (density-based clustering)					
Minimum cluster size	492	Minimum samples	68	Cluster selection method	eom
Distance metric (embedded space)	euclidean	Alpha	1.0454	Cluster-selection epsilon	0.0290

Table 3.4: Selected configuration for the clustering of acoustic responses at a 75 dB threshold

Unlike in the 55 dB case, this setup does not consistently yield the same number of clusters across repeated runs. To handle this variability, we executed the clustering pipeline several times and retained the run that produced the most visually separable structure. In this instance, the result consisted of 7 clusters (plus outliers). With this configuration fixed, we now turn to the embedded space, shown in Figure 3.25, where points are colored according to their assigned cluster ID:

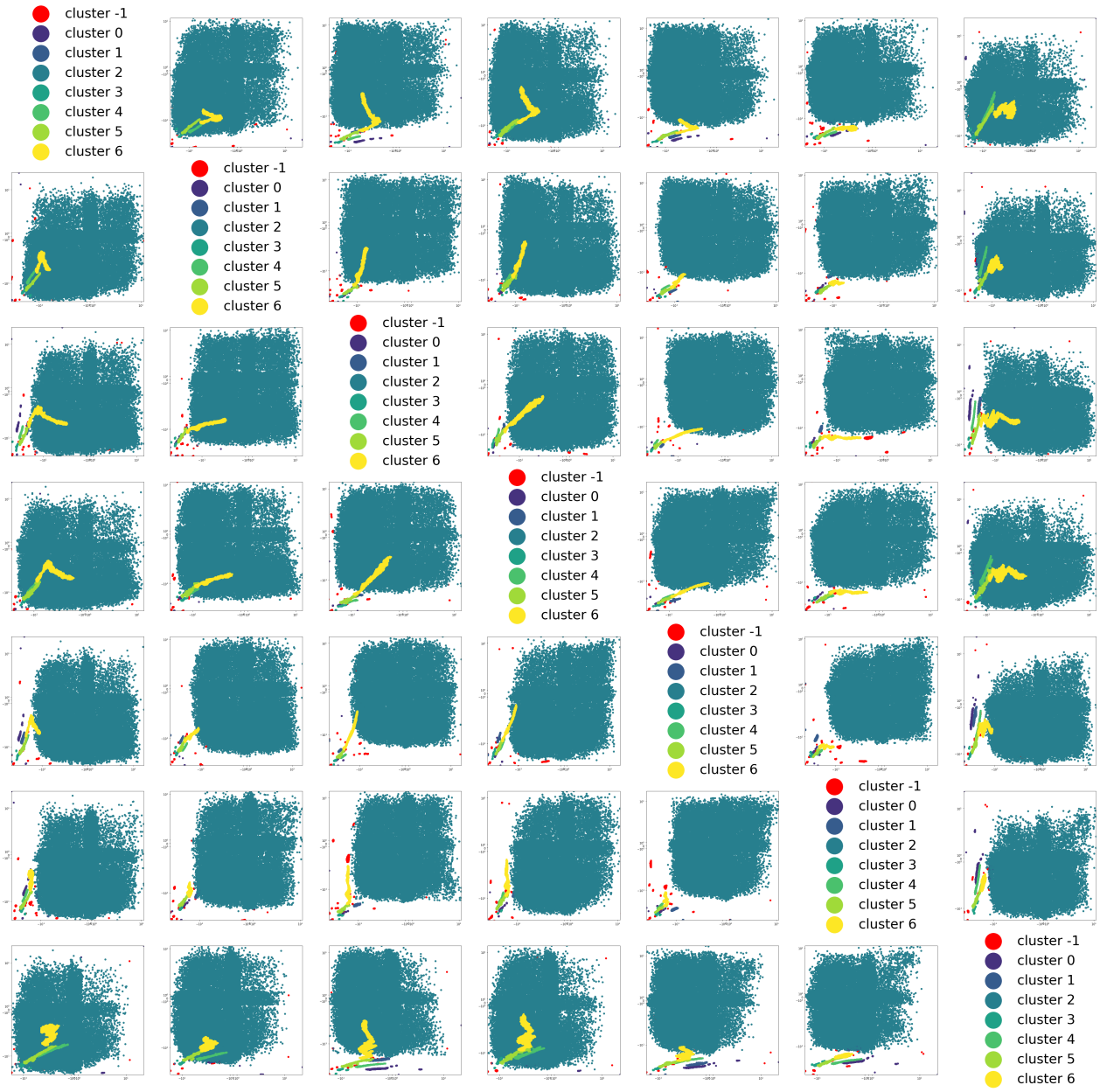
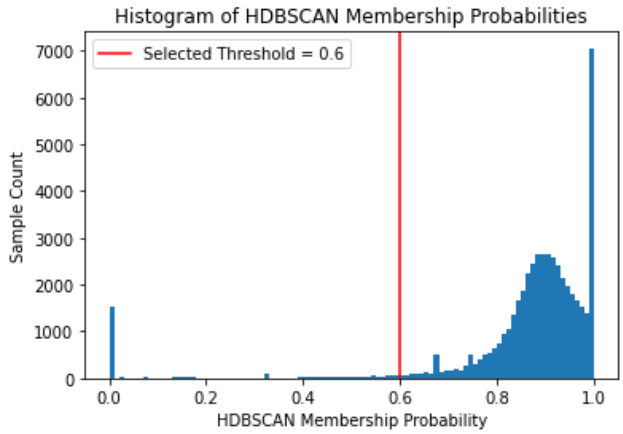
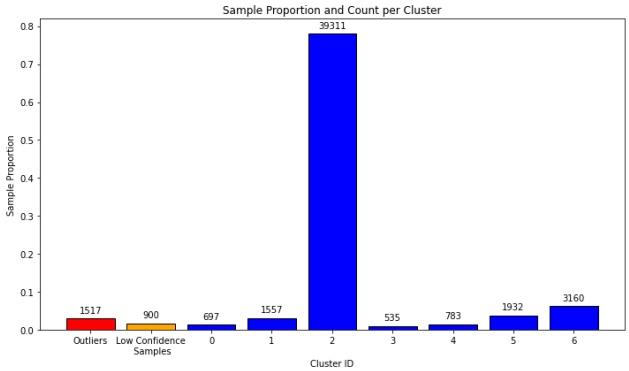


Figure 3.25: Embedded space colored by cluster ID. -1 cluster contains the outliers detected by HDBSCAN

The embedded space exhibits even poorer cluster separation than in the 55 dB case. Whether this is due to the GPU-based implementation of UMAP (as discussed earlier), the absence of a hyperparameter configuration yielding a consistent number of clusters, or some intrinsic structure of the data itself, remains difficult to determine. Next we examine cluster assignment and confidence as shown in Figure 3.26:



(a) HDBSCAN membership probabilities histogram.

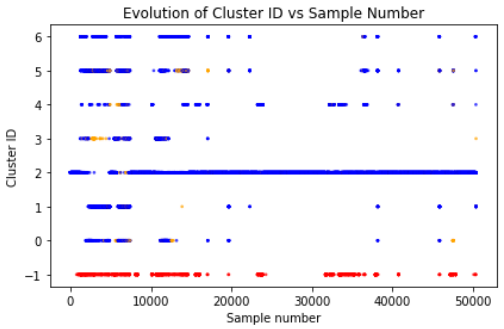


(b) Sample proportion and count per cluster.

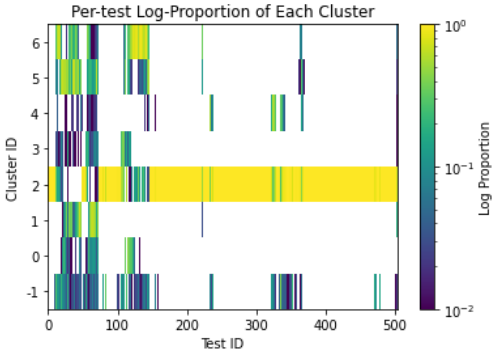
Figure 3.26: Clustering confidence and distribution.

The histogram of HDBSCAN membership probabilities shows a large proportion of highly confident assignments (near 1), followed by a truncated Gaussian-like curve that flattens around 0.6. Applying an elbow-type criterion, we set a cutoff at 0.6 to remove potentially unreliable points that could otherwise perturb the training process. The resulting histogram of cluster memberships reveals a strongly unbalanced distribution: cluster2 alone accounts for nearly 80% of all samples, while some clusters contain as few as 535 points. This imbalance is even more pronounced than in the 55dB case and is likely to affect training stability. On the positive side, the outliers and low-confidence samples constitute only a very small fraction of the dataset, which is acceptable.

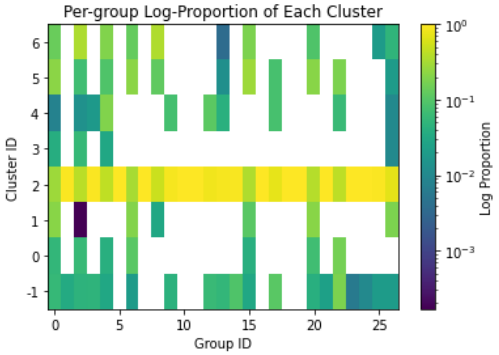
Finally, we analyse the evolution of cluster assignments across different ordering schemes to visualise potential patterns. Specifically, we examine their progression through ordered samples, ordered tests, and ordered test groupss, as illustrated in Figure 3.11.



(a) Cluster ID through time-ordered samples.



(b) Histograms of cluster ID distribution per ordered test.



(c) Histograms of cluster ID distribution per ordered test group.

Figure 3.27: Evolution of clustering through time, tests, and test groups.

Similarly to the 55dB threshold case, graph(a) exhibits a two-step regime: the less frequent clusters (all but cluster 2) occur more often in the first half of the protocol and become scarcer in the second half.
However, this does not suggest that the clusters arise from test group–specific background

noise. In contrast to the 55dB case, the less frequent clusters are now distributed across a larger number of test groups and tests. This observation supports our earlier assumption that group-specific background noise largely vanishes at the 75dB threshold, making the analysis more likely to capture genuine mechanical phenomena.

3.3.2.2 Cluster analysis

Having identified the cluster structure at the 75 dB threshold and similarly to before, we now investigate the characteristics of each cluster to assess their distinctiveness and potential physical meaning. We begin by observing the acoustic responses grouped by cluster, illustrated in Figure 3.28:

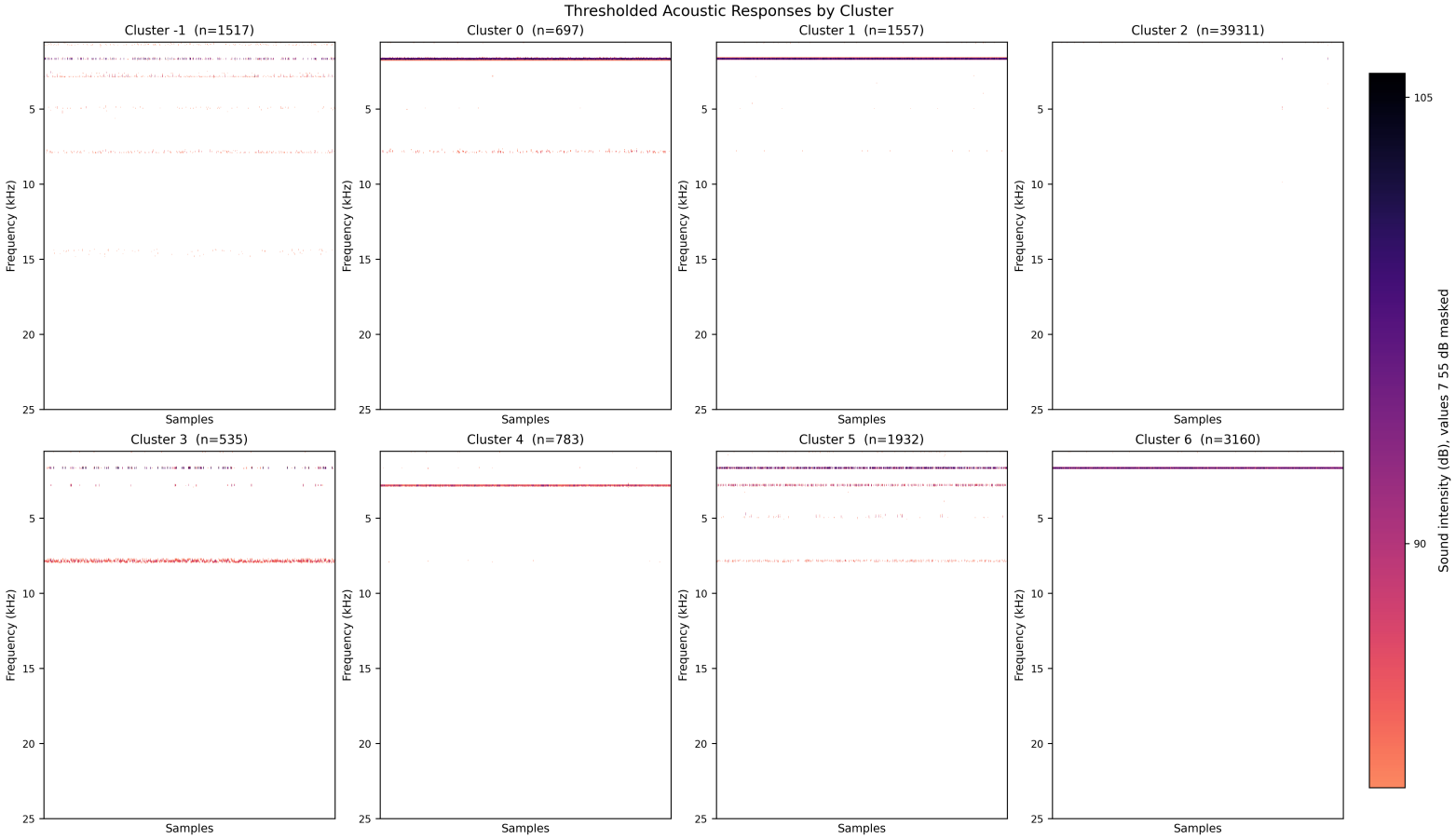


Figure 3.28: Thresholded acoustic responses grouped by cluster. Values below the 75 dB threshold are masked (white regions). Each subplot corresponds to one cluster, with samples along the horizontal axis and frequency (kHz) along the vertical axis.

The thresholded acoustic responses at 75dB are noticeably simpler than those at 55dB. They are composed primarily of high-energy events (mainly squeals) occurring in four distinct frequency bands: approximately 1300Hz, 2500Hz, 5000Hz, and 7500Hz. The clusters differ mainly by which of these bands are activated: cluster2 shows no activation at all (consistent with its dominance as the majority case), cluster4 is characterized by activation around 2500Hz, and cluster5 activates all four bands simultaneously. Even subtle differences are visible—for example, cluster1 and cluster6 both exhibit activation around 1300Hz, but cluster1 spans a slightly broader range (about 500 Hz wider). The noise cluster (-1) appears as a heterogeneous mixture lacking clear structure, which is consistent with expectations. These observations suggest that the clustering captures recurring, well-defined spectral signatures

of high-energy acoustic events. In the following, we investigate whether these spectral differences are also reflected in the auxiliary signals.

We start with the Thermal-related variable histograms (Figure 3.29).

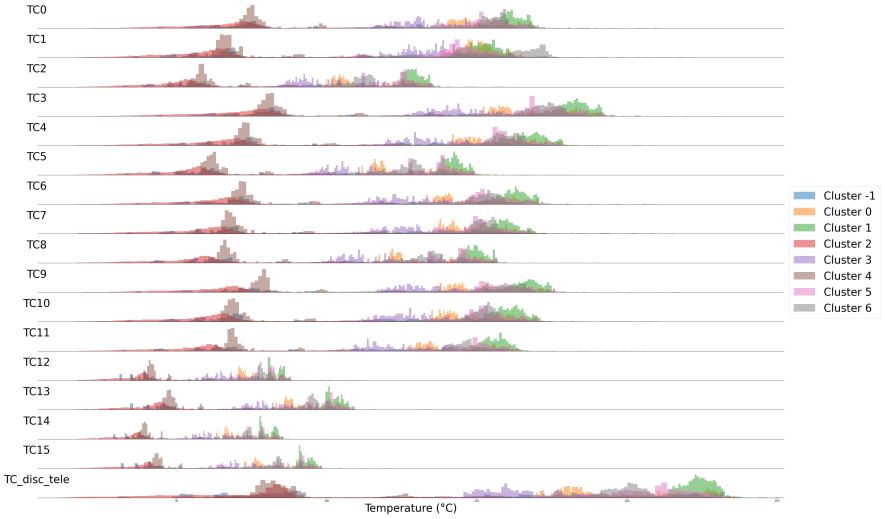


Figure 3.29: Thermal Variable Histograms colored by Acoustic Cluster.

An interesting phenomenon emerges when comparing thermal distributions across clusters. The clusters associated with higher temperatures—namely clusters 0, 1, 5, and 6, and to a lesser extent cluster3—are also those that activate the ∼1300Hz acoustic band. Among them, cluster1 spans the broadest frequency range and corresponds to the warmest temperatures, whereas cluster3 only weakly activates the ∼1300Hz band and is the coldest of this subset. In contrast, clusters2 and 4 are noticeably cooler overall, even though cluster4 reliably activates the ∼7.5kHz band (while cluster2 shows no activation at all). These observations suggest that the different identified frequency bands are linked to distinct thermal regimes, indicating that spectral activation patterns may directly reflect variations in underlying thermal properties. We continue with the Mechanical-related variable histograms (Figure 3.30).

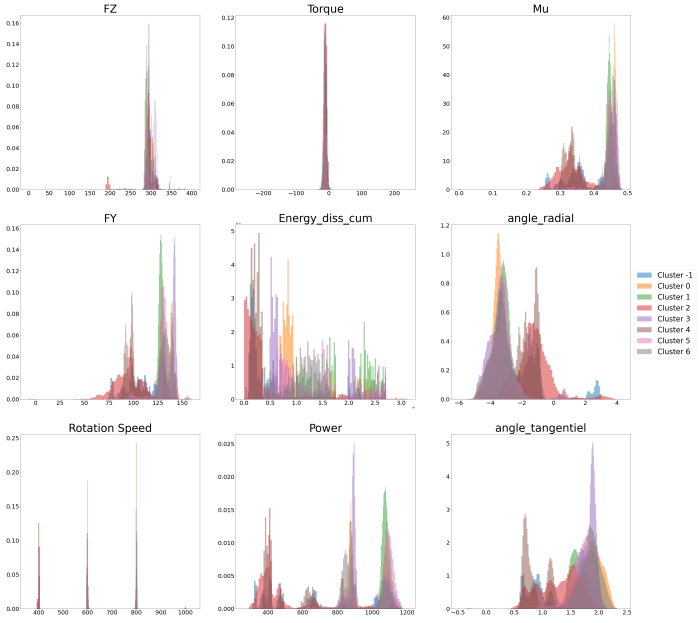
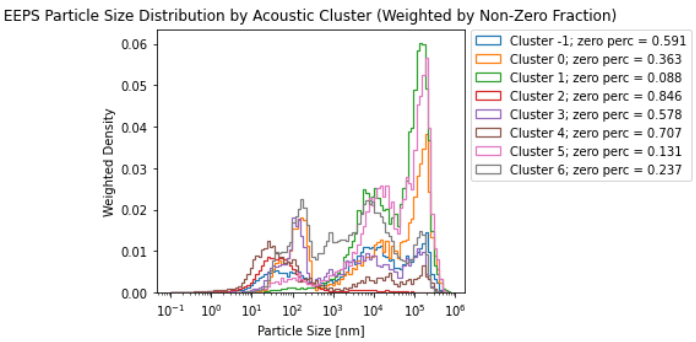


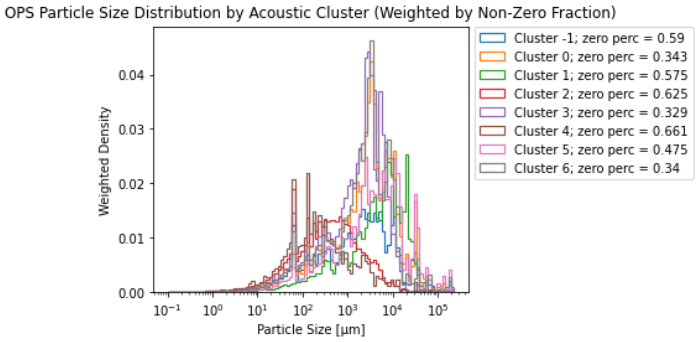
Figure 3.30: Mechanical Variable Histograms colored by Acoustic Cluster.

Regarding the mechanical variables, the separation is less pronounced. A similar grouping emerges to what was observed with temperature: clusters that activate the ∼1300Hz acoustic band tend to differ from those that do not, particularly in μ, radial angle, cumulative dissipated energy, F_Y, power, and tangential angle. However, unlike at the 55dB threshold, no single cluster stands out as being consistently distinct across these variables.

We now turn to particle emissions, whose distributions across acoustic clusters are shown in Figure 3.31.



(a) EEPS Distribution by cluster



(b) OPS Distribution by cluster

Figure 3.31: Flattened EEPS and OPS particle amount distributions colored by acoustic cluster. All size bins were combined into a single distribution.The Densities weighted by the proportion of non-zero readings to allow for better comparison.

Particle emissions exhibit even weaker clustering patterns than temperature or mechanical variables. In both the EEPS and OPS distributions, clusters2 and4—the ones that do not activate the ∼1300 Hz acoustic band—stand out as emitting less overall. Not only do they show somewhat reduced intensity when emissions are present, but they also have comparatively high zero-emission proportions (around 0.6–0.8), indicating that emissions are absent more frequently in these clusters. Nevertheless, the histogram differences remain small, making it difficult to establish this effect with statistical certainty.

