Dottorato di Ricerca in / PhD program Computer Science **Ciclo / Cycle** XXXVIII

Curriculum in Artificial Intelligence and Machine Learning

TITOLO TESI / THESIS TITLE

In-Processing Neuro-Fuzzy Approaches for Fairness-Aware and Explainable Machine Learning

Nome / Name SAHAR **Cognome / Surname** SHAH

Matricola / Registration number 899134

Supervisor: Prof. Italo Francesco Zoppis

Co-Supervisors: Prof. Davide Elio Ciucci, Prof. Sara Lucia Manzoni

Company Supervisor: Dr. Francesco Epifania

Coordinator: Prof. Leonardo Mariani

ANNO ACCADEMICO / ACADEMIC YEAR

2022/25

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani

Abstract

Machine learning (ML) now plays a direct role in decisions that shape people's access to healthcare, credit, employment, and justice. When these systems learn from data that carries social bias, they can unintentionally reinforce inequality rather than help to reduce it. This thesis focuses on improving fairness through in-processing strategies while keeping decision processes understandable to those affected by them. The central idea is to blend the transparency of fuzzy rules with the predictive strengths of neural networks (NNs). In doing so, the work explores how sensitive information influences model behaviour and proposes mechanisms to keep that influence under control.

The thesis unfolds in three main stages. The first contribution introduces an in-processing learning strategy that uses attribution signals to identify when sensitive features dominate predictions and suppresses such behaviour during training. The second contribution updates the Adaptive Neuro-Fuzzy Inference System (ANFIS) by integrating kernel functions to increase flexibility and by injecting attribution feedback to improve stability and rule usefulness. The third contribution builds on this idea by allowing both the rule base and membership functions to evolve while the network trains, making the system more responsive to how bias appears across different data contexts.

These developments are later unified into a single framework that works with commonly used attribution techniques, such as Integrated Gradients (IGs) and SmoothGrad. The framework reports fairness indicators alongside classification performance and offers interpretable, rule-level insights into where harmful bias may arise.

The methods are evaluated on simulated datasets and public benchmarks, including Adult, Correctional Offender Management Profiling for alternative sanctions (COMPAS), and German Credit. Across all cases, the proposed models reduce group-level disparity measures such as statistical parity difference and disparate impact, while maintaining accuracy comparable to standard baselines. The results confirm that it is possible to improve both interpretability and fairness at the same time when

neuro-fuzzy reasoning is embedded into the core of the learning process.

Dedication

Love and support have been my greatest strength. I dedicate this to my wife, Mahnoor Khan, my supportive brother, Mr. Tahir Shah, my adorable little boy, Mr. Raheel Shah, and my parents, Zulfat Shah and Saira Bibi.

Acknowledgements

I want to thank my perfect supervisor, **Prof. Italo Francesco Zoppis**, from the core of my heart, for his unwavering support, encouragement, and direction during my PhD adventure. His wise counsel and helpful criticism have greatly influenced my academic growth and research. My work was considerably enhanced by the insightful recommendations, patience, and unwavering drive of my co-supervisors, **Prof. Sara Lucia Manzoni** and **Prof. Davide Elio Ciucci**, for which I am equally grateful. I also want to thank **Dr. Francesco Epifania**, who has greatly improved the technical and practical aspects of this project. He is the CEO of Social Thingum Srl and the head of MudiLab, where I work in an exciting research environment. I am really appreciative of the PNRR (Piano Nazionale di Ripresa e Resilienza) funding support, which enabled this research.

On a personal note, I owe everything to my family. To my father, **Mr. Zulfat Shah**, for being my constant source of strength and encouragement; to my mother, **Mrs. Saira Bibi**, who has been my everything and the foundation of my life, and to my elder brother, **Mr. Tahir Shah**, whose unwavering belief in me and continuous support have shaped what I am today. My deepest love and gratitude go to my wife, **Mahnoor Khan**, and my cute son, **Raheel Shah**, whose smile and presence reminded me of the bigger picture and gave me the strength to move forward.

List of Publications

[1] S. Shah, D. E. Ciucci, S. L. Manzoni, and I. F. Zoppis, *Neural Networks Bias Mitigation Through Fuzzy Logic and Saliency Maps*, Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025), Porto, Portugal, ISBN 978-989-758-737-5.

[2] I. F. Zoppis, S. Shah, S. L. Manzoni, and D. E. Ciucci, *Kernelizing Adaptive Neuro-Fuzzy Inference for Bias Mitigation*, In 2025 IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1–6. IEEE, 2025.