Finally, similarly to the 55 dB case, we can get a naive evaluation of the contact location during each acoustic response via the barycentre of shallow thermocouples time deltas. The resulting two-dimensional barycentre positions can then be visualised as histograms for each cluster in Figure 3.32.

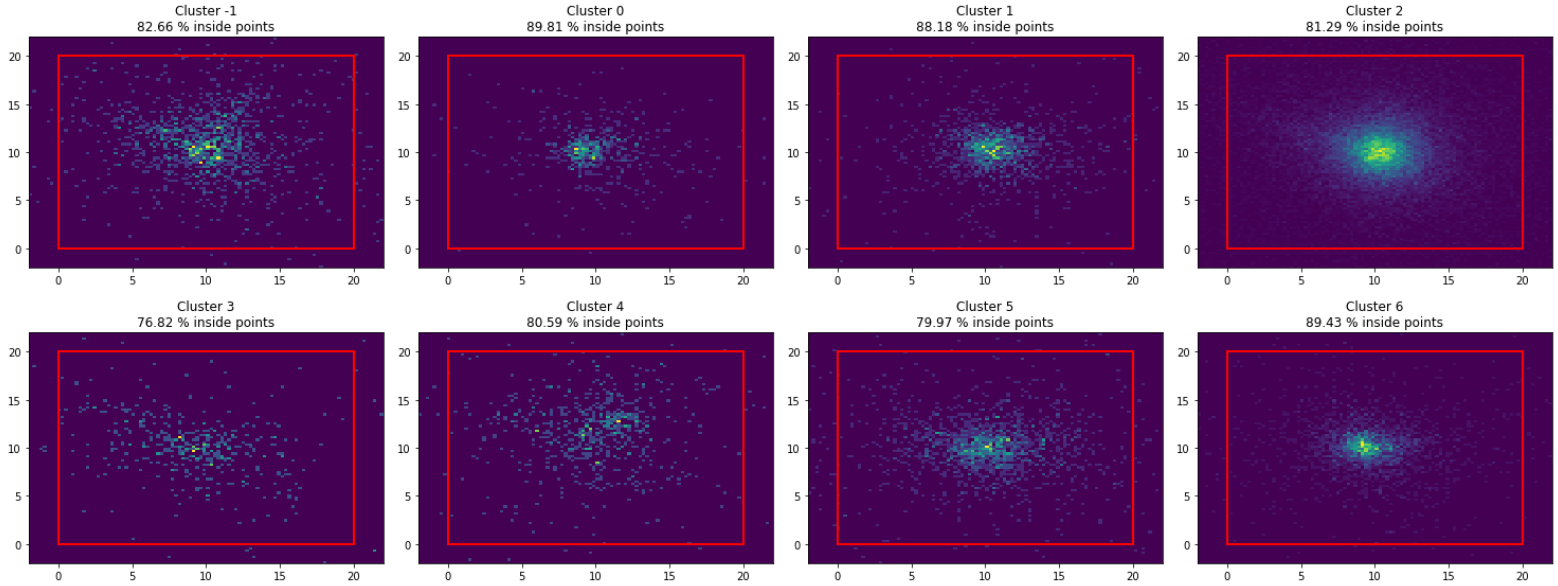


Figure 3.32: Surface temperature thermocouple time delta barycentres. Red rectangle represents the border of the pin. The percentage of inside point is the percentage of computed barycentres that is within the pin. The samples outside the pin can be explained by errors made during the capture of the data.

As we did in the 55 dB threshold case, we can now compare the distributions using a 2D Kolmogorov-Smirnov test as compairing these histograms visually is often hard. The resulting p-values are shown in Figure 3.33:

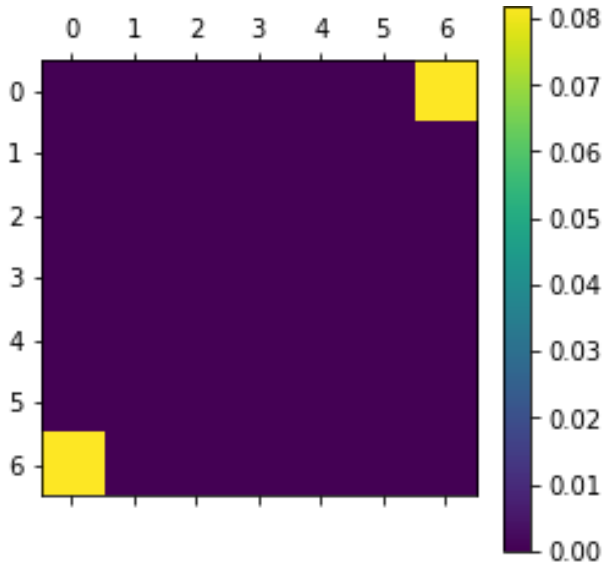


Figure 3.33: 2D Kolmogorov–Smirnov approximation test p-values by cluster ID.

All cluster pairs exhibit distinct differences in their thermal delta barycentre distributions, with the sole exception of the (0,6) pair. We attempted to interpret this similarity in terms of a physical mechanism, noting that both clusters activate the ∼1300 Hz acoustic band. However, this cannot account for the overlap, since clusters 1, 3, and 5 also activate the same band yet do not display comparable distributions. In the absence of a clear mechanical explanation, the similarity may instead result from the limited sample size, a chance effect of random sampling, or a shared distribution of "naive" contact locations. For all other pairs,

however, the results indicate that the clustering provides meaningful differentiation in our proxy for contact location .

In summary, the cluster analysis shows that the acoustic-based grouping at 75 dB is reflected differently than at 55 dB across the auxiliary signals. The observed differences arise more from cluster attributes (i.e., which frequency band is activated) than from cluster identity itself. Specifically, the thermal signal shows strong dependence, the mechanical signal moderate-to-strong, the particle signal weak, and the barycentre signal primarily reflects cluster identity. This suggests that clustering based solely on acoustic responses may not be the most appropriate approach; instead, the problem might be better framed as one of source activation. Despite these limitations, the clustering analysis still provides valuable insights into the structure of the data. Building on these observations, we now turn to the classification stage.

3.3.2.3 Classification

Building on the clustering results and analysis, we next examine the classification stage, starting with a summary of the model architecture in Figure 3.34.

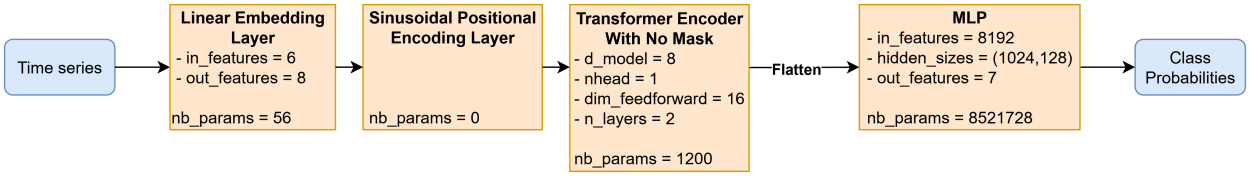
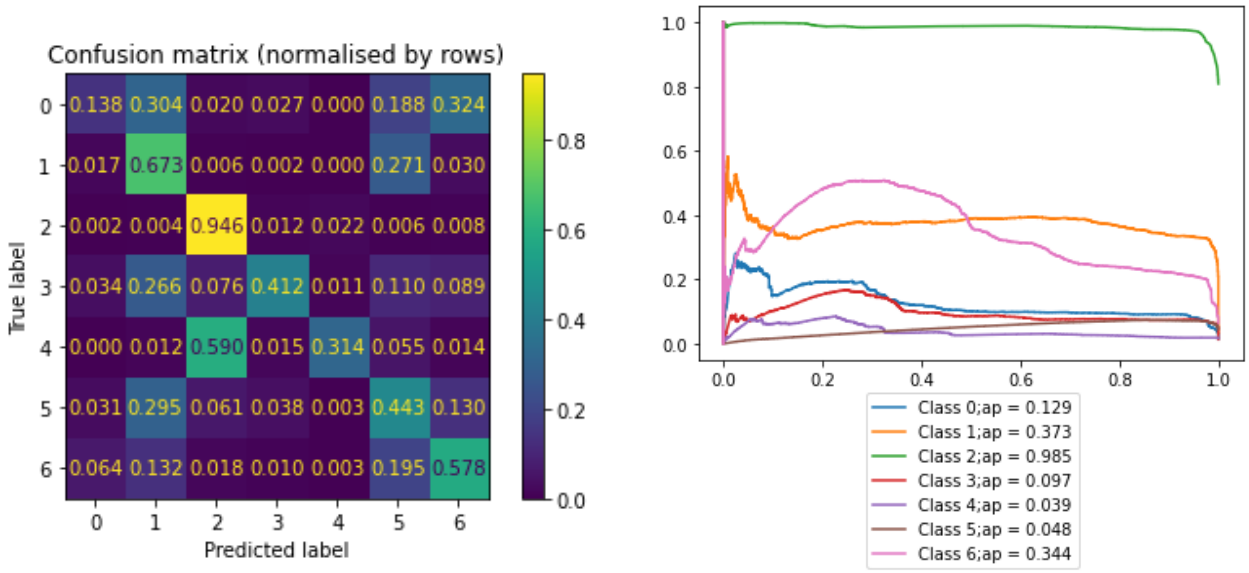


Figure 3.34: Overview of Classification Transformer model used for the 75 dB threshold

This architecture results in a total of 8,522,984 parameters, which is approximately 17 times fewer than the model used for the 55 dB classification. This scale is far more realistic given the size of our database. However, as will be shown later, its final evaluation metrics do not approach those of the 55 dB model. The fact that even architectures with capacity comparable to or greater than the 55 dB model failed to reach similar performance suggests that this classification task is inherently more difficult. At the same time, it also indicates that the model may not simply be overfitting to the experimental protocol, but rather capturing more meaningful mechanical phenomena.

We now turn to the evaluation of the developped classifiers. The results are presented in the form of the confusion matrix, one-vs-rest precision–recall curves (see Figure 3.35), and the full classification report (see Table 3.5):



(a) Classifier Confusion Matrix (normalised by True labels)

(b) OvR Precision Recall Curve with Average Precision

Figure 3.35: Classification results

Class	Precision	Recall	F1-score	Support
0	0.22	0.14	0.17	751
1	0.41	0.67	0.51	1607
2	0.98	0.95	0.96	38074
3	0.30	0.41	0.34	616
4	0.23	0.31	0.26	803
5	0.38	0.44	0.41	2130
6	0.65	0.58	0.61	3070
Accuracy			0.86	47051
Balanced accuracy			0.50	47051
Macro avg	0.45	0.50	0.47	47051
Weighted avg	0.88	0.86	0.86	47051

Table 3.5: Classification report

The results reveals a three-tier performance pattern. First, cluster 2 stands out as the only class predicted with consistently high accuracy (94.6%), reflected in its strong F1-score (0.96) and very high average precision (AP = 0.985). Second, clusters 1 and 6 achieve intermediate performance, with F1-scores of 0.51 and 0.61 and corresponding AP values of 0.37 and 0.34. While the classifier is able to identify these clusters in a majority of cases, the reliability of the predictions is limited. Finally, clusters 0, 3, 4, and 5 show much weaker results, with F1-scores ranging from 0.17 to 0.41 and AP values below 0.13 (clusters 3, 4, and 5 in particular). These classes are frequently confused with others, indicating that their acoustic signatures are less easy to predict from auxiliary signals.

This three-tier structure is also reflected in the global metrics. Although the raw accuracy is 0.86, the balanced accuracy drops to 0.50, highlighting that performance is dominated by the majority cluster 2. The classifier thus appears highly specialized in detecting the dominant class but struggles to provide consistent discrimination across the minority clusters.

Taken together, these results suggest that the classification task at this sound level is substantially more challenging than in the 55 dB case. The fact that the model predicts cluster 2 reliably but performs poorly on all other clusters indicates that it has effectively learned little more than a binary distinction between the dominant class and the rest—akin to an "is there noise or not" classifier. We should keep these shortcomings in mind when applying interpretation methods later on, as the poor decision boundaries of the classifier are likely to translate into unreliable or misleading explanations. Addressing these weaknesses could involve targeted data augmentation, improved training strategies, or even modifications to the experimental protocol to ensure more frequent representation of the less prevalent classes.

3.3.2.4 IG Calibration

Although the classifier has not demonstrated reliable performance, it is still worthwhile to complete the interpretation pipeline, as this can still provide useful insights, provided we remain cautious given the model's limited performance. As before, we begin with the computation of Permutation Feature Importance (PFI), which we evaluate using balanced accuracy rather than raw accuracy, given the highly imbalanced class distributions in our setting. The results are as shown in Figure 3.36 :

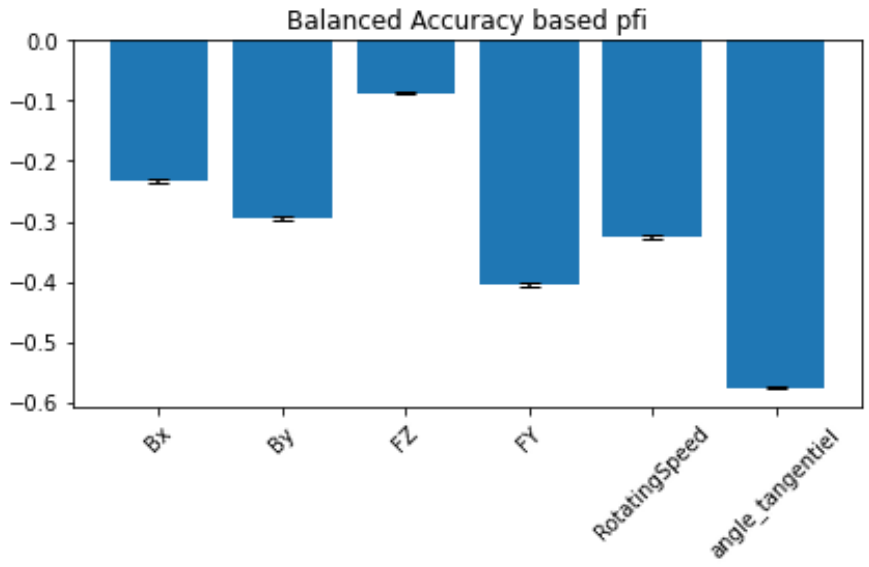


Figure 3.36: Balanced accuracy based PFI for the classification model at 75 dB threshold

Based on these scores, the next step is to identify the most suitable IG baseline. To do so, we calculate the absolute aggregated attributions for each variable under the different baseline candidates and then assess their alignment with the PFI scores.The results are reported in Table 3.6.

Method\Variable	B_x	B_y	F_z	F_y	Rotation speed	Tangential angle	Pearson r
PFI	-0.14 (6)	-0.24 (2)	-0.09 (5)	-0.22 (3)	-0.26 (1)	-0.18 (4)	
Time-dependent mean	3.49×10^{-3} (6)	5.78×10^{-3} (4)	2.97×10^{-3} (5)	6.92×10^{-3} (2)	1.15×10^{-2} (1)	6.07×10^{-3} (3)	-0.84
Constant mean	7.33×10^{-3} (4)	7.86×10^{-3} (2)	9.52×10^{-3} (1)	7.31×10^{-3} (3)	1.60×10^{-2} (0)	5.15×10^{-3} (5)	-0.36
Time-dependent median	3.99×10^{-3} (6)	5.85×10^{-3} (4)	2.78×10^{-3} (5)	7.07×10^{-3} (2)	7.78×10^{-3} (3)	6.19×10^{-3} (1)	**-0.93**
Constant median	9.45×10^{-3} (2)	6.15×10^{-3} (4)	7.35×10^{-3} (3)	6.62×10^{-3} (5)	1.73×10^{-2} (1)	4.91×10^{-3} (6)	-0.36
Zeros	3.32×10^{-2} (3)	4.93×10^{-2} (1)	4.14×10^{-2} (2)	4.06×10^{-2} (0)	1.10×10^{-2} (5)	3.11×10^{-2} (4)	0.30
Ones	3.30×10^{-2} (1)	1.63×10^{-2} (3)	1.45×10^{-2} (5)	2.60×10^{-2} (2)	3.08×10^{-2} (0)	1.37×10^{-2} (4)	-0.26
Midpoint	4.05×10^{-3} (6)	1.17×10^{-2} (2)	1.37×10^{-2} (1)	1.16×10^{-2} (3)	1.37×10^{-2} (0)	6.33×10^{-3} (5)	-0.27
Physical baseline 1	1.98×10^{-2} (2)	4.86×10^{-3} (3)	2.23×10^{-4} (5)	5.29×10^{-4} (4)	5.85×10^{-5} (6)	1.96×10^{-2} (1)	0.28
Physical baseline 2	1.41×10^{-2} (2)	7.80×10^{-3} (3)	2.39×10^{-4} (5)	5.80×10^{-4} (4)	6.44×10^{-5} (6)	2.03×10^{-2} (1)	0.16

Table 3.6: Permutation Feature Importance (PFI) scores and absolute IG aggregated values per variable, along with the Pearson r between PFI and IG for each baseline. Parentheses indicate the rank of each variable according to the method. Note that PFI ranks decrease with higher importance (negative values), while IG-based values increase.

Based on the Pearson r correlation values, the time-dependent median again emerges as the most suitable baseline. A comparison of the global attributions obtained with PFI and Time-dependant median IG is presented in Figure 3.37.:

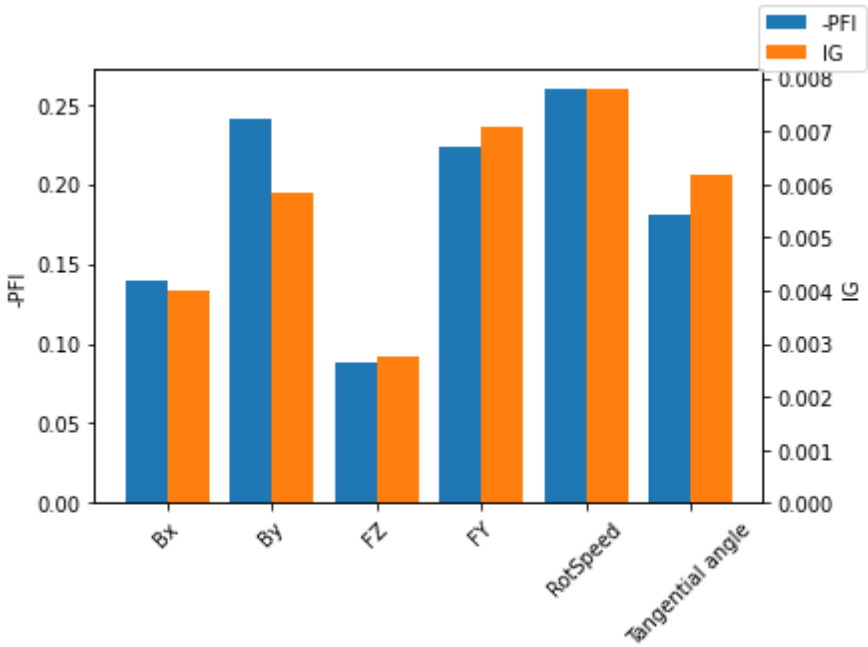


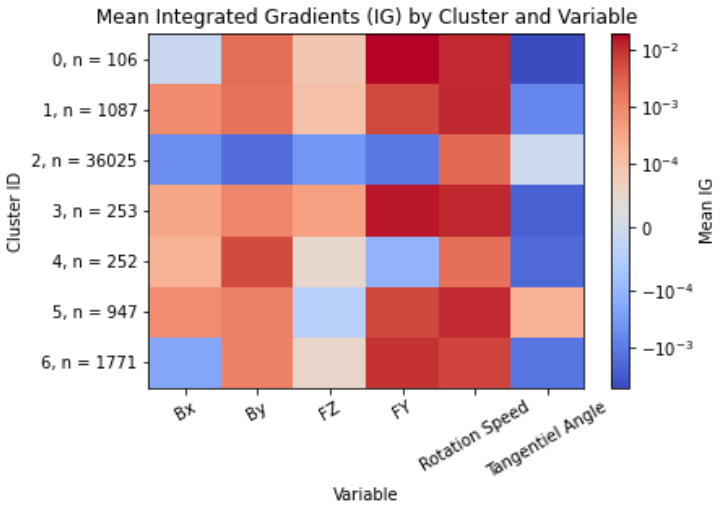
Figure 3.37: PFI vs Time-dependant median guided IG global scores per variables

PFI and IG attributions exhibit a clear overall consistency. Apart from B_y and the tangential angle, the relative importance and ranking of variables are largely preserved across the two approaches. This indicates that the selected IG baseline produces attributions that broadly track the model-agnostic PFI scores. Still, the weak predictive performance of the classifier limits the strength of this conclusion: the apparent agreement is encouraging, but it should not be taken as evidence that the explanations are fully reliable.

3.3.2.5 Interpretation

With the baseline identified as appropriate, we now apply IG to investigate the model's decision process. As an initial step, we consider the mean attributions of each variable across clusters, examining them both in absolute terms and with their original sign, as illustrated in Figure 3.38.

(a) Within-cluster normalized mean absolute IG values, reflecting the relative importance of variables inside each cluster.



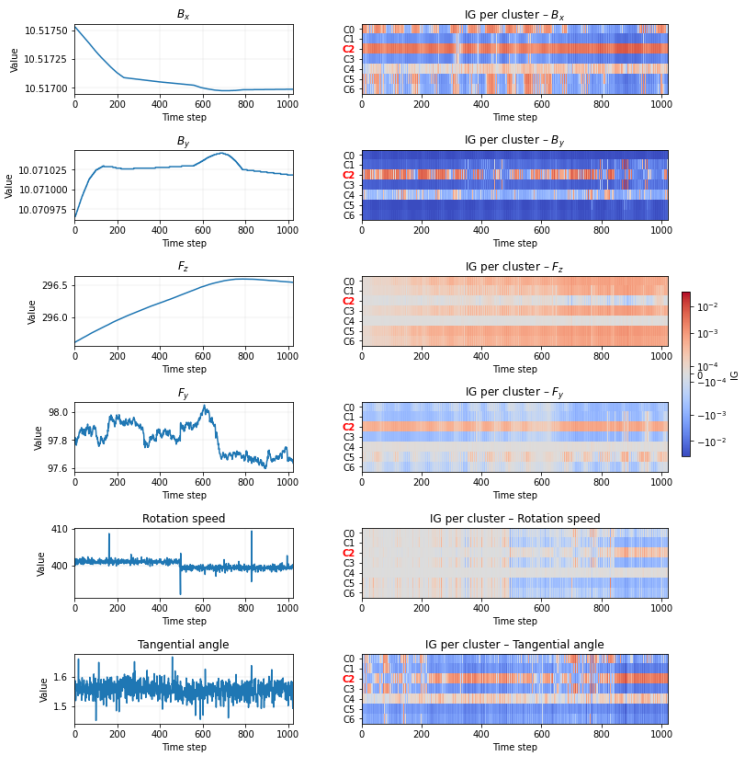
(b) Mean signed IG values, indicating whether variables push the prediction towards or away from each cluster in general.

Figure 3.38: Cluster-level attribution analysis. (a) Relative importance of variables per cluster based on normalized mean absolute IG. (b) Directional effect of variables based on mean signed IG.

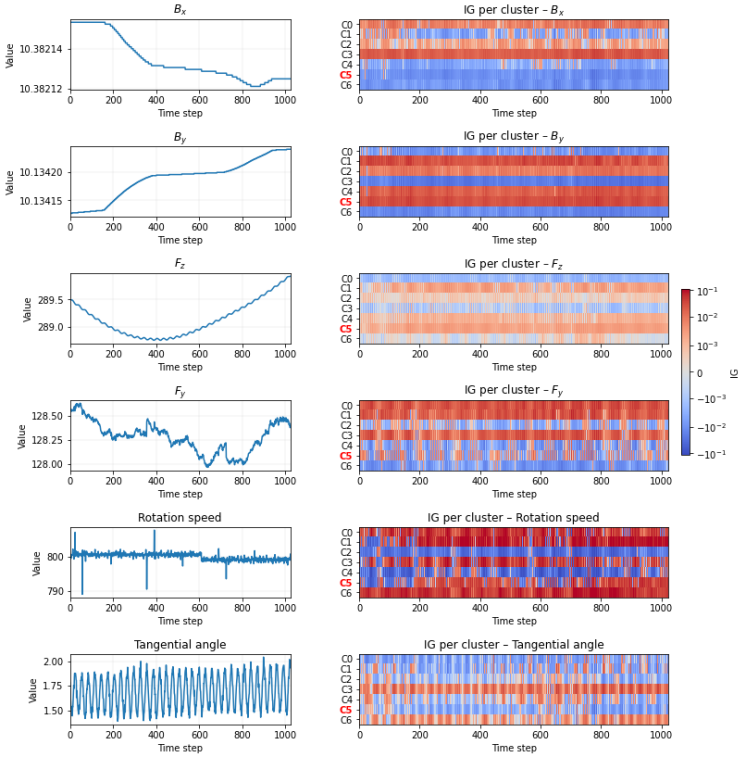
The graphs are interpreted in the same way as before (see 3.3.1.5). From graph (a), three groups can be distinguished. Cluster2 does not appear to be strongly influenced by any single variable, which suggests that the model may be representing it largely through the bias term of the final MLP layer: it becomes the default prediction unless other variables push the decision toward a different cluster, which is coherent with it's "no emission" nature. Cluster4 shows a clear dependence on B_y, rotation speed, and the tangential angle. Finally, clusters 0, 1, 3, 5, and 6 share a similar attribution profile, being primarily shaped by rotation speed and F_y; interestingly, they also coincide as the clusters associated with the ∼1300kHz emission band, suggesting that this frequency may be linked to the influence of these variables.
Graph (b) adds complementary information: nearly all of the variables identified as important in graph (a) display positive mean signed IG values, with the tangential angle for cluster 4 being the main exception. This pattern suggests that the classifier does not sharply discriminate between opposing effects, but instead tends to accumulate positive attributions across variables. Such convergence may reflect the difficulty of the classification task itself and could contribute to the model's poor performance.

These observations provide useful intuition, but they should be interpreted with caution: the small number of available samples and signs of poor convergence mean that the patterns seen in graphs (a) and (b) may not reliably reflect the underlying mechanical reality.

After having interpreted the model at a global and cluster level, we can once again turn to individual examples (Figure 3.39) to see how IG explains specific predictions:



(a) Instance-level IG visualization for sample 9961.



(b) Instance-level IG visualization for sample 20120.

Figure 3.39: Inputs (left of each panel) and signed IG per cluster over time (right of each panel) for two contrasting samples. A shared symmetric–log color scale is used across variables and clusters.

The first example corresponds to a cluster 2 prediction. As the colorbar indicates, the attribution magnitudes are overall lower than in the following example, reinforcing the impression that no variable contributes strongly, with little evidence of any single one exerting a dominant positive effect. This suggests that the classifier predicts cluster2 largely by default rather than being actively "pushed" toward it by particular inputs and is consistent with what we saw earlier in the global interpretations, where cluster2 stood out as having no clearly decisive variables.

The second example illustrates a prediction in cluster5. The colorbar shows higher attribution magnitudes than in cluster2, and oscillatory patterns in the attributions closely track oscillations in the raw inputs, most clearly for the tangential angle. As in the 55dB case, the classifier appears to adapt directly to input values rather than capturing more intricate temporal dynamics. Unlike at 55dB, however, the weak predictive performance here does not point to overfitting; instead, it may reflect either the absence of richer temporal dependencies in the data or insufficient convergence of the model. Finally, as in previous examples, the model selects relevant variables in a context-dependent manner, with different features becoming important for different samples.

It is worth emphasizing once again that simpler tabular classifiers trained on summary statistics of the time series performed worse. This suggests that the model may in fact be capturing subtler patterns that are not readily visible in the IG attributions—either because of the way they are represented, or due to limitations introduced by our baseline choice.
Overall, this subsection offers useful, though limited, intuition about the global and instance-

level mechanical factors underlying emissions at the 75 dB level. The analysis has been restricted to correctly classified examples, yet these still originate from a model with modest predictive performance. The consistency between IG and PFI and the recurring patterns across clusters suggest that some meaningful signals are being captured, but the explanatory power remains constrained. These results should therefore be viewed as providing preliminary insight rather than definitive conclusions, highlighting both the promise of the approach and the need for stronger modelling to support more reliable interpretations.

3.3.2.6 Conclusion

To conclude the analysis at the 75 dB threshold, we summarize the main findings.
First, the Optuna study did not yield stable clusters across runs. However, the clusters are no longer grouped by test or test group, which is consistent with our preliminary analysis suggesting that background signal disappears at this higher threshold.

Second, the resulting clusters are highly visually differentiable and appear to be characterized primarily by the activation or non-activation of four distinct frequency bands. Cluster analysis revealed clear links between these frequency bands and both thermal and mechanical variables, with a weaker but still visible relation to particle emissions. The naive contact-location proxy also showed differences aligned with cluster ID.

Third, classification performance was poor. Only the majority "no-emission" cluster (cluster 2) was predicted reliably, while all others were confused, undermining the reliability of subsequent interpretation.

Fourth, despite these limitations, we carried out IG calibration and interpretation. Calibration once again identified the time-dependent median as the most suitable baseline. The IG analysis suggested that cluster 2 is effectively represented through a bias term, while the remaining clusters depend on different combinations of factors, sometimes coherently linked to their associated frequency bands. Nonetheless, the poor classification performance means these interpretations should be treated with caution.

Overall, the 75 dB results reinforce the importance of threshold choice: while background influence is reduced and clusters become more physically interpretable, their scarcity and imbalance make classification and subsequent explanations fragile. The pipeline at this threshold thus provides promising but limited insight.

The whole pipeline took $\sim$ 18.5 Hours of compute time in total. See Appendix A.1 for details on the software, hardware, and computational resources used.

A possible avenue for future work would be to move away from clustering and instead adopt a source-separation perspective. Algorithms such as Non-negative Matrix Factorisation (NMF [33]) could be used to decompose the acoustic signals into distinct sources—here, likely corresponding to the observed frequency bands. Once extracted, these sources could then be linked back to auxiliary variables through classification or regression models, allowing us to determine when and to what extent each source is active. Such an approach would disentangle the overlapping clusters associated with the 1300 Hz band and could lead to more robust modelling by focusing directly on the causal drivers of each frequency component.

3.4 Conclusion

This chapter set out to move beyond raw prediction of emission levels and instead focus on *understanding* the mechanisms that underlie acoustic emissions. By clustering spectrogram segments and then linking them back to auxiliary variables via classification and explainability tools, we aimed to uncover structured relations between sound emissions, operating conditions, and physical states.

Several lessons emerged. At the **55 dB threshold**, the pipeline produced stable clusters that were reflected across thermal, mechanical, and (to a lesser extent) particle signals. However, much of this structure likely reflected the experimental protocol itself rather than intrinsic mechanisms. Classification performed well, but high performance was probably driven in part by this protocol-specific signal. At the **75 dB threshold**, background influence largely disappeared and clusters became more interpretable, being defined primarily by activation of a small number of frequency bands. These showed clearer links to thermal and mechanical variables. Yet, the severe class imbalance and overall scarcity of samples meant that classification failed except for the majority "no-emission" case, leaving the interpretation stage fragile.

Across both thresholds, the results point toward a key insight: grouping acoustic responses often corresponds to meaningful differences in thermal and mechanical conditions, which strongly suggests that the clustering approach captures real aspects of the emission mechanisms. At the same time, the approach revealed its limits. Classification was inconsistent across thresholds, interpretation was undermined by poor performance at 75 dB, and background signals risked skewing results at 55 dB.

Equally important are the **methodological reflections**. The pipeline combined a series of 'homemade' components:

- UMAP + HDBSCAN tuned with Optuna, using stability and silhouette as heuristic objectives,
- Integrated Gradients calibrated by aligning global attributions with PFI, a heuristic criterion rather than a principled guarantee.
- Physics-driven baselines for Integrated Gradients, which did not perform well in practice but remain a promising direction since they provide theory-based reference signals against which experimental deviations could be interpreted.

These design choices were pragmatic and made the analysis feasible, but they remain fragile. Small changes in hyperparameters, baseline selection, or model convergence could alter the interpretations. The method is therefore best viewed as an exploratory tool that organizes noisy data into interpretable structures, rather than a definitive framework for uncovering ground truth mechanisms.

Looking forward, the results suggest that clustering may not be the most appropriate formulation of the problem. Instead, **source-separation approaches** (e.g., Non-negative Matrix Factorisation, NMF) could decompose emissions into underlying frequency-band sources. In particular, at the **75 dB threshold**, where clusters became tangled and imbalanced, NMF could simplify the representation by reducing the task to tracking the activation of *four sources* (the observed frequency bands) instead of arbitrary combinations of them. Once separated, these sources could then be more robustly linked to auxiliary conditions through

regression or classification, providing a clearer mapping between physical drivers and acoustic emissions.

In sum, this chapter shows both the **promise and the limitations** of using explainable ML to study emission mechanisms. Even if the classifiers were not always reliable, the combination of clustering, classification, and attribution highlighted recurring links between acoustic bands, thermal regimes, and mechanical conditions. More importantly, it pointed the way toward improved methodologies that disentangle experimental artefacts from genuine mechanisms. Thus, while the present analysis should be interpreted cautiously, it provides a constructive step toward building more robust models of emission generation.

Chapter 4

From Understanding to Control Strategies: Reinforcement Learning in a FEM Test Bench

The previous chapters explored the use of machine learning for predicting braking emissions and for understanding acoustic patterns. While these studies provided partial successes and valuable insights, they also revealed persistent limitations. Prediction models captured some correlations but struggled to generalize across conditions, and the interpretability pipeline highlighted links between acoustic states and operating variables without yielding definitive answers. In short, progress was made toward understanding, but full explanatory control over the system remained out of reach.

This motivates a shift in perspective. Instead of aiming to perfectly model or interpret the system, we now ask a more pragmatic question: *can we train agents to directly optimize braking behavior, even without a complete physical understanding?* Reinforcement learning (RL) provides a natural framework for this challenge: by interacting with an environment, a RL agent can iteratively improve its control strategy with respect to a performance objective, sidestepping the need for a full understanding of the system.

To pursue this idea, we design a finite element method (FEM) simplified test bench. It captures essential mechanical and thermal aspects of the pin–disk contact while enabling large-scale exploration of control strategies. Within this setting, we cast braking as a sequential decision problem and use reinforcement learning to search for strategies that optimize the contact pressure distribution during sliding.

In short, this chapter shifts the focus from predicting or interpreting emissions to actively exploring how reinforcement learning can be used to search for optimized braking strategies, without requiring fundamental knowledge of the underlying emission processes or mechanisms.

4.1 Method

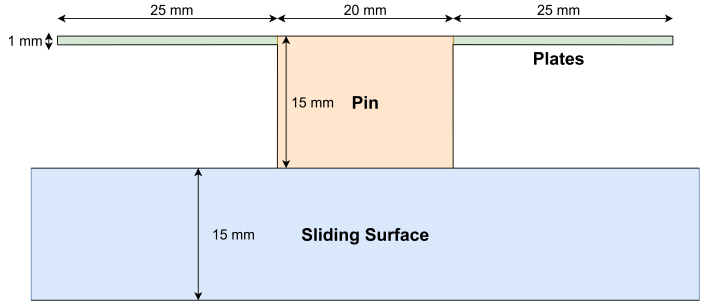
The methodological approach combines two components: a simplified finite element (FEM) simulation that provides a physically grounded environment for experimentation, and a reinforcement learning (RL) framework that optimizes control strategies within this environment. The FEM model serves as the substrate on which control decisions act, while the RL agent

adapts its policy based on the evolving system response. In what follows, we first describe the FEM environment, then formalize the control problem, and finally detail the RL training setup and reward formulation.

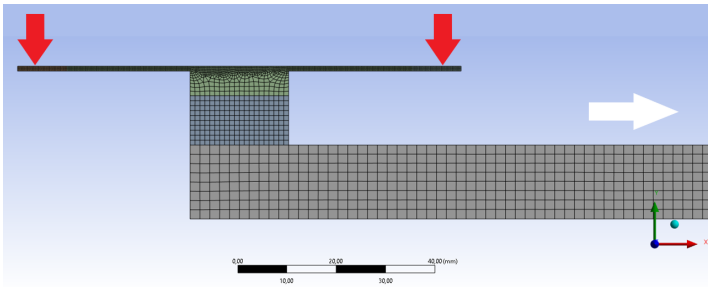
4.1.1 FEM Environment

The goal of this chapter is to develop optimized braking strategies. Because this requires a large number of adaptive decisions, relying on pre-computed databases is unsuitable, as they cannot capture the full range of operating conditions. Instead, we design a simplified finite element method (FEM) environment, tailored to be computationally efficient while retaining the essential physics needed for training and testing. This proof-of-concept model does not aim at high-fidelity reproduction of all phenomena (e.g., acoustic emissions or particulate generation), but rather at providing a tractable framework for control-oriented studies.

The FEM setup is based on a two-dimensional pin-on-disk configuration, represented in an unrolled view. The geometry consists of a pin sliding against a disk surface under controlled normal loading. The normal loading is applied on two plates. Illustrations of the environment are provided in Fig. 4.1.



(a) Schematic representation of the different bodies considered in the FEM simulation.



(b) ANSYS visualization of the FEM model. Red arrows indicate the locations where two normal forces are applied, while the white arrow denotes the sliding direction.

Figure 4.1: Illustrations of the FEM simulation environment.

After an initialization phase during which contact is established, the pin slides across the disk at a constant speed of 1mm/s. The simulation also incorporates a coupled thermal analysis, allowing us to capture the friction-induced heat generation and its dissipation into the surrounding material.

The principal physical parameters of the FEM environment are summarized in Table 4.1.

All simulations were run in parallel using 8 independent environments. On average, eight parralel environments require approximately 68 mins for episodes of 100 timesteps on our setup.

Quantity	Symbol	Value
Pin sliding speed	v	1 mm/s
Normal force (per pin)	F_N	300N
Coefficient of friction	μ	0.4
Thermal conductivity Steel/Pad	k	43/12 W/m.K /
Heat capacity Steel/Pad	C_p	445/900 J/kg.K
Density Steel/Pad	ρ	7800/2500 kg/m^3
Young Modulus Steel/Pad	E	200/2 GPa
Thermal expansion Steel/Pad	α	1,2e-5/2e-4 K^{-1}
Poisson ration Steel/Pad	ν	0.3/O.2

Table 4.1: Main physical parameters used in the FEM simulation.

4.1.2 Control Problem Formulation

Having described the FEM environment in which the agent operates, we now turn to the control problem itself. While the FEM model captures mechanical and thermal effects, it does not allow for accurate representation of acoustic or particulate emissions. As a proof of concept, we therefore focus on a surrogate control objective: achieving a uniform pressure distribution across the pin nodes in contact with the sliding surface.

The control problem can be stated as follows: given the pressure distribution across 21 contact nodes since the beggining of the simulation and the model's past decisions, determine the optimal percentage of the total force (fixed at 300N) to apply to each of the two plates at each timestep.

Before control begins, the FEM environment executes a first step during which mechanical contact between pin and disk is established. To encourage diversity in the phenomena that may appear, we then impose the force ratio at the second timestep, which also marks the onset of sliding motion. From the third timestep onward, the agent determines the force repartition at each step, thereby controlling the subsequent evolution of the system.

4.1.3 Reinforcement Learning Approach

The control task defined above is sequential and long-term in nature: each action influences not only the immediate system response but also the future evolution of the dynamics. Reinforcement learning (RL) provides a principled framework to address such problems by explicitly optimizing policies with respect to the cumulative return over time.

Among RL algorithms, we use the *Soft Actor–Critic* (SAC[20]) algorithm. At a high level, SAC learns two components: an **actor**, which selects actions given the current state, and **critics**, which estimate the long-term value of state–action pairs. The critics provide learning signals to improve the actor, while the actor balances exploiting high-value actions with maintaining sufficient exploration. A schematic overview of the optimization and interaction mechanisms is provided in Section A.12, which details the technical formulation.

SAC is particularly well suited to our setting as interactions with the finite element (FEM) environment are computationally expensive, which makes **sample efficiency** crucial. By reusing past experience through a replay buffer, SAC reduces the number of simulations

required, and it has been shown in the literature to achieve competitive sample efficiency compared to other state-of-the-art algorithms.

The following section details the training setup and reward formulation used for the control experiments.

4.1.4 Training Setup

With the RL framework established, we now describe the concrete training setup used for this chapter:

State space : The state representation (i.e., the information available to the agent when making decisions) consists of the past and present normalized pressures across the 21 contact nodes, together with the past force repartition.

Action space : The action space (i.e., the form of the agent's decisions) is continuous and one-dimensional. At each timestep, the agent selects the proportion of the total normal force (fixed at 300 N) applied to the first plate; the second plate receives the complement. Formally,

$$a_t \in [0, 1], \quad F_t^{(1)} = 300\, a_t, \quad F_t^{(2)} = 300\, (1 - a_t).$$

Reward : The reward (i.e., the signal that guides learning) is based on the entropy of the normalized pressure distribution, complemented by a progress term to prevent static policies. Details of the formulation are given in Section 4.1.5.

Environment initialisation : During training, the initial force repartition at the second timestep is sampled randomly in order to promote robustness across different initial conditions. During evaluation, however, we use a fixed grid of 40 initial repartitions equally spaced between 0.2 and 0.8. This allows for a systematic assessment of the learned policies across a range of representative scenarios.

Parralelisation : To accelerate data collection, we run eight environments in parallel. This parallelization is synchronous at the timestep level: all environments advance one step, their results are gathered and appended to the replay buffer, the networks are optimised, and then the next step is executed. Unlike more advanced distributed approaches that merge gradients across asynchronous learners, our setup only parallelizes simulation. While less sophisticated, this approach is straightforward to implement and proved sufficient to obtain stable learning behavior in our setting.

Actor and critic architectures : Both consist of an LSTM encoder followed by a multilayer perceptron (MLP). The critic receives both the last LSTM hidden state and the evaluated action as inputs to the MLP. A schematic overview is shown in Fig. 4.2.