[3] S. Shah, F. Zaman, F. Es Sabery, F. Epifania, and I. F. Zoppis, *Fine-Tuning of Distil-BERT for Continual Learning in Text Classification: An Experimental Analysis*, IEEE Access 12 (2024): 104964–104982.

[4] S. Shah, D. E. Ciucci, S. L. Manzoni, and I. F. Zoppis, *Dynamic Neuro-Fuzzy Regularization for Fairness-Aware Neural Networks Using Saliency Attribution*, AIMS 2025 (co-located with FLLM 2025), Vienna, Austria, 25–28 Nov 2025.

Submitted / Under Review Papers

[1] I. F. Zoppis, S. Shar, S. L. Manzoni, G. Lazzarinietti, D. Malchiodi, and D. E. Ciucci, *In-processing Neuro-Fuzzy Approaches for Bias Mitigation*, submitted to *Journal of Artificial Intelligence Research (JAIR)*, ACM, October 2025. *(Submitted).*

List of Acronyms

Acronym	Meaning
AI	Artificial Intelligence
ML	Machine Learning
XAI	Explainable Artificial Intelligence
NN(s)	Neural Network(s)
ANFIS	Adaptive Neuro-Fuzzy Inference System
LIME	Local Interpretable Model-agnostic Explanations
SHAP	SHapley Additive exPlanations
IG(s)	Integrated Gradient(s)
Grad-CAM	Gradient-weighted Class Activation Mapping
TCAV	Testing with Concept Activation Vectors
DP	Demographic Parity
EO	Equalized Odds
TP	True Positive
NLP	Natural Language Processing
DL	Deep Learning
SMOTE	Synthetic Minority Over-sampling Technique
ANOVA	Analysis of Variance
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
EU AI Act	European Union Artificial Intelligence Act
ACC	Accuracy
DI	Disparate Impact
SPD	Statistical Parity Difference
EOD	Equal Opportunity Difference
AOD	Average Odds Difference
CR(s)	Classification Rule(s)

Acronym	**Meaning**
BR(s)	Bias Rule(s)
γ^{last}	Final normalized bias control parameter
ΔF	Change in F-measure relative to baseline (M0)
F	F-measure (harmonic mean of precision and recall)
$F_\beta(\beta = 1)$	F1-score at ($\beta = 1$)
Bias	Learned bias regularization value
Non-Reg	Non-Regularized model
Saliency-Reg	Saliency-Regularized model
Fuzzy-Reg	Fuzzy-Regularized model
FL	Fuzzy Logic
PR	Post-hoc Reweighting (model variant)
ν^*	Normalized bias index
β	Beta parameter controlling recall–precision trade-off
M0	Baseline model (no fairness intervention)
M1–M10	Model variants or ablations used in comparisons
M_FL_DP	Fuzzy-Logic debiasing tuned for Demographic Parity
M_FL_EO	Fuzzy-Logic debiasing tuned for Equalized Odds
M_PR	Post-hoc Reweighting model variant

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Trend

Over the past twenty years, ML has moved from research labs into everyday applications. It now supports decision-making in areas such as finance, health care, education, employment, and criminal justice [1]. This rapid spread has been made possible by the increased computing power, the rise of large public datasets, and the development of new neural architectures. As these systems became more capable, they also began to influence people's lives in direct ways—deciding who gets a loan, which candidate is shortlisted for a job, or how a medical condition is classified. Along with these benefits came growing concern about how fair and transparent these automated processes really are [2].

Research over the past decade has shown that ML models trained on historical data often reproduce the same social biases present in those data [3]. Well-known examples include the COMPAS tool for recidivism prediction, which produced unequal error rates across racial groups, and hiring systems that downgraded resumes from female applicants because of past gender imbalances. These examples demonstrate that ML models are socio-technical systems rather than impartial instruments, and their results are influenced by the data, objectives, and limitations that inform them. As a result, the area now addresses fairness, interpretability, and responsibility rather than just accuracy. This relationship is summed up in Figure 1.1: the degree to which fairness, interpretability, and accountability overlap and support one another determines how credible Artificial Intelligence (AI) is.