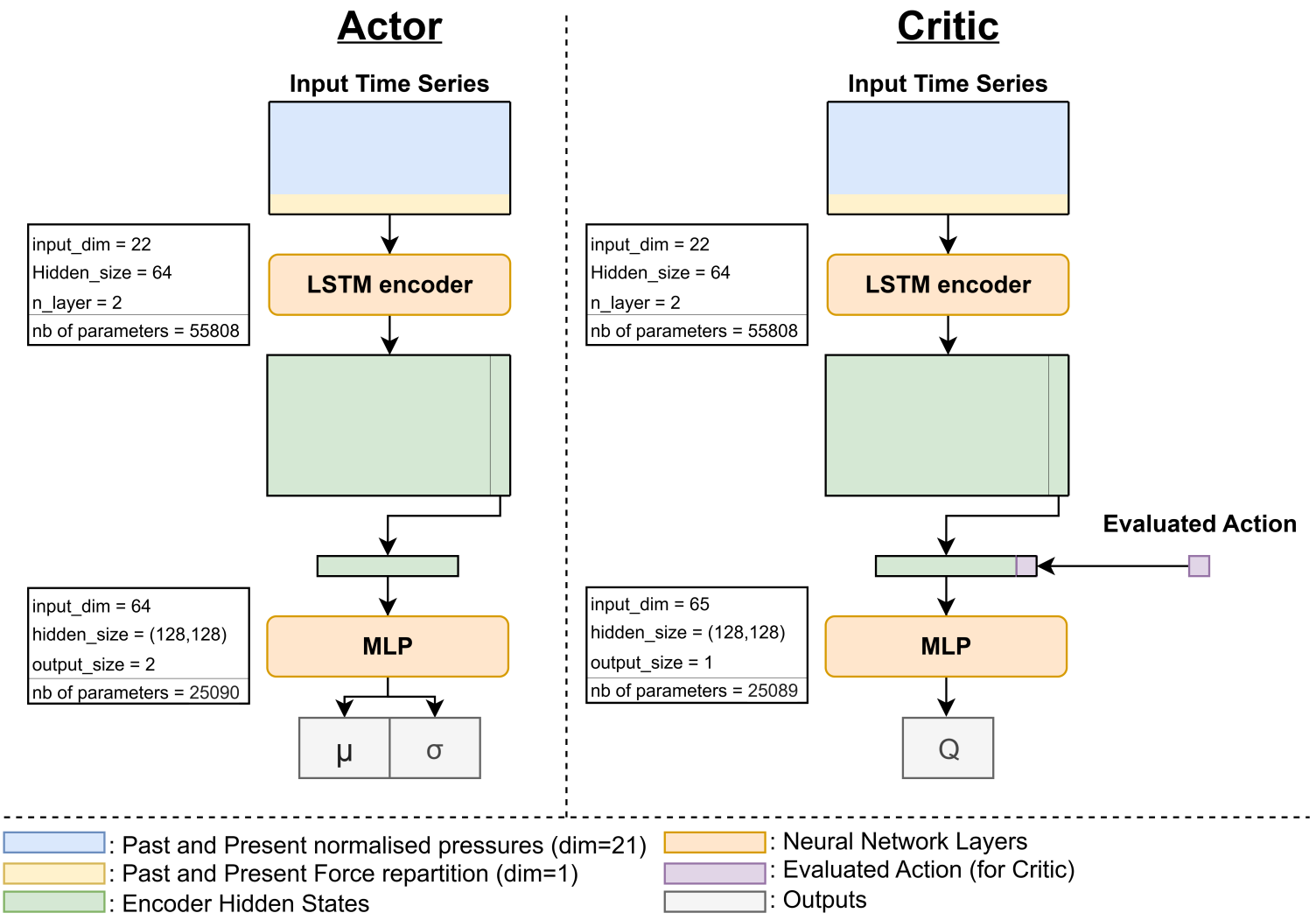


Figure 4.2: Architectures of the actor and critic networks. Both use an LSTM encoder, than the last hidden state is fed to an MLP; in the critic, the action is concatenated with the MLP input.

The use of an LSTM encoder allows the policy and value networks to capture sequential patterns in the evolution of pressures and actions, while keeping the number of parameters moderate. This is particularly advantageous in our setting, where data is limited due to the high computational cost of FEM simulations. In addition, representing the state by the last hidden state of the LSTM provides a fixed-size encoding independent of the episode length, removing the need to define an arbitrary window size while still leveraging long-term temporal dependencies.

Hyperparameters : The training hyperparameters are reported in Table 4.2.

Component	Specification
Optimizers	Adam
Policy lr	$1e^{-4}$
Critic lr	$1e^{-5}$
Entropy lr	$1e^{-4}$
Replay buffer size	$3 * 10^5$ transitions
Batch size	256
Discount factor γ	0.99
Target smoothing coefficient τ	0.005
Updates per step	20
Target update interval	1
Entropy coefficient α	Automatically tuned
Number of parallel environments	8
Training horizon	80000 steps
Maximum episode length	100 timesteps
Average training time	$\sim$ 125 hours (see A.1)

Table 4.2: Hyperparameters used in SAC training.

4.1.5 Reward Formulation

In the previous section, we only outlined the reward at a high level. We now provide its detailed formulation, based on an entropy score complemented by a progress term designed to prevent convergence to static policies.

Normalized entropy score :
Let $\mathbf{p}_i \in \mathbb{R}^n$ denote the pressure values at step i. We smooth and normalize:

$$\tilde{\mathbf{p}}_i \;=\; \frac{\mathbf{p}_i + \varepsilon \mathbf{1}}{\sum_{k=1}^{n}(p_{i,k} + \varepsilon)}, \quad \varepsilon > 0.$$

The entropy and its maximum are

$$H(\tilde{\mathbf{p}}_i) \;=\; -\sum_{k=1}^{n} \tilde{p}_{i,k} \, \log \tilde{p}_{i,k}, \qquad H_{\max} = \log n.$$

The normalized entropy score is then

$$r_i = \left(\frac{H(\tilde{\mathbf{p}}_i)}{H_{\max}} \right)^{\alpha}, \quad r_i \in [0, 1].$$

Here $\alpha > 0$ is a hyperparameter controlling the shaping of the entropy signal: $\alpha > 1$ emphasizes high-entropy regimes, while $\alpha < 1$ smooths differences in lower-entropy regions.

Progress bonus :
From experiments, we observed that using only the entropy score as a reward often led to static policies that did not evolve over time. To address this, we introduce a progress term that rewards improvement and penalizes regress relative to the previous score r_{i-1}. With $d_i = r_i - r_{i-1}$,

$$d_i^+ = \max(d_i, 0), \qquad d_i^- = \max(-d_i, 0),$$

and

$$p_i^+ = \frac{d_i^+}{1 - r_{i-1} + \varepsilon}, \qquad p_i^- = \frac{d_i^-}{r_{i-1} + \varepsilon}.$$

The terms p_i^+ and p_i^- measure the fraction of *headroom* gained or lost: p_i^+ is the relative progress toward the maximum achievable score, while p_i^- quantifies the relative regress compared to what had already been achieved. The progress bonus is then

$$\text{bonus}_i = c\left(p_i^+(1 - r_i) \; - \; p_i^- r_i\right),$$

where $c \geq 0$ is a hyperparameter that controls how strongly progress (or regress) influences the reward compared to the entropy score.

Final reward:
The step reward is

$$R_i = \begin{cases} -1, & \text{if the episode terminates at step } i, \\ r_i + \text{bonus}_i, & \text{otherwise,} \end{cases}$$

which remains in $[0, 1]$ for non-terminal steps. The final reward thus balances the instantaneous uniformity of the pressure distribution (via r_i) with a measure of temporal improvement (via the progress bonus).

4.2 Results

We now turn to the results obtained with the proposed reinforcement learning framework. First, we consider the *reference study*, which consists of optimizing the braking decision on a flat sliding surface as described above. We then extend the analysis to a perturbed surface with sinusoidal deformations, designed to introduce additional complexity and provide a more stringent test of the proposed methodology.

4.2.1 Flat sliding surface

We begin with the flat sliding surface, which constitutes the base control problem studied in this chapter. The goal here is to assess whether the reinforcement learning agent can discover a force repartition strategy that leads to a good pressure distribution in the environment described above. Training follows the setup of Section 4.1.4, and performance is evaluated using the entropy-based reward introduced in Section 4.1.5 with $\alpha = 1$ and $c = 0.2$.

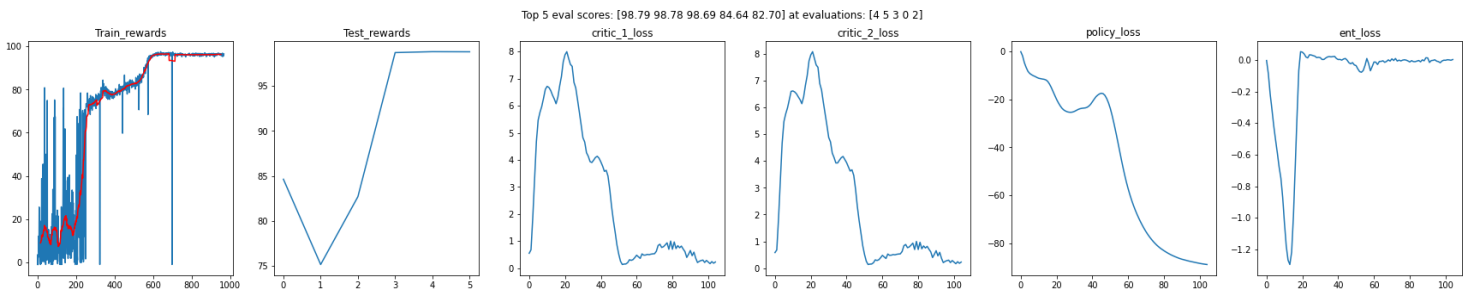
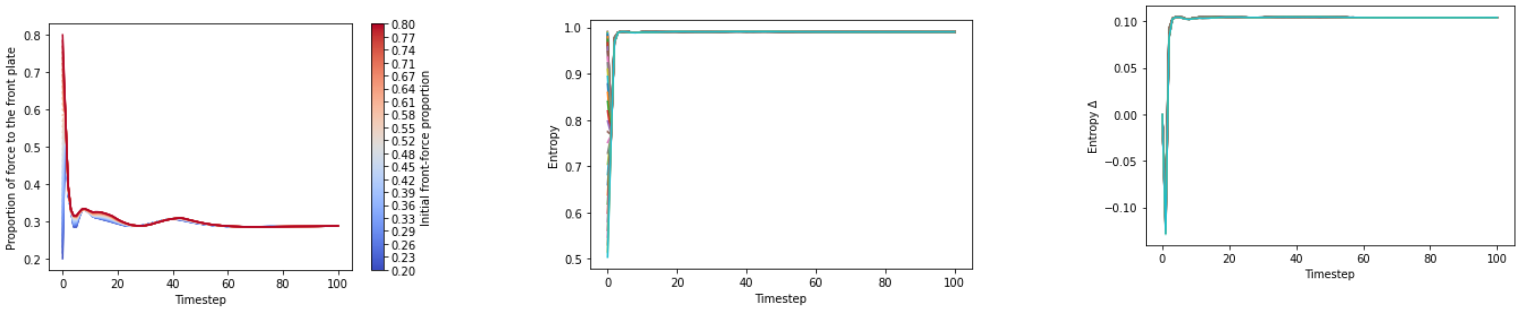


Figure 4.3: Training diagnostics for the flat-surface environment.

As shown in Fig. 4.3, the smoothed training rewards increase rapidly and then plateau near the task maximum of 100. The evaluation rewards closely follow this trajectory, with recent values around 98–99, indicating stable convergence. The critics' losses exhibit the canonical pattern of an initial transient peak followed by decay toward zero, with only minor late-stage oscillations. The policy loss continues to decrease, suggesting residual scope for optimizing the objective; nevertheless, because rewards are already near their ceiling, any further progress is expected to yield diminishing returns and, correspondingly, weaker learning signals. The entropy term stabilizes after an initial dip and rebound, indicating sustained exploration rather than collapse. Since the reward incorporates a normalized-entropy shaping component $\hat{H}^{\alpha}$ (with $\hat{H} \in [0, 1]$), increasing the reward exponent α could accentuate differences among high-performing behaviours, potentially translating modest policy-loss reductions into measurable gains in reward and, ultimately, improved behaviours. We did not however have the time to perform a study on the effect of the reward exponent and simply selected the best performing evaluation from the training above.

We now summarize the learned behaviors across the 40 evaluation initializations in Figure 4.4:



(a) Action vs Timestep

(b) Entropy vs Timestep

(c) Entropy Δ vs Timestep for 50/50 baseline

Figure 4.4: Learned behaviors over the 40 evaluation initialisations and over time.

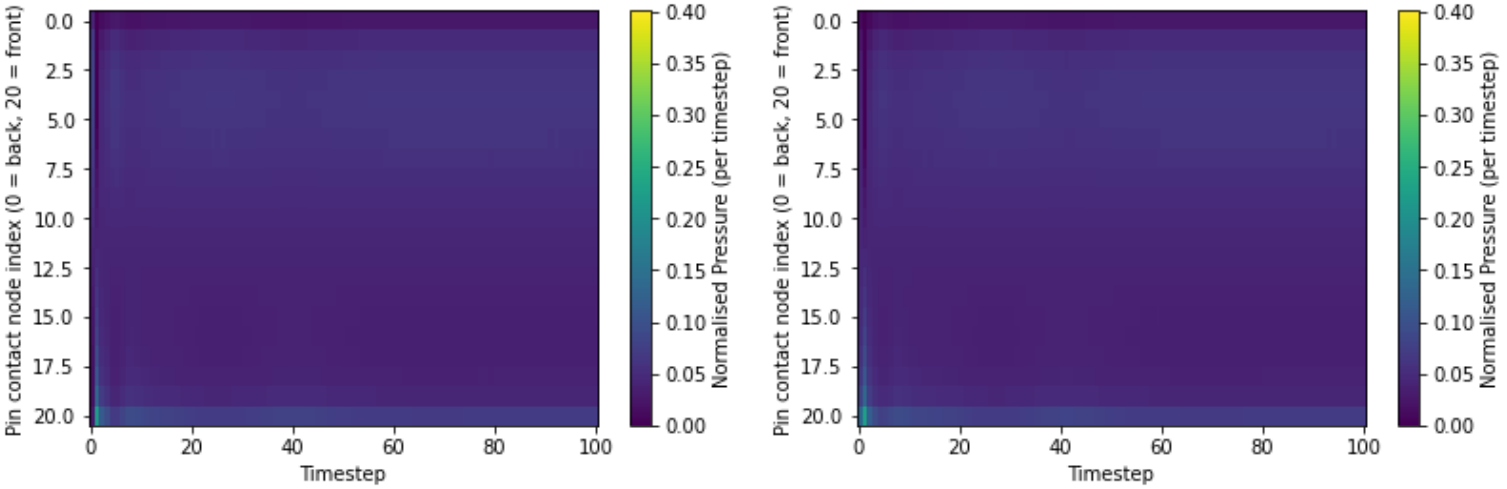
Figure (a) shows that, irrespective of the initial front–plate force proportion, the trajectories converge to essentially the same long–term allocation. Small variations are observed, and these depend smoothly on the initialization.

Figure (b) shows that the entropy of the pressure distribution increases rapidly over the first few timesteps and then stabilizes near its optimum value, remaining steady thereafter.

Figure (c) reports the entropy gain relative to a fixed 50/50 force allocation. The initial Δ is modest—and occasionally negative—but becomes positive within a few timesteps and plateaus close to 0.1.

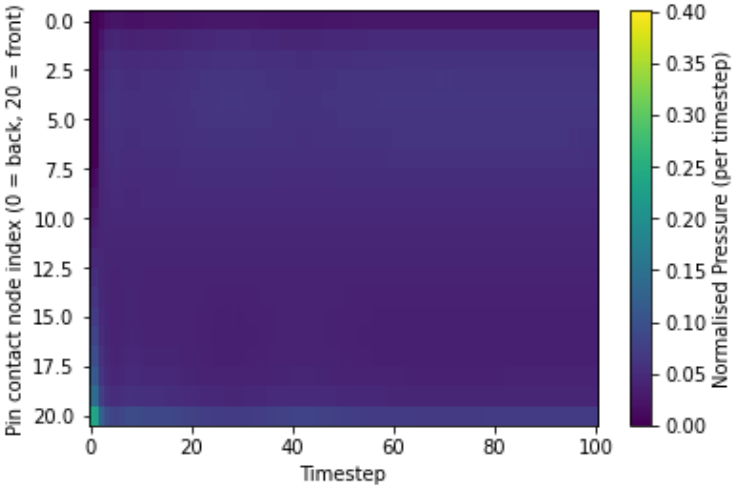
Taken together, these results indicate that the agent produces near–optimal pressure distributions after only a brief transient. However, the nearly identical action–timestep curves suggest that the policy is not strongly adaptive, tending instead to allocate approximately one third of the force to the front plate. At the same time, the smooth dependence on the initialization is encouraging, as it indicates that the learned responses remain physically consistent rather than erratic.

While the entropy curves provide a compact measure of uniformity, they do not reveal how the pressure distribution evolves across the pin surface. To gain more physical insight into the learned behaviors, we therefore examine a few representative examples of pressure evolution under the learned controller in Figure 4.5.

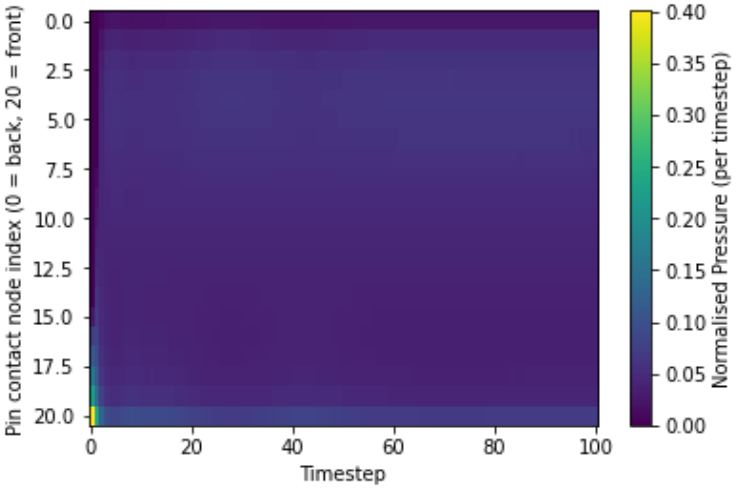


(a) Initial repartition: 20% front

(b) Initial repartition: 40% front



(c) Initial repartition: 60% front



(d) Initial repartition: 80% front

Figure 4.5: Representative examples of pressure distribution evolution over time under the learned controller. Each subplot corresponds to a different initial repartition of the normal force between the two plates, ranging from strongly backloaded (0.2) to strongly frontloaded (0.8).

Across all runs (Fig. (a)–(d)), the spatiotemporal pressure maps exhibit a short transient followed by a highly uniform distribution that remains essentially stationary over time.

Slightly elevated pressures persist at the very front node and along a broad band near the back of the contact surface (approximately pin indices 2–10), indicating localized structure that survives the transient. These features are consistent across initializations.

Visual differences between panels are subtle, which is expected given the similarity of the underlying control trajectories.

Conclusion : The policy displays mild initialization–dependent variation but largely collapses to a single strategy, allocating roughly one third of the total force to the front plate. From a physical perspective, this quasi-constant behavior can be attributed to the limited thermo-mechanical coupling in the environment. Although a coupled thermal analysis is included, the resulting temperature rise remains only a few degrees—even with an artificially increased thermal expansion coefficient. As a consequence, the system behaves almost linearly: pressure variations are small, and feedback between temperature and contact mechanics is weak. Under such conditions, the control problem effectively reduces to a static allocation task, for which a constant force repartition is already near-optimal. To probe adaptability and promote richer behaviors, we therefore consider a perturbed sliding surface in the next experiment.

4.2.2 Perturbation of sliding surface

While the flat-surface experiments demonstrate that reinforcement learning can discover effective and robust strategies, the environment remains relatively simple: once the transient phase is overcome, the controller essentially just allocates a third of the pressure to the front plate. This motivates us to investigate a more demanding setup, where the controller must deal with non-trivial dynamics.

In order to perturb the sliding surface, we impose a sinusoidal deformation. The body is divided into two regions: a flat entry zone for $x < x_c$, and a modulated zone for $x \geq x_c$. Unlike a purely geometric surface corrugation, the deformation is applied to all material points of the body and scales with depth. Denoting by $y_{\min}$ the lowest vertical coordinate, each point at depth y is displaced according to a depth–dependent factor

$$\eta(y) = \frac{y - y_{\min}}{0 - y_{\min}} \in [0, 1],$$

so that the displacement vanishes at the bottom plane ($y = y_{\min}$) and reaches full amplitude at the surface ($y = 0$). The resulting geometry is thus given by

$$z(x, y) = \begin{cases} 0, & x < x_c, \\ \eta(y)\, A \sin\!\left(\frac{2\pi}{\lambda}(x - x_c)\right), & x \geq x_c, \end{cases}$$

where A and λ denote the amplitude and wavelength of the modulation, respectively. This construction introduces a smoothly transmitted sinusoidal perturbation that propagates from the surface into the bulk of the material. Note that the agent does not receive the amplitude or wavelength of the sinusoidal perturbation as part of its state. Instead, it must infer the underlying surface geometry indirectly from the evolving pressure distribution. In practice, this modification increases the diversity of conditions encountered by the agent.

To make use of this increased variability, we adjust the initialization strategy. Specifically, we define ranges for the initialization parameters: (0.2, 0.8) for the first action, (0, 0.3)mm for the amplitude, (50, 150)mm for the wavelength, and $x_c = 30$mm. Based on these ranges, we proceed as follows:

- **During training:** the random initialization of force repartitions is preserved, but now combined with random values of amplitude and wavelength within the specified ranges.

- **During evaluation:** A regular grid with only 40 points in three dimensions would be too sparse and unevenly distributed, making it unsuitable. Instead, we generate 40 initialization points from the Sobol[56] sequence (3D, no scrambling), covering the dimensions of first action, amplitude, and wavelength. These points are then rescaled into the respective parameter intervals to provide a dense and quasi-random coverage of the initialization space.

With this perturbed setup, the control task becomes significantly harder: the agent must not only achieve uniform pressure distributions, but also adapt to continuously varying conditions induced by the surface geometry. The training behavior is shown in Figure 4.6:

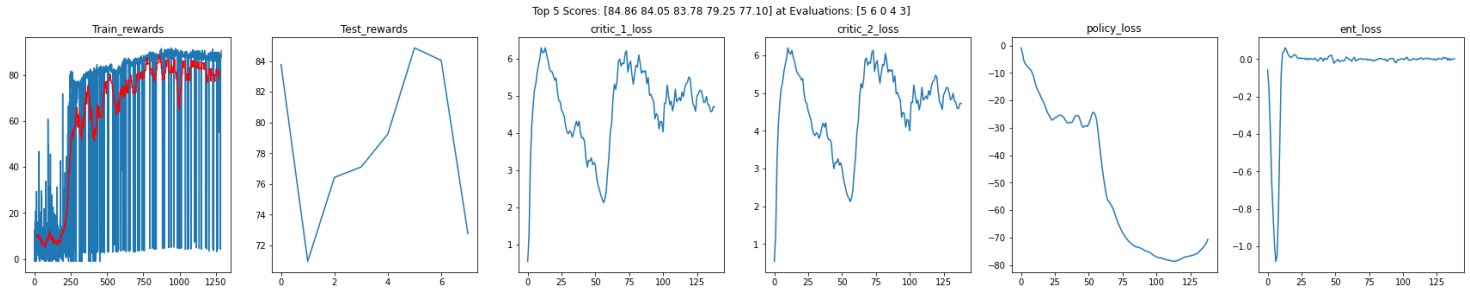
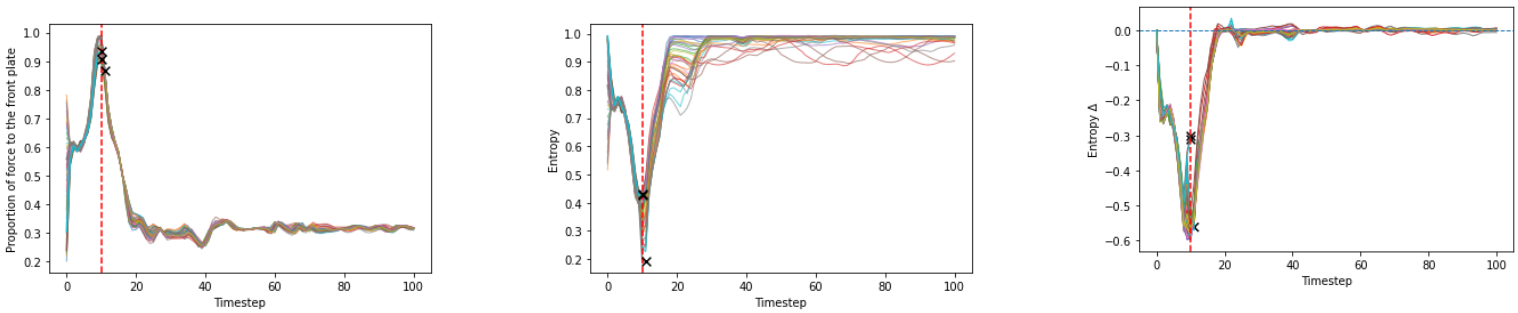


Figure 4.6: Training diagnostics for the perturbed environment.

Figure 4.6 shows the training diagnostics for the sinusoidally perturbed surface. Compared to the flat case, the reward curves are noisier and plateau at lower levels, with evaluation scores stabilizing between 77 and 85. In the best-performing run, three out of forty evaluation episodes terminate prematurely at around timestep 10, immediately after entering the perturbed region—highlighting the increased difficulty of this task.
The critics' losses remain elevated and oscillatory after the initial transient, reflecting the fact that the agent does not observe the perturbation parameters and must infer them indirectly from pressures. The policy loss decreases overall but with intermittent sharp drops, and the entropy term stabilizes after an early dip. Taken together, these curves suggest that the agent does learn to exploit some patterns in the environment, but it remains unclear whether this translates into genuinely non-trivial control strategies. To address this question, we now examine the evaluation behaviors.



(a) Action vs timestep

(b) Entropy vs timestep

(c) Entropy Δ vs timestep (relative to 33/67 baseline)

Figure 4.7: Learned behaviors over the 40 evaluation initializations. The red dashed line marks the entry into the perturbed surface region; black crosses indicate premature terminations.

Figure 4.7 provides this assessment. Figure (a) shows that, irrespective of initialization, all trajectories converge to approximately the same force split (33/67) after a brief $\sim$ 20-timestep transient in which the front plate is heavily overloaded upon entering the perturbed

zone. Figure (b) confirms that entropy drops sharply during this transient but then recovers to values close to one. Figure (c), which compares the agent against a fixed 33/67 baseline, highlights that the agent consistently underperforms during the transient phase before converging to parity. Post-transient mean entropies are virtually identical: 0.9736 for both the agent and the baseline.

From a physical perspective, the convergence toward a constant 33/67 split likely results from the modest amplitude of the imposed perturbations. The induced pressure and displacement gradients remain weak, keeping the system close to the quasi-linear regime already observed for the flat surface. In such conditions, as shown in the previous experiment, the optimal control is effectively static, so a near-constant force repartition is expected. However, this does not explain the poor transient behavior observed at the start of the episodes. It remains unclear why this occurs, although the training curves suggest that learning was still ongoing toward the end of the run. It is therefore possible that the training duration was insufficient for the policy to fully stabilize its transient response.

Improving upon this result would likely require changes both to the physical setup and to the training procedure. On the physical side, increasing the perturbation amplitude or introducing more pronounced geometric or thermal variations would create a stronger control challenge, encouraging the agent to develop genuinely adaptive strategies. On the algorithmic side, possible avenues include curriculum learning (gradually increasing perturbation severity or episode length), the use of larger or more expressive models, or teacher–student approaches in which a privileged agent with access to hidden perturbation parameters guides the training of a restricted agent—for example, through policy distillation via KL minimization.

4.3 Conclusion

In this chapter, we introduced a finite element (FEM) setup as a proof of concept for applying reinforcement learning to brake tribology contact conditions. The framework provides a simplified yet physically grounded environment that enables systematic exploration of contact dynamics without the prohibitive cost of experimental campaigns. Within this setting, we formulated a control problem aimed at regulating the contact pressure distribution through force allocation and trained Soft Actor–Critic (SAC) agents using an entropy-based reward.

On a flat sliding surface, the approach proved successful: the agent systematically outperformed a naive 50/50 split and smoothly converged to a near-optimal 33/67 allocation. This constant control strategy is physically consistent with the weak thermo-mechanical coupling of the simplified environment : temperature rises remain small, feedback effects are minimal, and a static allocation is effectively optimal. These results demonstrate that reinforcement learning can uncover meaningful and robust strategies directly from interaction data, validating the viability of the proposed setup under quasi-linear conditions.

On a perturbed sliding surface, however, the picture was less favorable. While the agent again converged to the 33/67 allocation in steady state, it consistently underperformed a simple static baseline during the initial transient and failed to exhibit genuinely adaptive behavior. Physically, this outcome likely reflects the modest amplitude of the imposed perturbations, which did not generate sufficiently strong gradients to alter the system's response, making a constant allocation strategy remain near-optimal. Algorithmically, the poor transient performance may indicate that training had not fully converged. Future studies could therefore

explore larger perturbation amplitudes or longer training durations to expose the agent to stronger dynamic variations.

These findings open several paths for future work. On the methodological side, more challenging environments and curriculum-based training protocols could help elicit stronger strategies and provide a clearer demonstration that reinforcement learning can be effectively applied to brake control. On the formulation side, alternative reward functions—particularly if predictive models such as those developed in Chapter 2 become more accurate—could shift the objective from pressure uniformity toward application-relevant metrics such as emission reduction. On the experimental side, the same approach could be deployed in real time on physical test benches, moving from simulation-only demonstrations toward practical validation.

Taken together, these directions would strengthen the case for reinforcement learning as a tool for exploring and optimizing complex frictional systems.

General conclusions

This thesis has investigated how machine learning can contribute to the study and optimization of brake tribology, addressing key challenges of prediction, interpretation, and control. Two parts of the work relied on a multimodal experimental dataset combining acoustic, thermal, mechanical, and particle-emission measurements, which enabled the development of predictive and interpretative models of emissions and acoustic mechanisms. A third part, by contrast, employed a finite-element brake-like simulator as a proof of concept for reinforcement learning–based control strategies. These different avenues show that machine learning can adapt to the nature of the available information, whether derived from physical experiments or numerical models, and open distinct paths toward advancing tribological research.

Our first contribution focused on the prediction of braking emissions, both acoustic and particulate. Unlike most previous studies, which target only aggregate outcomes, we explored time-resolved acoustic spectra and size-resolved particle distributions. To our knowledge, this formulation is novel in the tribology literature and demonstrates that machine learning can address emissions at a finer granularity than commonly attempted. These results underline the potential of data-driven models to act as virtual sensors, offering real-time estimations of emissions that could one day support both laboratory testing and vehicle-level monitoring.

The second contribution moved beyond prediction toward interpretation. By applying clustering and classification techniques to spectrograms of acoustic emissions, complemented by explainability methods such as permutation feature importance and integrated gradients, we identified recurring acoustic states and examined the physical variables that appear to influence them. This represents a first step toward addressing one of the main limitations of existing machine learning approaches in tribology: the lack of interpretability. While preliminary, these results suggest that unsupervised and explainable ML methods could provide new insights into the mechanisms behind squeal and other noise phenomena, and they point to a possible path for bridging the gap between raw signal analysis and tribological understanding.

The third contribution investigated control strategies through reinforcement learning applied to a finite-element brake-like simulator. In this proof-of-concept study, our reinforcement learning algorithms were able to find acceptable solutions for stabilizing pressure distributions on a flat sliding surface, showing that adaptive control can emerge from data-driven training. On perturbed surfaces, however, the learned policies did not surpass a constant baseline. This outcome points to avenues for improvement in algorithm design and training environments. If such improvements can be achieved, reinforcement learning could become a compelling candidate for exploring adaptive control strategies in more realistic tribological systems.

Taken together, these three axes—prediction, interpretation, and control—show how machine learning can expand the methodological landscape of brake tribology. They directly address

current research gaps in granularity, interpretability, and methodological breadth, and illustrate the variety of applications through which ML can help advance the understanding and improvement of brake tribology.

Naturally, the work presented here also has limitations. The prediction- and interpretation-oriented models were trained on a laboratory-scale dataset, which, although multimodal, cannot fully capture the diversity of in-use braking conditions. The reinforcement learning study, for its part, relied on a highly simplified finite-element simulator, far from the complexity of real brake systems. Also, across all three axes—prediction, interpretation, and control—the performance remains below the threshold required for industrial deployment. These constraints reflect the exploratory nature of the research and point to the need for larger datasets, more realistic simulations, and closer integration with mechanical understanding of brake systems.

Looking ahead, several perspectives emerge. At the methodological level, incorporating physical knowledge throughout the learning pipeline—from data formatting and feature engineering to model design through constraints, coupled mechanical formulations, or physics-informed neural networks—offers a promising way to balance interpretability and predictive power. At the experimental level, scaling up data acquisition—whether through more diverse laboratory campaigns or instrumented vehicles—would enable models to generalize across materials, environments, and duty cycles. At the application level, embedding ML into vehicle monitoring systems could provide real-time diagnostics of noise and emissions, helping manufacturers and regulators meet tightening environmental and comfort standards. Finally, reinforcement learning and other control-oriented methods could evolve into digital co-pilots for braking systems, enhancing their stability and adaptability in real-world conditions.

In conclusion, this thesis demonstrates that machine learning, when carefully integrated with tribological knowledge and experimental rigor, can contribute to the prediction of emissions, the understanding of mechanisms, and the control of braking systems. While much remains to be done before such approaches reach industrial maturity, the results presented here establish a foundation and open promising directions for future research at the interface of brake tribology and data science.

Bibliography

[1] Farhad Ahmadijokani, Akbar Shojaei, Mohammad Arjmand, Yasaman Alaei, and Ning Yan. Effect of short carbon fiber on thermal, mechanical and tribological behavior of phenolic-based brake friction materials. *Composites Part B: Engineering*, 168, 12 2018.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[3] Dragan Aleksendrić and David Barton. Neural network prediction of disc brake performance. *Tribology International*, 42:1074–1080, 07 2009.

[4] D. Antanaitis. Application of machine learning models to enable virtual development of high-performance brake systems. In *SAE Technical Paper*, 2024.

[5] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.

[6] L. Bălăsoiu, M. Ilie, D. Rusu, and F. Marin. Artificial neural networks for predicting tribological behavior of polymer composites. *Polymers*, 16(24):3588, 2024.

[7] F. Bonini. Braking torque estimation through machine learning algorithms. In *Materials Research Proceedings, AIMETA*, volume 26, pages 213–218, 2023.

[8] F. Bonini, A. Rivola, and A. Martini. Estimation of braking torque for motogp class motorcycles with carbon braking systems through machine learning algorithms. In *IEEE International Workshop on Metrology for Automotive*, pages 1–6, 2021.

[9] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[10] J. Cao, J. Bao, Y. Yin, W. Yao, T. Liu, and T. Cao. Intelligent prediction of wear life of automobile brake pad based on braking conditions. *Industrial Lubrication and Tribology*, 75(2):157–165, 2022.

[11] Wei Chen, Jiliang Mo, Renxia Wang, Zhicheng He, Chunguang Zhao, and Song Zhu. Fully coupled thermo-mechanical-wear analysis for brake interface of high-speed train. *Wear*, 556-557:205510, 07 2024.

[12] Kyunghyun Cho, Bart Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 06 2014.

[13] A. Choudhuri and P. Shekhar. Predicting pad wear with ml. In *EuroBrake Conference*, 2020.

[14] Y. Desplanques et al. Analysis of tribological behaviour of pad–disc contact in railway braking. part 1. laboratory test development, compromises between actual and simulated tribological triplets. *Wear*, 2007.

[15] Mikael Eriksson, Fredrik Bergman, and Staffan Jacobson. On the nature of tribological contact in automotive brakes. *Wear*, 252(1-2):26–36, 2002.

[16] Mikael Eriksson and Staffan Jacobson. Tribological surfaces of organic brake pads. *Tribology International*, 33(12):817–827, 2000.

[17] C. Cotici N. Gheorghita Andrei-Florin Hristache Daiana Alina Ionescu G. Ipate, A. Cristescu. Evaluation of the tribological behavior of a brake disc-pad friction pair using a fuzzy inference model based on an adaptive network (anfis). *echnium: Romanian Journal of Applied Sciences and Technology*, 14:9695, 2023.

[18] C. Geier, F. Hoffmann, M. Stender, and P. Dufrenoy. Machine learning-based state maps for complex dynamical systems: applications to friction-excited brake vibrations. *Nonlinear Dynamics*, 111(21):22137–22151, 2023.

[19] Theodoros Grigoratos and Giorgio Martini. Brake wear particle emissions: a review. *Environmental science and pollution research international*, 22, 10 2014.

[20] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

[21] J. Han, H. Ling, X. Sun, and L. Zou. Commercial vehicle disc brake temperature prediction model construction based on machine learning. In *Journal of Physics: Conference Series*, volume 2825, page 012016, 2024.

[22] M. S. Hasan, A. Kordijazi, P. K. Rohatgi, and M. Nosonovsky. Triboinformatic modeling of dry friction and wear of aluminum base alloys using machine learning algorithms. *Tribology International*, 161:107065, 2021.

[23] Holger Hetzler and Kai Willner. On the influence of contact tribology on brake squeal. *Tribology International*, 46:237–246, 2012.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[25] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933.

[26] H.J. Hwang, S.L. Jung, K.H. Cho, Y.J. Kim, and H. Jang. Tribological performance of brake friction materials containing carbon nanotubes. *Wear*, 268(3):519–525, 2010.

[27] N. Jardine and R. Sibson. The construction of hierarchic and non-hierarchic classifications. *The Computer Journal*, 11(2):177–184, 08 1968.

[28] R. Jegadeeshwaran and V. Sugumaran. Fault diagnosis of automobile hydraulic brake system using statistical features and svm. *Mechanical Systems and Signal Processing*, 52–53:436–446, 2015.

[29] Zhencai Zhu Yan Yin Jiusheng Bao, Minming Tong. Intelligent tribological forecasting model and system for disc brake. In *Proceedings of the 24th Chinese Control and Decision Conference (CCDC)*, pages 3104–3109, 2012.

[30] Mohamed Kchaou, Amira Sellami, Jamel Fajoui, Rafal Kus, and Riadh Elleuch. Tribological performance characterization of brake friction materials: What test? what coefficient of friction? *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, 2019.

[31] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115, 2017.

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[33] Daniel Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, 11 1999.

[34] H. Li, S. Liu, Y. Liu, and Q. Zhou. Tribological properties study and prediction of qbe2 beryllium bronze and 7075-t6 aluminium alloy pairs under grease lubrication. *Proceedings of the IMechE, Part J: Journal of Engineering Tribology*, 2025.

[35] J. Li, S. Jiang, M. Li, and J. Xie. A fault diagnosis method of mine hoist disc brake system based on machine learning. *Applied Sciences*, 10(5):1768, 2020.

[36] F. Limmer, D. Barton, C. Gilkeson, P. Brooks, and S. Kosarieh. Development of a small-scale test bench for innovative braking systems. In *EuroBrake 2021 Technical Programme*, 2021.

[37] F. Limmer, P. Brooks, C. Gilkeson, and S. Kosarieh. Tribo-oxidation of a brake friction couple under severe thermal conditions. *Tribology International*, 2023.

[38] Yezhe Lyu, Francesco Varriale, Vilhelm Malmborg, Martin Ek, Joakim Pagels, and Jens Wahlström. Tribology and airborne particle emissions from grey cast iron and wc reinforced laser cladded brake discs. *Wear*, 556-557:205512, 2024.

[39] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[40] V. Magnier, J.F. Brunel, and P. Dufrénoy. Impact of contact stiffness heterogeneities on friction-induced vibration. *International Journal of Solids and Structures*, 51(7-8):1662–1669, 2014.

[41] James B. McDonald and Yexiao J. Xu. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1):133–152, 1995.

[42] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[43] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[44] Nikzad Motamedi. *Vers la prédiction et la compréhension des effets tribologiques sur les performances systèmes par l'intelligence artificielle*. PhD thesis, 2023. Thèse de doctorat dirigée par Magnier, Vincent et Wannous, Hazem Mécanique, énergétique, génie des procédés, génie civil Université de Lille (2022-....) 2023.

[45] G. P. Ostermeyer. New insights into the tribology of brake systems. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2008.

[46] J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 03 1983.

[47] Leonardo Pelcastre, Lisa-Marie Weniger, and Jens Hardell. On the low temperature tribological behaviour of brake block materials for railway applications under dry and icy conditions. *Wear*, 523:204764, 2023.

[48] Danishtah Quamar and Chiranjit Sarkar. Modelling of performance parameters of phenolic base resins non-asbestos organic (nao) friction material in brake pad using machine learning algorithms. *Tribology International*, 191:109188, 2024.

[49] Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.

[50] Domenico Antonio Rita, Stefano Candeo, Priyadarshini Jayashree, Ana Paula Gomes Nogueira, Emiliano Rustighi, and Giovanni Straffelini. Comparative analysis of pin-on-disc and inertia-dynamometer sliding tests on a friction material. *Wear*, 558-559:205552, 2024.

[51] A. Rosenkranz, M. Marian, F. J. Profito, N. Aragon, and R. Shah. The use of artificial intelligence in tribology—a perspective. *Lubricants*, 9(1):2, 2021.

[52] H. Sellami, A. Ouhrouche, A. Benaicha, and I. Boukli-Hacene. Explainable multi-target regression for simultaneous prediction of wear and friction coefficient. *Acta Polytechnica Hungarica*, 21(11):151–167, 2024.

[53] Manoharan Sembian, Vijay R, D. Lenin Singaravelu, and Mohamed Kchaou. Experimental investigation on the tribo-thermal properties of brake friction materials containing various forms of graphite: A comparative study. *Arabian Journal for Science and Engineering*, 44, 10 2018.

[54] Saravanakumar Sengottaiyan, Sathiyamurthy Subbarayan, Vinoth Viswanathan, and Pathmanaban Pugazhendi. Optimized machine learning with hyperparameter tuning and response surface methodology for predicting tribological performance in biocomposite materials. *Polymer Composites*, 45:9421–9439, 04 2024.

[55] Vishal Reddy Singireddy, Rohit Jogineedi, Sai Kalyan, et al. On scaled-down bench testing to accelerate the study of disc wear. *Tribology International*, 2022.

[56] I.M Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967.

[57] K. Song, L. Chen, Z. Wang, and H. Xu. Brake noise cae prediction enhanced by machine learning. In *SAE Technical Paper*, 2025.

[58] A. T. Sose, S. Y. Joshi, L. K. Kunche, F. Wang, and S. A. Deshmukh. A review of recent advances and applications of machine learning in tribology. *Physical Chemistry Chemical Physics*, 25(3):2368–2386, 2023.

[59] M. Stender, P. Dufrenoy, and J. Kuhlmann-Wilsdorf. Nonlinear brake vibrations analyzed by recurrence quantification and machine learning. *Nonlinear Dynamics*, 98:1189–1205, 2019.

[60] M. Stender, P. Dufrenoy, and J. Kuhlmann-Wilsdorf. Deep learning for brake squeal detection and characterization. *Mechanical Systems and Signal Processing*, 149:107181, 2020.

[61] Steffen Sturm. Universal brake disc analysis with new high-speed thermographic systems for automated test bench solutions. In *EuroBrake 2021 Technical Programme*, 2021.

[62] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[63] F. Varriale, S. Candeo, G. Riva, J. Wahlström, and A. Wahlström. A brake system coefficient of friction estimation using 3d pvt maps. *Lubricants*, 10(7), 2022.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[65] Cédric Villani. *Optimal Transport: Old and New*. Springer, 2008.

[66] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey, 2023.