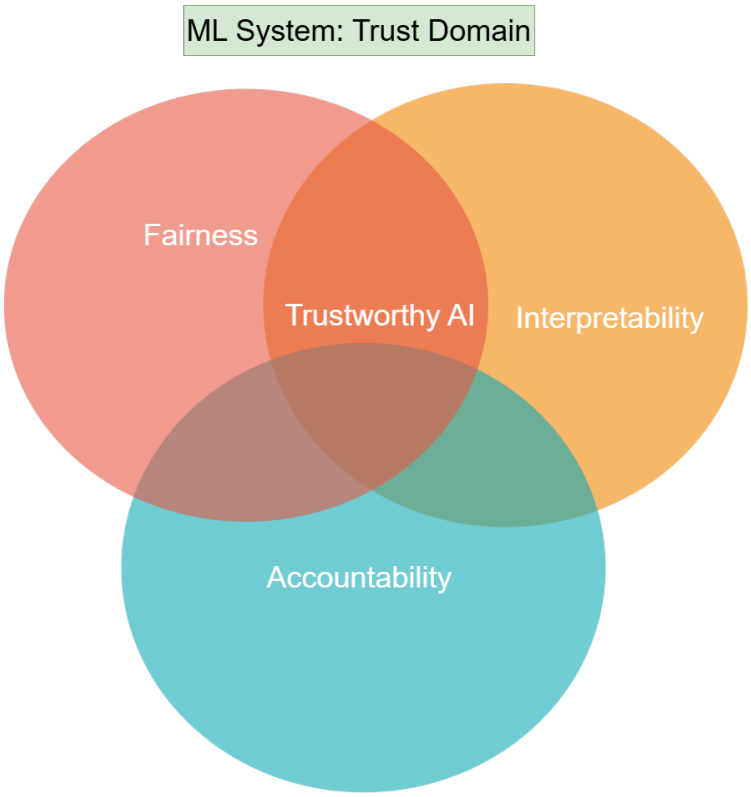

Figure 1.1: Trust domains in ML: fairness, interpretability, and accountability.

Institutions and governments have started to respond to these problems. For example, high-risk AI applications must be subject to independent evaluations for fairness and transparency under the European Union Artificial Intelligence Act (EU AI Act) [4]. Ethical principles that demand human oversight, traceability, and non-discrimination in automated systems have also been released by professional associations and policy groups [5]. These steps point in a clear direction: bias prevention must result in models that end users can understand, audit, and trust, in addition to lowering inequalities. Figure 1.2 demonstrates how Explainable AI (XAI) techniques can enhance automated results by connecting black-box predictions with human comprehension.

Technically speaking, the machine learning community has put out a number of strategies to lessen bias. These fall into three categories: in-processing techniques that include fairness in the learning process directly, post-processing techniques that affect model outputs, and pre-processing techniques that alter training data [6]. Although they are frequently simpler to implement, pre- and post-processing obscure the basic

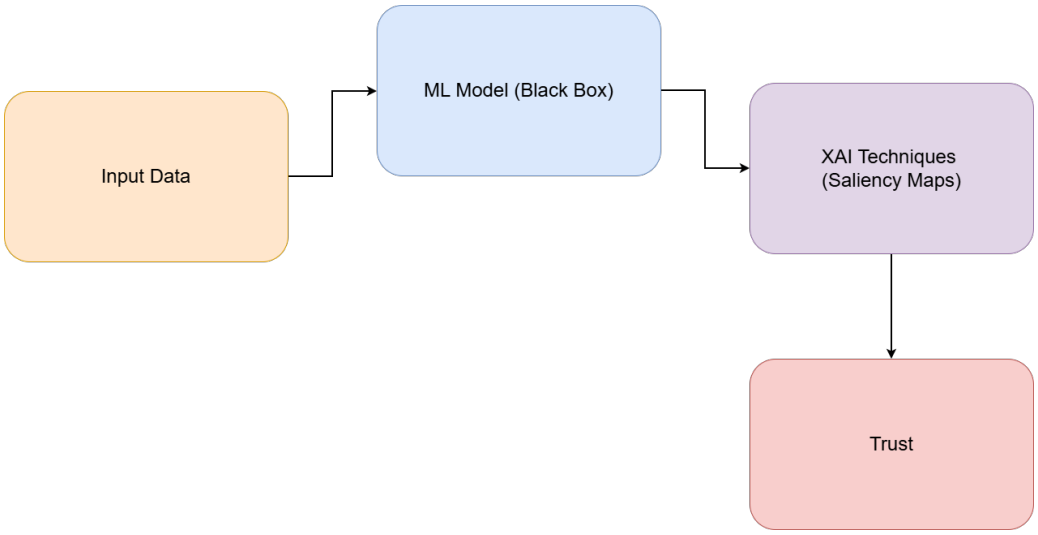

Figure 1.2: Explainable AI (XAI) pipeline.

idea. On the other hand, in-processing techniques alter the learning goal itself, directing the model to adhere to fairness restrictions. They are therefore appropriate for situations requiring both performance and transparency. However, the majority of in-processing techniques used today, such as fairness-constrained optimization and adversarial debiasing, continue to function as black boxes [7]. Their implementation in high-stakes choices is limited because, while they may enhance statistical metrics of fairness, they provide little insight into how those gains are made.