[67] Jin-Kun Xiao, Shu-Xian Xiao, Juan Chen, and Chao Zhang. Wear mechanism of cu-based brake pad for high-speed train braking at speed of 380 km/h. *Tribology International*, 150:106357, 2020.

[68] Yelong Xiao, Y. Cheng, H. Zhou, W. Liang, M. Shen, P. Yao, H. Zhao, and G. Xiong. Evolution of contact surface characteristics and tribological properties of a copper-based sintered material during high-energy braking. *Wear*, 488-489:204163, 2022.

[69] M. Yang, W. Jiang, J. Bao, and C. Zhang. Complex modal optimization of the disk brake based on thermal–structural coupling. *Mechanics Based Design of Structures and Machines*, 52(12):10422–10438, 2024.

[70] Z. Yang, Y. Zhou, L. Wang, and H. Zhang. Reliability-based robust optimization design of vehicle braking systems under multiple failure modes based on high-precision surrogate models. *Proceedings of the IMechE, Part O: Journal of Risk and Reliability*, 2025.

[71] Jiabao YIN, Chun LU, and Jiliang MO. Comprehensive modeling strategy for thermo-mechanical tribological behavior analysis of railway vehicle disc brake system. *Friction*, 12(1):74–94, 2024.

[72] N. Yin, Y. Xu, X. Zhao, J. Liu, and H. Zhang. Frictional signal processing and machine learning-based fault diagnosis and forecasting of braking systems. *Mechanical Systems and Signal Processing*, 226:112349, 2025.

[73] N. Yin, P. Yang, S. Liu, S. Pan, and Z. Zhang. Artificial intelligence in tribology: present and future. *Friction*, 2024.

[74] Xin Zhang, Yongzhen Zhang, Sanming Du, Tiantian He, and Zhenghai Yang. Influence of braking conditions on tribological performance of copper-based powder metallurgical braking material. *Journal of Materials Engineering and Performance*, 27, 07 2018.

Appendix A

Technical Appendix

A.1 Compute capabilities, software and compute usage

This section reports the hardware/software environment used across all experiments and the compute usage per task. Table A.1 summarizes the environment; Table A.2 summarizes per-task usage.

Hardware	
GPU	NVIDIA A100 (PCIe), 80 GB VRAM
CPU	2× Intel Xeon Gold 5317 @ 3.00 GHz
RAM	256 GB
Storage	NVMe SSDs
Software	
GPU Driver/CUDA	NVIDIA driver 570.133.20 (CUDA 12.8)
PyTorch	2.2.1
RAPIDS cuML	23.8.0 (`cuml-cu12`)
ANSYS	23.1

Table A.1: Hardware and software configuration

Runtime and memory usage per task				
Task	**Time taken (hh:mm:ss)**	**Peak RAM (GB)**	**Peak VRAM (GB)**	**Notes**
Chapter 2				
Single-Target Pollution Prediction				
Train + Eval	69:18:30	∼80	3.85 (per MIG)	3-way MIG: 2g.20gb, 2g.20gb, 2g.20gb
PFI	13:51:30	∼80	1.45 (per MIG)	3-way MIG: 2g.20gb, 2g.20gb, 2g.20gb
Multi-Target Pollution Prediction				
Train + Eval	*151:33:20*	∼80	4.3 (per MIG)	3-way MIG: 2g.20gb, 2g.20gb, 2g.20gb
Chapter 3				
55 dB threshold				
Clustering	00:24:20	∼12	3.20	no parallel; Optuna trials sequential
Classifier training & eval	08:28:24	∼45	4.56	no MIG
PFI	05:24:40	∼45	0.97	no MIG
IG computations	20:00:36	∼45	22.17	no MIG
75 dB threshold				
Clustering	05:20:40	∼12	3.49	no parallel; Optuna trials sequential
Classifier training & eval	06:58:24	∼45	0.38	no MIG
PFI	02:37:30	∼45	0.18	no MIG
IG computations	03:18:24	∼45	3.42	no MIG
Chapter 4				
RL training (flat)	∼125:00:00	∼12	0.50	time spent mostly computing FEM simulations
RL training (perturbed)	∼125:00:00	∼12	0.50	time spent mostly computing FEM simulations

Table A.2: Runtime and memory usage per task, grouped by chapter.

A.2 Balancing Cross-Validation Splits Using Wasserstein Distance

Standard k-fold cross-validation is a widely used practice for model evaluation. However, it inherently assumes that the dataset is independently and identically distributed (i.i.d.), and that input parameter distributions are reasonably balanced across folds. In our case, this assumption does not hold: as shown in Figure 1.3, several experimental parameters—such as contact duration and initial disc temperature—exhibit strong skewness and saturation. These imbalances can introduce systematic bias, whereby certain operating conditions are overrepresented in some folds and underrepresented or entirely absent in others.

To mitigate this issue, we adopt a distribution-aware partitioning strategy based on the Wasserstein distance [65], also known as the Earth Mover's Distance. This metric quantifies the minimum "effort" required to transform one probability distribution into another, taking into account the geometry of the parameter space. Unlike other divergences (e.g., KL divergence or Jensen–Shannon), the Wasserstein distance remains well-defined even when the distributions have disjoint supports—an important property in small or sparse datasets.

Although an exact optimization method for minimizing the Wasserstein distance between folds may exist, we employ a pragmatic approach: we generate a large number of random splits and select the one with the lowest Wasserstein distance across key input parameter distributions. While this heuristic does not guarantee a globally optimal solution, it proves effective in practice, especially when the number of candidate splits is sufficiently high.

Admittedly, this approach introduces a degree of bias, as it selects folds based on input distribution similarity rather than pure randomness. However, it remains the most practical solution we've found for achieving robust and stable model evaluation in the presence of unbalanced parameter distributions. In the long term, expanding the dataset to cover a broader parameter space would reduce the need for such corrective strategies.

A.3 Cross-Validation Splits and Test Parameter Distributions

To ensure that our models generalize across different types of pollution-related objectives, we defined several cross-validation strategies. Each strategy focuses on maintaining balanced test parameter distributions given the available data specific to a given objective. Below, we present the different cross-validation splits, along with visualizations of their respective test parameter distributions:

A.3.1 Sound-Focused Split

This cross-validation split is designed for scenarios where the objective is based exclusively on sound pollution, as measured by the microphone data. After applying outlier removal, a total of 472 valid tests remain. The resulting test parameter distributions cdfs per-split are shown below in Figure A.1:

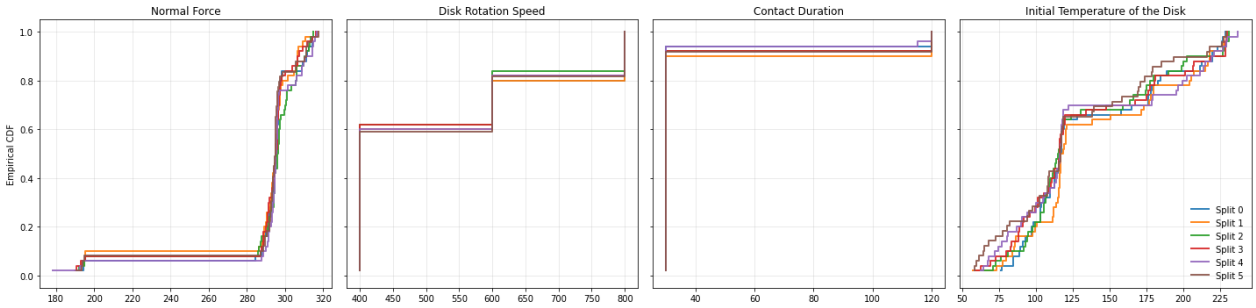


Figure A.1: Sound focused split test parameters

A.3.2 EEPS Focused Split

This cross-validation split is designed for scenarios where the objective is based exclusively on nm-scale particle emission, as measured by the EEPS sensor. After applying outlier removal and taking out the tests where EEPS data is unavailable, a total of 337 valid tests remain. The resulting test parameter distributions cdfs per-split are shown below in Figure A.2:

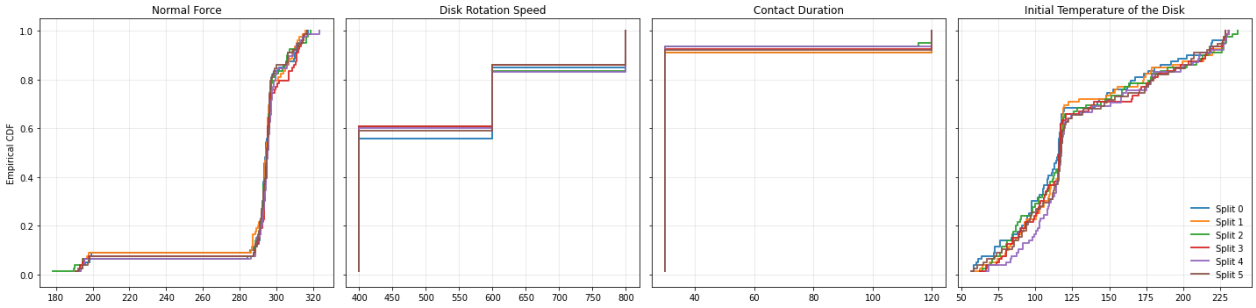


Figure A.2: EEPS focused split test parameters

A.3.3 OPS Focused Split

This cross-validation split is designed for scenarios where the objective is based exclusively on μm-scale particle emission, as measured by the OPS sensor. After applying outlier removal and taking out the tests where OPS data is unavailable, a total of 299 valid tests remain. The resulting test parameter distributions cdfs per-split are shown below in Figure A.3:

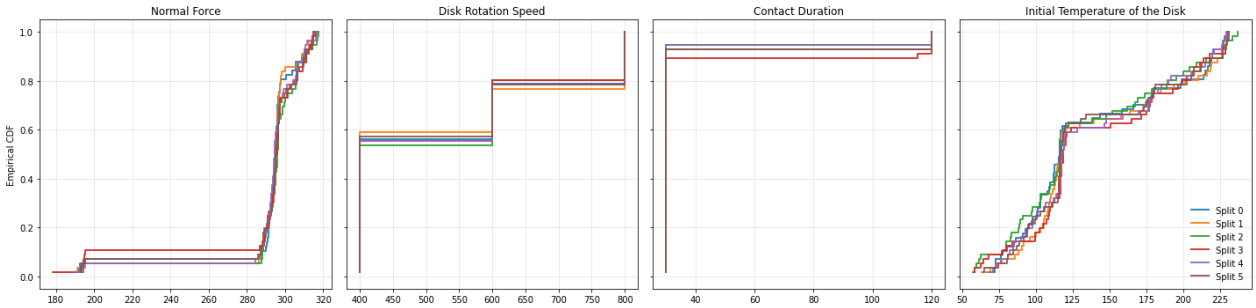


Figure A.3: OPS focused split test parameters

A.3.4 Global Split (All Objectives)

This final cross-validation split considers all objectives simultaneously: sound pollution, nanoscale particle emissions (EEPS), and microscale particle emissions (OPS). While sound data has been consistently recorded, the EEPS and OPS measurements are occasionally missing. To account for this, two additional binary test parameters—*"has EEPS"* and *"has OPS"*—were introduced. Alongside the usual test parameters, these indicators are distributed as evenly as possible across the splits to ensure a balanced evaluation. The resulting test parameter distributions for each split are presented below in Figure A.4:

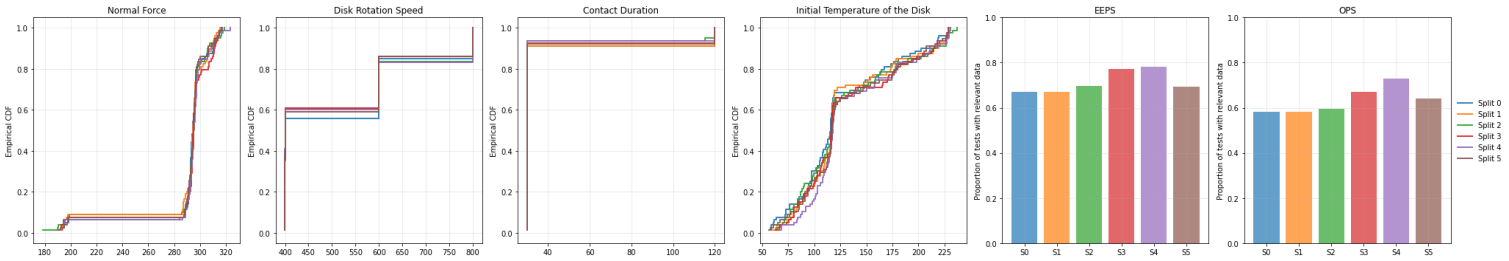


Figure A.4: Split test parameters

A.4 Sound Spectrograms formatting

During this thesis, we frequently worked with sound data, which is often noisy and lacks expressiveness in its raw form. To enhance the signal's expressivity, we applied the Short-Time Fourier Transform (STFT) that converted the raw sound signals into spectrograms. The STFT was computed using parameters commonly applied in the lab: a FFT size of 1024 and a step size of 10.

These settings offer a good balance between frequency resolution and time windowing, allowing for a clear representation of the signal without overloading models with excessive information.

Finally, to reduce noise from the test bench motor and other lab equipment, we discarded frequencies below 500 Hz, as they are believed to primarily capture these non-signal sources. In addition, we experimented with various denoising strategies at the spectrogram level (e.g., *Wavelet, Wiener, thresholding*) in an attempt to suppress irrelevant background components. However, none of these methods produced clean signals consistently and often introduced artifacts, which could degrade downstream performance more than the original background noise itself.

A.5 Last Observation Carried Forward

The Last Observation Carried Forward (LOCF) method is a simple technique used to handle missing values in time series data. It replaces any missing value with the last observed value prior to the missing point. This method is particularly useful in preventing data leakage by ensuring that future information is not used for imputation. Mathematically, if $y(t_m)$ is missing, it is replaced by $y(t_{m-1})$, where t_{m-1} is the last valid observation before t_m. Figure A.5 illustrates how it doesn't leak data compared to Linear interpolation.

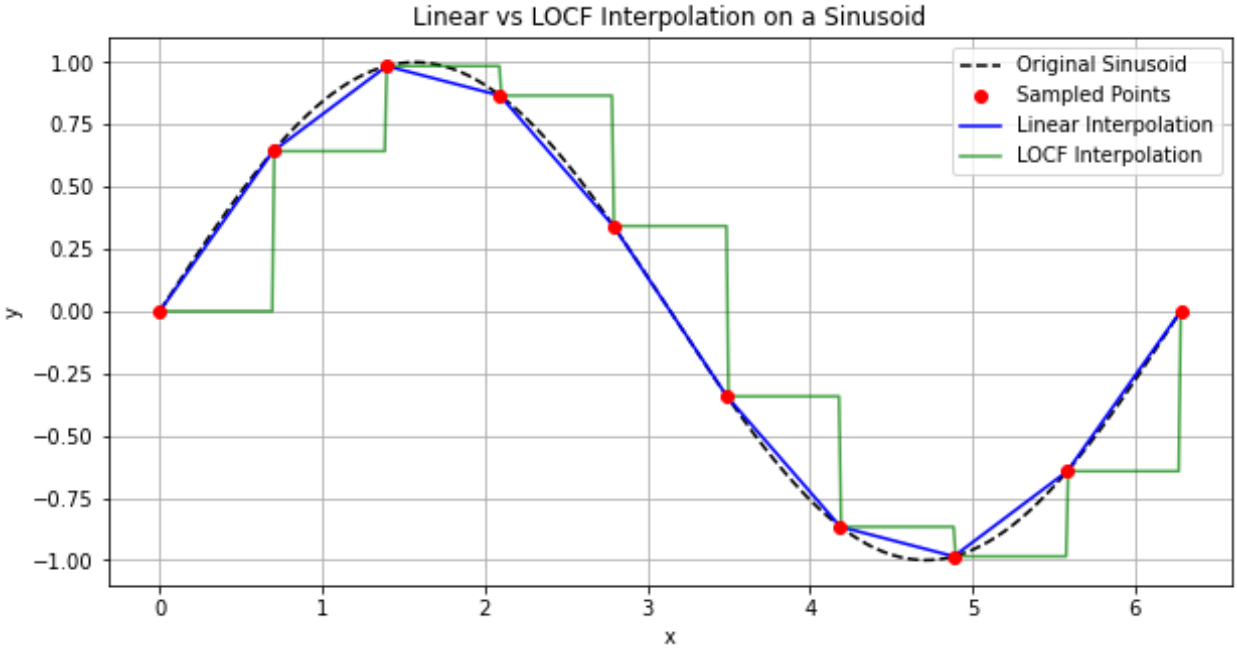


Figure A.5: Linear vs LOCF Interpolation on a Sinusoid

While computationally efficient, the method assumes that the missing values are similar to the last observed value, which may not always be accurate. LOCF is ideal when we want to prevent data leakage at all cost, but it may introduce bias if trends or seasonality are present in the data.

A.6 Standard Deep Learning Data Scalers

Most deep learning methods achieve better and faster results when scaling both inputs and outputs, as they are optimized using gradient descent. Table A.3 is a brief list of common data scaling techniques :

Data Scaler	**Formula**	**Parameter**	**Use Case**
StandardScaler	$X' = \frac{X-\mu}{\sigma}$	μ: Mean σ: Standard deviation	When data follows a normal distribution.
MinMaxScaler	$X' = \frac{X-X_{\min}}{X_{\max}-X_{\min}}$	$X_{\min}$: Minimum value $X_{\max}$: Maximum value	When data has known bounds and needs to be scaled to [0,1].
AbsMaxScaler	$X' = \frac{X}{\lvert X\rvert_{\max}}$	$\lvert X\rvert_{\max}$: Absolute max value	When preserving the sign and relative magnitude is important.

Table A.3: Common Data Scalers, Their Parameters, and Use Cases

A.7 Time Series Forecasting Architectures

The following schematics illustrate single-timestep (See A.6) and sequence-to-sequence (See A.7) forecasting models training and evaluation procedures:

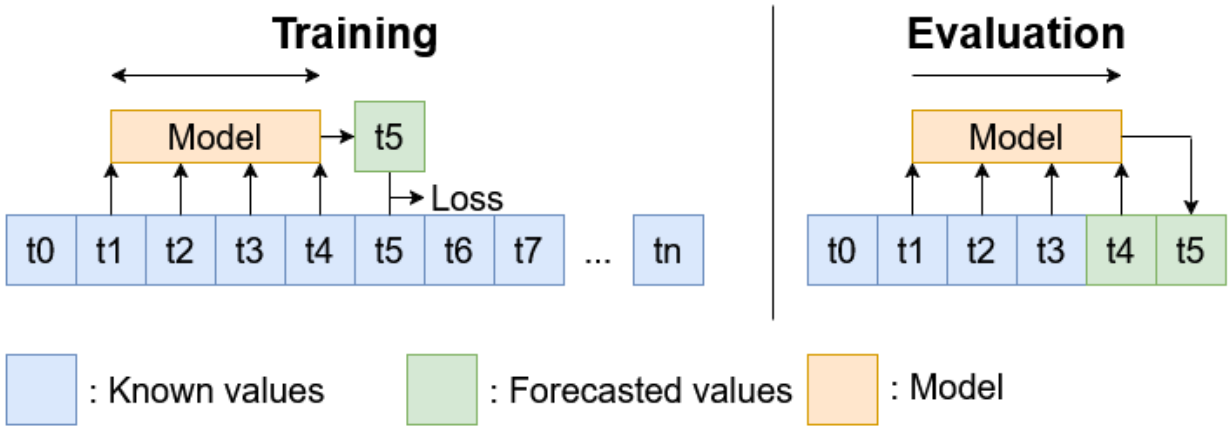


Figure A.6: Single-timestep forecasting models training and evaluation process

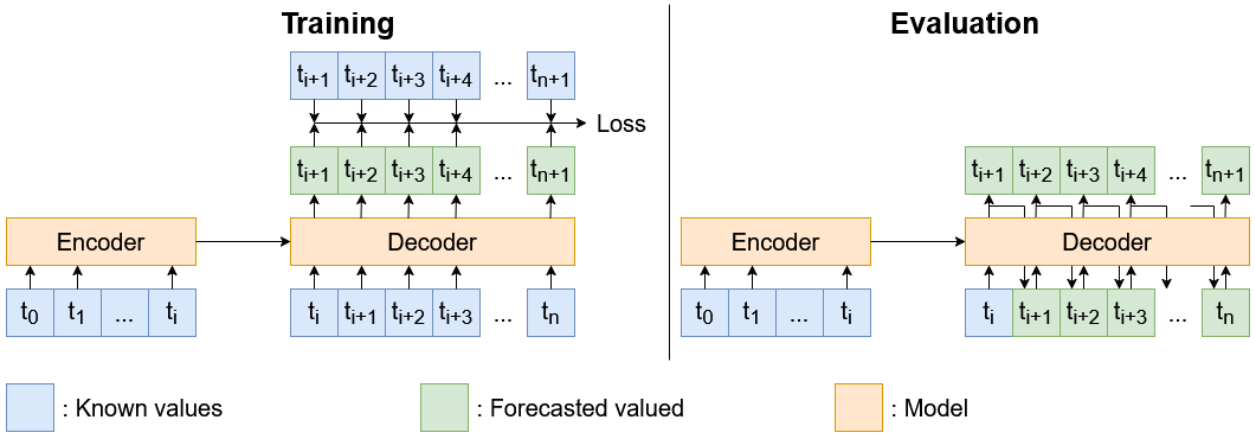


Figure A.7: Seq2seq forecasting models training and evaluation process.

Both figures illustrate the key difference between the single time-step and sequence-to-sequence approaches: in the single time-step setup, the input is updated for each individual prediction, whereas in the sequence-to-sequence model, predictions are generated recursively from an initial set of inputs.

A.8 From Regression to Classification: Label Discretization

An important aspect of classification training is maintaining a relatively balanced distribution of classes. This section outlines the methodologies we employed to discretize pollution emission values, thereby transforming regression tasks into classification tasks while preserving class balance to the best of our ability.

A.8.1 EEPS and OPS readings

EEPS and OPS readings are defined over the interval $[0, \infty)$ but are heavily dominated by zero values, which correspond to the absence of particle emissions.

Due to this strong imbalance, traditional quantile-based binning is not suitable for defining emission classes across the entire distribution. To address this, we treat exact zeros as a distinct class, and then use quantiles computed solely from the non-zero values to define the remaining classes. The results are shown in Figure A.8 and A.9 for 8 classes (which is the number used in our work) :

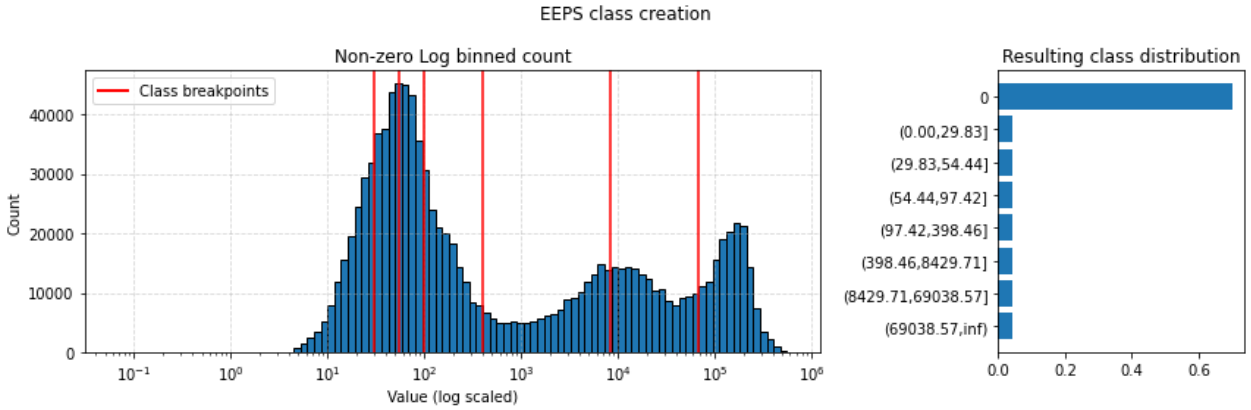


Figure A.8: Visualisation of EEPS emission class creation

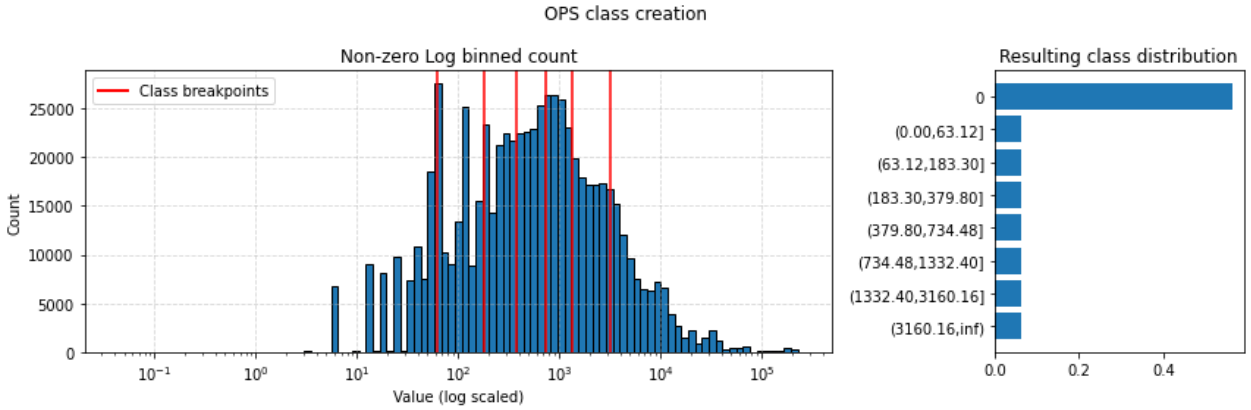


Figure A.9: Visualisation of OPS emission class creation

A key strength of this method is that it is distribution-agnostic and easy to set up. We experimented with distribution-based class construction—such as applying a Gaussian Mixture Model in log space for the EEPS emissions—but these approaches introduced numerous hyperparameter choices and frequently resulted in instability. Consequently, we chose to retain our original method, even though the resulting breakpoints may be less interpretable than those derived from well-fitted distributions.

A.8.2 Sound Spectrograms

Sound spectrograms are also defined over the interval $[0, \infty)$. However, due to inherent noise in the recording environment and the superposition of multiple sound sources, spectrograms almost never contain exact zeros; instead, their values tend to remain strictly positive. As a result, we do not encounter the same issues as with the EEPS and OPS readings.

A different challenge arises from the high spectral resolution we have chosen: high-intensity behavior typically occurs at only a few specific frequencies, while the vast majority of frequencies assume average (yet non-zero) values. This leads to a distribution with a sharp peak, where most quantiles cluster closely around the peak. Consequently, standard quantile-based binning becomes ineffective for creating meaningful and well-separated classes.

To address this, we opted to define the classes using a brute-force search to find the optimal exponent n that yields the most balanced distribution across m classes. These classes are defined as follows:

$$
\begin{aligned}
C_1 &= \left[0, \left(\frac{1}{m}\right)^n\right) \\
C_2 &= \left[\left(\frac{1}{m}\right)^n, \left(\frac{2}{m}\right)^n\right) \\
C_3 &= \left[\left(\frac{2}{m}\right)^n, \left(\frac{3}{m}\right)^n\right) \\
&\vdots \\
C_m &= \left[\left(\frac{m-1}{m}\right)^n, \infty\right)
\end{aligned}
$$

The results for m = 8 are shown in Figure A.10:

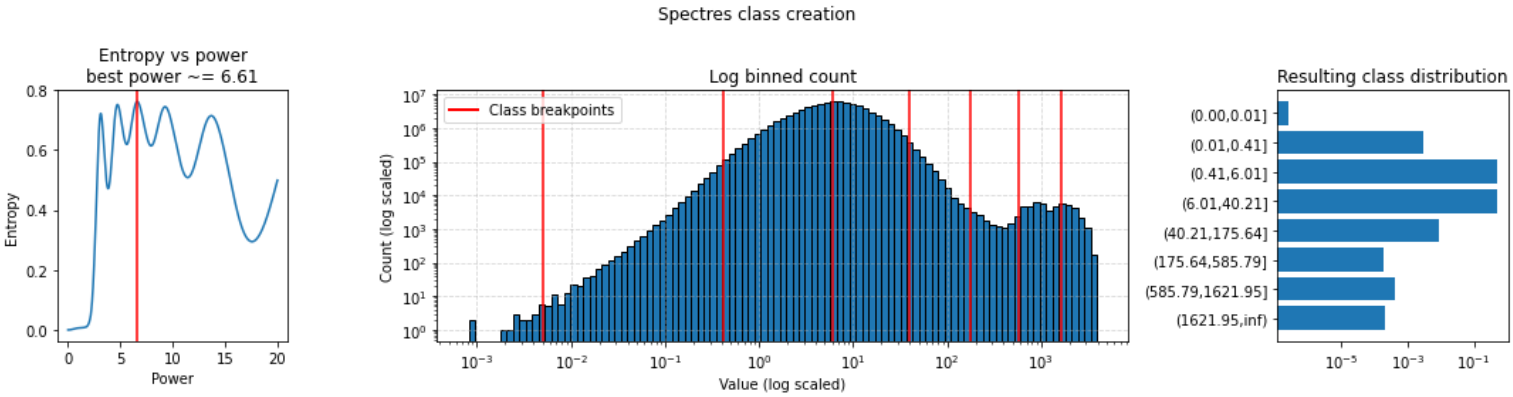


Figure A.10: Visualization of Sound Spectrogram emission class creation.

While classes 3 and 4 still account for approximately 97% of the samples, the remaining "out-of-peak" classes are relatively well balanced—except for the first class, which contains only 20 samples. Merging this class with the second not only improves the overall class balance but also yields a set of intervals that align more naturally with perceptually meaningful divisions on the decibel scale as shown in Figure A.11:

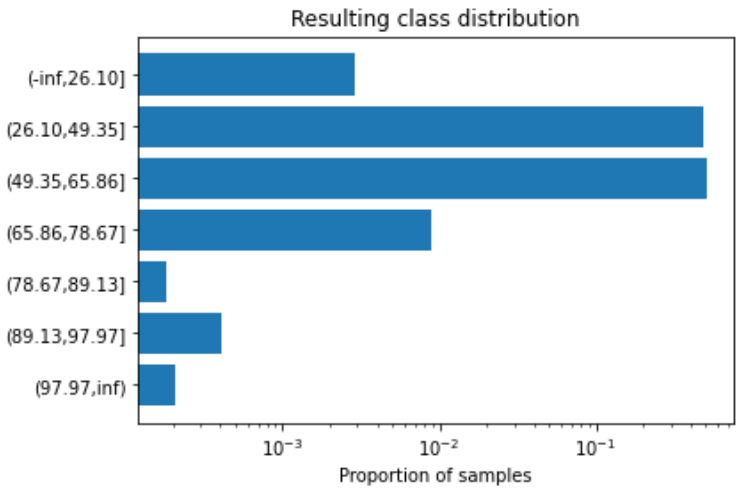


Figure A.11: Final Sound mission levels distribution in decibels

Although this distribution falls short of our goal of achieving a balanced dataset for classification, it still ensures that each class contains at least 10,000 sample points. Combined with the use of appropriate class weights during training, this may be sufficient to mitigate the effects of the highly skewed distribution of sound emissions. Unlike the previous method, this approach is far from optimal, and alternative interval definitions could certainly be considered. The main advantage, however, lies in the fact that the resulting intervals remain easily interpretable on the decibel scale.

A.9 GB2 Distribution: Definition and Optimization

Distributional Regression refers to the task of predicting the parameters of a probability distribution to best fit observed data. Ideally, this is done with prior knowledge about the data's true distribution. However, in the absence of such knowledge, one may attempt to fit multiple candidate distributions and select the best one—an approach that is both time-consuming and without guarantees. An alternative, which we adopt here, is to choose a highly general distribution that can flexibly model a wide range of behaviors.

A.9.1 BG Distribution Family and the GB2 Definition

In our case, the three variables under study—Sound Spectrogram, EEPS, and OPS readings—are all non-negative and of support $[0, \infty)$. Given this, we focus on the Generalised Beta (GB) distribution, which is known for its flexibility and generality in modeling positive-valued data.
A generalised beta random variable Y is defined by the following probability density function (PDF):

$$\mathrm{GB}(y; a, b, c, p, q) = \frac{|a|\, y^{ap-1} \left(1 - (1-c)\left(\frac{y}{b}\right)^a\right)^{q-1}}{b^{ap}\, B(p,q) \left(1 + c\left(\frac{y}{b}\right)^a\right)^{p+q}}, \quad \text{for } 0 < y^a < \frac{b^a}{1-c}$$

where $a \neq 0$, $c \in [0, 1]$, and $b, p, q > 0$.

The five parameters of the generalised beta distribution make it an extremely flexible family for modeling positive data. Figure A.12 from [41] illustrates the wide range of distributions encompassed within this family:

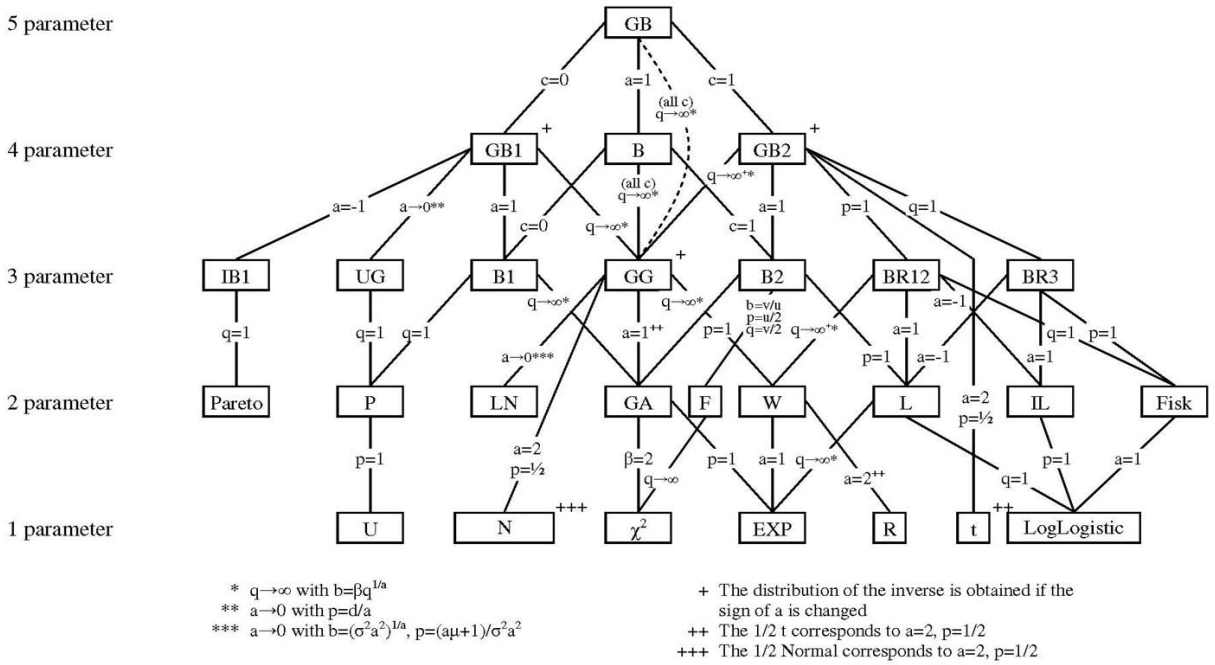


Figure A.12: Generalised beta distribution and its sub-distributions

As shown, many common distributions—including the exponential, log-normal, gamma, and chi-squared—are special cases of the generalised beta family.
However, a challenge arises when training neural networks using gradient descent on the negative log-likelihood (NLL) of this distribution. Specifically, since the density can be zero for certain values $x > 0$, the NLL can become infinite, preventing training.
To circumvent this, we set $c = 0$, which eliminates the zero-density region and yields the GB2 distribution, a subfamily of the GB distribution. This variant remains highly expressive while ensuring the PDF is strictly positive for all $x > 0$.
The GB2 distribution, when reformulated, is also known as the *generalised beta prime distribution*, a well-studied model. Its probability density function is defined as:

$$f(x;\alpha,\beta,p,q) = \frac{p\left(\frac{x}{q}\right)^{\alpha p-1}\left(1+\left(\frac{x}{q}\right)^{p}\right)^{-\alpha-\beta}}{q\,\mathrm{B}(\alpha,\beta)}, \forall x > 0$$

where $\alpha, \beta, p, q > 0$.

A.9.2 GB2 Optimisation and prediction

Starting from the previous definition, we can compute the negative log-likelihood (NLL) of the GB2 distribution (parameterized as the generalized beta prime) as:

$$\text{NLL}(x;\alpha,\beta,p,q) = -\log p - (\alpha p - 1)\log\left(\frac{x}{q}\right) + (\alpha+\beta)\log\left(1+\left(\frac{x}{q}\right)^{p}\right) + \log q + \log \text{B}(\alpha,\beta)$$

This expression is well-defined as long as $\alpha, \beta, p, q > 0$, for all $x > 0$. For $x = 0$, the probability density function (PDF) is not defined, and thus the NLL is also undefined. A solution, while not mathematically perfect, is to add a small offset to the zero observations in the term $(\alpha p - 1)\log\left(\frac{x}{q}\right)$.

With this adjustment, the expression becomes well-defined again. While this introduces a bias towards avoiding predictions of zero, using a sufficiently small offset has been shown to yield good results.

Finally, for prediction purposes (to compare to single regression), we summarize the distribution by its mode. This is more suitable than mean as the GB2 distribution can be highly skewed. It is known that the mode is defined as:

$$q\left(\frac{\alpha p - 1}{\beta p + 1}\right)^{\frac{1}{p}} \quad \text{if} \quad \alpha p \geq 1$$

When $\alpha p < 1$, the mode is undefined; however, the corresponding distributions diverge asymptotically as $x \to 0$, with the density tending to infinity. Consequently, we define the predicted value as zero in these cases.

In summary, the GB2 distribution offers a flexible and expressive framework for modeling strictly positive, skewed data. By carefully handling edge cases and leveraging its mode for prediction, we obtain a principled and practical alternative to conventional regression.

A.10 Permutation Feature Importance

Permutation Feature Importance (PFI) is a model-agnostic technique used to assess the contribution of input variables to a model's decision-making process. This method is especially valuable when working with non-linear models that do not inherently provide feature importance metrics. PFI operates by randomly permuting the values of a single feature (or a group of features) and measuring the resulting decrease in the model's performance score [9].

A.10.1 Definition

Permutation Feature Importance is most commonly defined in the context of tabular data. The importance assigned to a given variable is quantified as the decrease in a model performance metric—such as accuracy or precision—when the values in the corresponding column are randomly shuffled. This disruption breaks the relationship between the feature and the target, revealing the extent to which the model depends on that feature for its predictions. The basic principle behind this method is illustrated in Figure A.13.

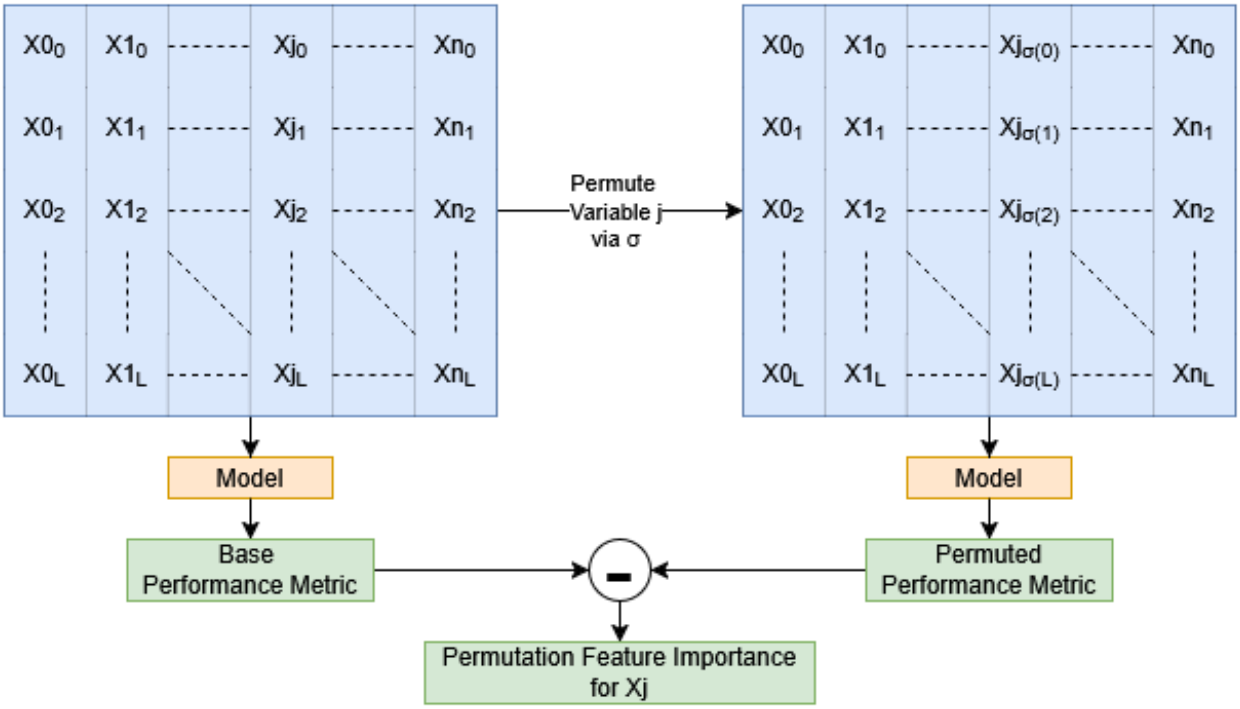


Figure A.13: Schematic illustration of Permutation Feature Importance (PFI) for tabular data.

PFI is often preferred over techniques such as feature occlusion, as it preserves the statistical distribution of the input features. This results in more reliable estimates of feature relevance, particularly for datasets where it is difficult to define an appropriate "uninformative" replacement value.

One limitation of PFI arises when features are correlated. In such cases, shuffling a single feature may not significantly degrade model performance, as the model can still access the same information through correlated variables. To address this issue, sets of highly correlated features are permuted together, allowing the method to more accurately capture the collective importance of interdependent variables. It is important to note that this fix depends heavily on the choice of correlation measure and is often heuristic in nature.

A.10.2 Extension to Time-Series Data

While PFI is traditionally defined for tabular data, it can be naturally extended to time series. Instead of permuting individual entries in a column, we permute time series corresponding to a given variable across different samples.

A challenge arises when time series have unequal lengths. A practical solution is the following procedure: let T_i be a time series sample receiving the permuted variable, and let $T_{\sigma(i)}$ be the source sample providing the replacement for variable j. We first truncate the entire T_i (i.e., all variables of sample i) to a length $L = \min(\text{len}(T_i), \text{len}(T_{\sigma(i)}))$. Then, we replace $T_{i,j}$ with the first L values of $T_{\sigma(i),j}$. This ensures that the replacement is dimensionally consistent while preserving the structure of the source time series. See Figure A.14 for an illustration:

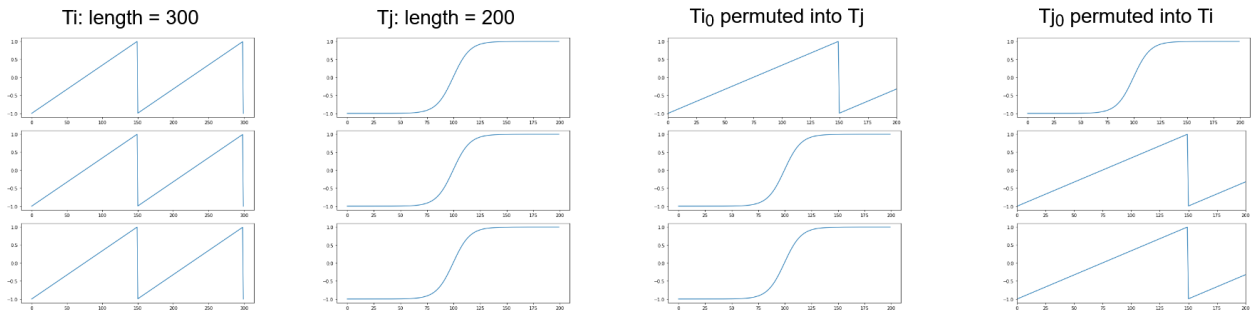


Figure A.14: Schematic illustration of truncation process for unevenly sized time-series PFI computation.

This procedure is not applicable to all tasks—for instance, classification of entire time series may be adversely affected by truncation—but it is suitable for time series regression tasks such as the one presented in Chapter 1.

Finally, when dealing with highly variable time series lengths, this method may under-represent the longest sequences: for a timestep to be included, its corresponding time series must be paired with another series of equal or greater length. Although a weighting scheme could be devised to mitigate this bias, this remains an open area for future improvement.

A.11 Integrated Gradients

Integrated Gradients (IG) [62] is a gradient-based attribution method designed to explain the predictions of deep neural networks. It improves upon simple gradient saliency maps, which directly use the gradient of the output with respect to the input features as a measure of importance. While intuitive, simple gradients often fail in practice due to saturation: once the model output plateaus with respect to an input, the gradient may vanish even though the feature remains influential. This makes raw gradients unreliable indicators of feature contribution [62].

A.11.1 Definition

Let $F : \mathbb{R}^n \to \mathbb{R}$ denote a differentiable model output function (e.g., the probability assigned to a class). For an input $x \in \mathbb{R}^n$ and a chosen baseline $x' \in \mathbb{R}^n$ representing the "absence" of features, the attribution for feature i is defined as:

$$\mathrm{IG}_i(x) = (x_i - x_i') \cdot \int_0^1 \frac{\partial F\big(x' + \alpha(x - x')\big)}{\partial x_i}\, d\alpha.$$

Instead of looking only at the gradient at x, IG integrates gradients along a straight-line path from the baseline input x' to the actual input x. This captures how the model output changes as the feature moves from "absent" to "present." Multiplying by $(x_i - x_i')$ scales the attribution to reflect the magnitude of that change.

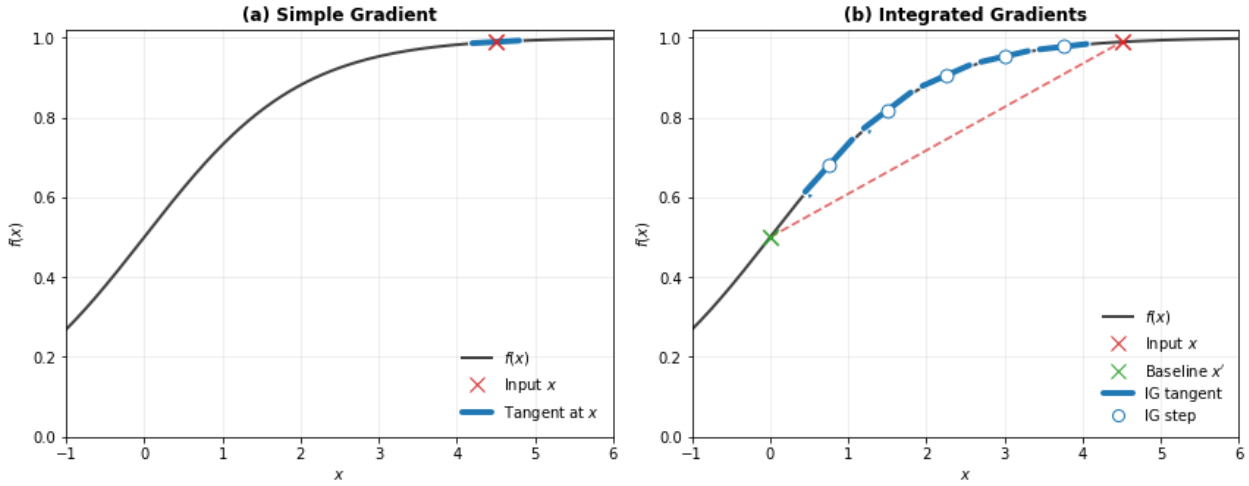


Figure A.15: Schematic illustration of Integrated Gradients: accumulating gradients along the path from a baseline input to the actual input. The example shows how, when the input lies in a saturated region of the model response, simple gradients yield near-zero attribution, whereas IG can recover the true contribution by integrating along the path—provided that an appropriate baseline is chosen.

The baseline x' is a crucial design choice. For images, a black image is often used; for tabular data, zero or mean values are common. However, when no natural baseline exists, this choice introduces variability into the results.

A.11.2 Properties and Limitations

IG satisfies desirable axioms:

- **Sensitivity**: if changing a feature from baseline to input changes the output, the feature receives a non-zero attribution.
- **Implementation invariance**: two functionally equivalent models yield identical attributions.

Despite these advantages, IG has limitations. Approximating the path integral requires multiple gradient evaluations (often 20–300 steps per input), making the method not computationally trivial. Moreover, results can vary depending on the choice of baseline, which may not always have a clear interpretation. Finally, as a gradient-based method, IG may still be sensitive to local irregularities in highly non-linear models.

Current research explores alternative baselines, non-linear interpolation paths, and variance-reduction strategies to improve both the efficiency and robustness of Integrated Gradients explanations. Compared to other attribution methods such as Layer-wise Relevance Propagation (LRP), DeepLIFT, or SHAP, IG is often preferred in practice due to its conceptual simplicity, modest implementation effort, and the fact that it requires access only to model gradients—without additional retraining or specialized backpropagation rules. While SHAP provides theoretically grounded attributions based on cooperative game theory, it is typically more computationally demanding. LRP and DeepLIFT, on the other hand, can be more sensitive to model architecture and hyperparameters. In this landscape, IG offers a pragmatic balance: it is simple, broadly applicable across neural architectures, and grounded

in axiomatic principles, which explains its widespread adoption as a default gradient-based attribution method.

A.12 Soft Actor–Critic: technical details

Soft Actor–Critic (SAC) is an off–policy reinforcement learning algorithm designed for continuous control. It achieves high sample efficiency by decoupling *data collection* from *policy optimization*: experience is stored in a replay buffer and reused many times for training.

The method relies on two main components:

- the **Critic**, a pair of action–value functions $Q_{\theta_1}, Q_{\theta_2}$ that estimate the long–term return of a state–action pair,
- the **Actor**, a stochastic policy $\pi_\phi(a|s)$ that outputs a distribution over actions.

The critics provide learning signals to improve the actor, while the actor's stochasticity (regularized by entropy) encourages efficient exploration. We now detail the optimization process.

A.12.1 Optimization

A.12.1.1 Critic update

The following figure illustrates the critic update.

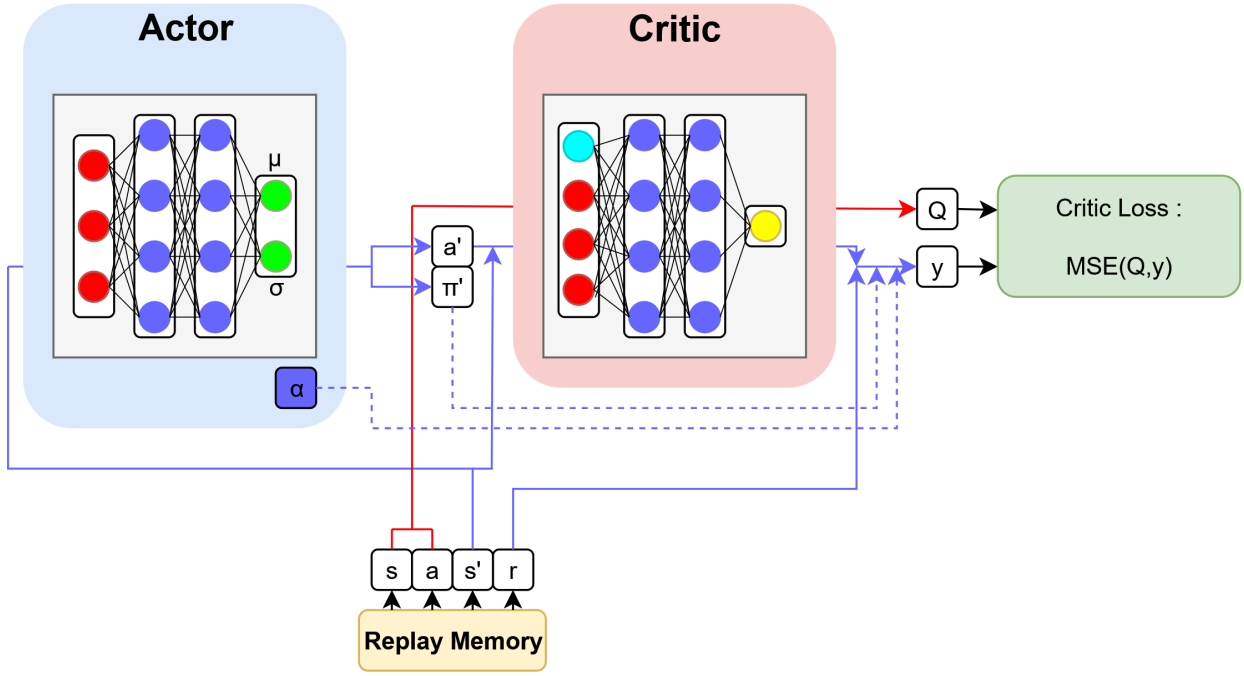


Figure A.16: SAC Critic update process. For clarity, our schematic shows only a single critic, but in practice SAC uses two critics (to reduce overestimation bias) and slowly updated target networks Q' for stability.

A minibatch (s, a, r, s') is sampled from the replay buffer, and the critic is trained to minimize the mean–squared Bellman error between its estimate $Q_\theta(s, a)$ and a bootstrapped target y:

$$y = r + \gamma \Big[\min_{i=1,2} Q'_{\bar{\theta}_i}(s', a') - \alpha \log \pi_\phi(a'|s') \Big], \quad a' \sim \pi_\phi(\cdot|s').$$

The target involves a one–step lookahead, but because the target critics Q' are themselves trained with the same equation, recursive substitution yields the *soft Bellman equation*. Consequently,the critic learns to approximate the full expected return

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r_t - \alpha \log \pi(a_t|s_t) \big) \,\Big|\, s_0 = s, a_0 = a \right].$$

A.12.1.2 Policy update

The following figure illustrates the actor and entropy updates.

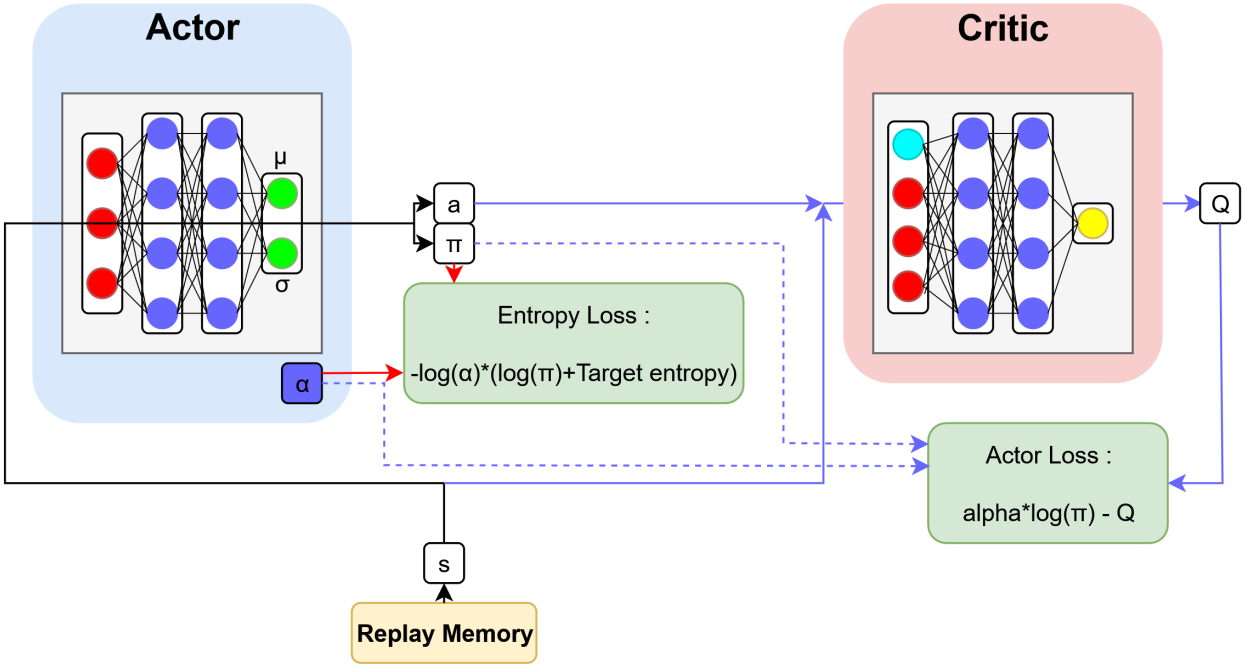


Figure A.17: SAC Policy update process. For clarity, our schematic shows only a single critic, but in practice SAC uses two critics (to reduce overestimation bias) and slowly updated target networks Q' for stability.

The actor is optimized to maximize the critics' values while preserving exploration through entropy:

$$J_\pi(\phi) = \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi_\phi}\Big[\alpha \log \pi_\phi(a|s) - \min(Q_{\theta_1}, Q_{\theta_2})(s, a) \Big].$$

Entropy is controlled by the temperature α, which is automatically tuned by minimizing

$$J(\alpha) = \mathbb{E}_{a\sim\pi_\phi}\Big[-\alpha\big(\log \pi_\phi(a|s) + \mathcal{H}_{\text{target}} \big) \Big].$$

This prevents premature convergence to deterministic policies.

A.12.1.3 Update order

Each gradient step proceeds in the following order:

1. critic update
2. actor update
3. α update
4. soft update of target critics

This ordering is natural: the actor update depends on the critics, so the critics must be refreshed first. The entropy coefficient α is updated only after observing the most recent policy behavior, hence it comes after the actor step. Finally, the target critics are updated last, ensuring that the bootstrapped targets used in the next iteration remain stable during the current updates.

A.12.2 Data collection and evaluation

The following figure illustrates the mechanisms used for *data collection* during training and for *evaluation* once a policy is learned:

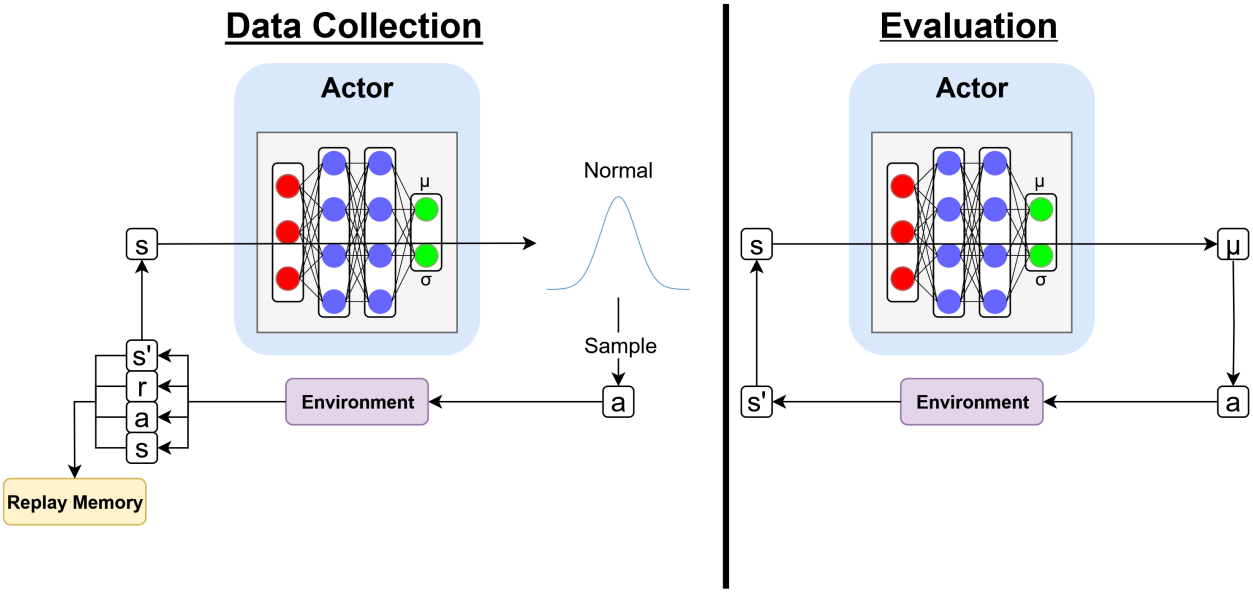


Figure A.18: SAC interaction mechanisms. On the left is the training setup, used to fill the replay buffer, and on the right is the evaluation setup, used to measure performance after training.

During **Data collection** (left), the actor samples $a \sim \pi_\phi(\cdot|s)$, the environment returns (s', r), and the tuple (s, a, r, s') is stored in the replay buffer. Stochastic sampling from the policy ensures sufficient exploration.
During **Evaluation** (right), the actor behaves deterministically, choosing $a = \mu_\phi(s)$ without sampling or entropy regularization. This provides a consistent measure of policy performance.

A.12.3 Conclusion

In summary, SAC combines an off–policy replay buffer, double critics, and entropy–regularized actor updates. This design yields stable learning, high sample efficiency, and robust performance across continuous control tasks.

Machine Learning in Brake Tribology: Pollution Prediction, Mechanism Analysis, and Control Strategies

---

Résumé

Les systèmes de freins à disque illustrent la complexité des contacts tribologiques, où des phénomènes mécaniques, thermiques, acoustiques et chimiques interagissent de manière fortement couplée et évolutive. Ces interactions soulèvent des enjeux pratiques, notamment le bruit généré et les émissions particulaires, qui sont difficiles à appréhender uniquement par des approches physiques classiques. Cette thèse explore comment le Machine Learning peut enrichir l'analyse et l'optimisation de la tribologie des freins. À partir d'un jeu de données multimodal combinant des mesures acoustiques, thermiques, mécaniques et d'émissions de particules, nous examinons d'abord des modèles prédictifs pour le bruit émis et les émissions particulaires. Nous allons ensuite au-delà de la prédiction en appliquant des méthodes de clustering et de classification à des spectrogrammes acoustiques. Associées à des techniques d'interprétabilité, elles révèlent des états de signal récurrents ainsi que les variables qui les influencent. Enfin, nous couplons un simulateur par éléments finis d'un système de frein avec du reinforcement learning dans une étude de faisabilité, montrant que cette approche peut découvrir des politiques de contrôle adaptatives améliorant la maîtrise et la stabilité des freins à disque. Bien que ces travaux demeurent exploratoires et encore éloignés des standards industriels — qu'il s'agisse de la prédiction en temps réel, de la compréhension des mécanismes ou du contrôle — ils ouvrent des perspectives prometteuses et laissent entrevoir la possibilité de généraliser de tels modèles à l'échelle du véhicule réel. Dans leur ensemble, ces contributions abordent des challenges liés à la granularité, à l'interprétabilité et à la diversité méthodologique dans le domaine du Machine Learning appliqué à la tribologie du freinage, illustrant comment le Machine Learning peut compléter les outils expérimentaux et numériques pour améliorer la conception et le fonctionnement des systèmes de freinage.

Mots-clés : Machine learning, Tribologie du freinage, Modélisation interface de contact, Problèmes multi-échelle et multi-physique

Abstract

Disk brake systems exemplify the complexity of tribological contacts, where mechanical, thermal, acoustic, and chemical phenomena interact in strongly coupled and evolving ways. These interactions give rise to practical concerns such as noise and particulate emissions, which remain difficult to capture with physics-based approaches alone. This thesis investigates how machine learning can extend the analysis and optimization of brake tribology. Using a multimodal experimental dataset that records acoustic, thermal, mechanical, and particle-emission signals, we first examine predictive models for braking noise and pollution. We then move beyond prediction by applying clustering and classification methods to acoustic spectrograms, complemented by interpretability techniques, which reveal recurring signal states and the variables that influence them. Finally, we couple a finite-element brake-like simulator with reinforcement learning in a proof-of-concept study, demonstrating that reinforcement learning can discover adaptive control policies that enhance the control and stability of disk brakes. While these contributions remain exploratory and not yet at the level of industrial standards—whether for real-time prediction, mechanism understanding, or control—they highlight promising directions and suggest that data-driven approaches could ultimately be generalized to vehicle-scale applications. Taken together, the work addresses open challenges of granularity, interpretability, and methodological breadth in machine learning applied to brake tribology research, illustrating how machine learning can complement experimental and numerical tools in advancing the design and operation of braking systems.

Keywords: Machine learning, Brake Tribology, Contact interface modeling, Multi-scale and multi-physics