As a result, recent research tends to integrate interpretability and fairness. While interpretability included within the model itself explains how adjustments take place, explainability aids in determining whether a model depends too much on sensitive features. Using rule-based models, causal structures, or fuzzy logic, hybrid solutions have the potential to develop ML systems that are socially responsible [8]. Because it can convey rules that are comprehensible by humans while yet fitting into gradient-based training, fuzzy logic is particularly helpful [9]. This characteristic makes it perfect for integrating model transparency with fairness methods. The Neuro-Fuzzy XAI architecture, as illustrated in Figure 1.3, combines NN feature extraction with fuzzy inference and explainability layers to provide decisions that are transparent and human-understandable.

All things considered, the present ML trend is a result of a confluence of policy pressure, public concern, and technological advancement. In addition to accuracy, models are being evaluated on their capacity for fair behavior, justification, and

accountability. New algorithms and frameworks that make those methods transparent and verifiable are both necessary to address bias in neural systems.

By creating in-processing neuro-fuzzy approaches that incorporate interpretability into fairness mitigation, the work presented in this thesis moves in this direction and advances the larger objective of reliable and socially conscious ML.

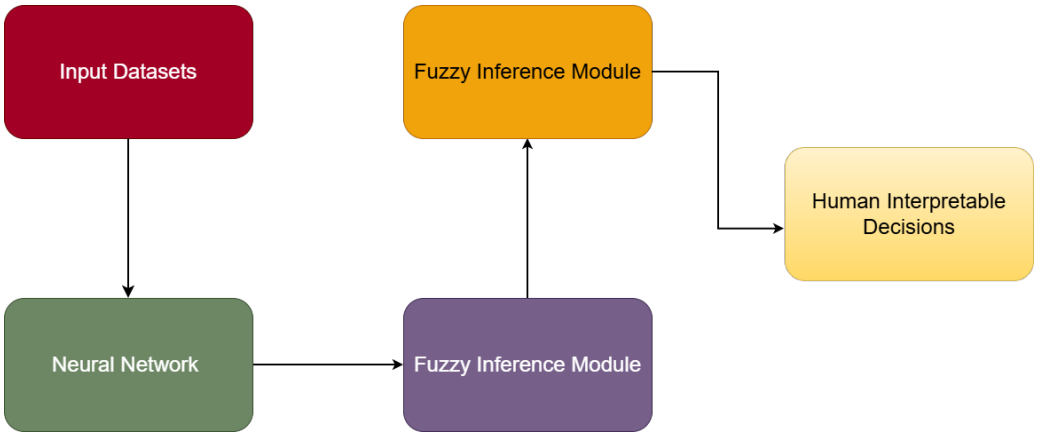

Figure 1.3: Conceptual Neuro-Fuzzy XAI framework.

1.2 Thesis Objectives and Motivations

The primary objective of this research is to create ML models that, without compromising accuracy, produce predictions that are transparent and equitable. Because ML now informs real decisions that affect people, models must be both reliable and interpretable. While conventional NNs are effective at learning intricate patterns, they function as black boxes, providing little information about how they arrive at their findings. It is challenging to determine whether a model is introducing bias by using sensitive data, like gender, ethnicity, or age, because of this lack of interpretability.

To address the challenges mentioned above, this thesis explores a set of neuro-fuzzy learning strategies that bring the interpretability of fuzzy rules into NN training. The idea is to ensure that fairness is not introduced later as a correction but becomes a natural feature of the model from the very beginning. By guiding the model during learning instead of after deployment, the system can avoid basing decisions too heavily on sensitive attributes. The fuzzy rule structure is well-suited for this because it interacts smoothly with gradient-based optimization while still expressing decision logic in a form that humans can inspect and understand.

The overall aim is to integrate the goals of fairness and interpretability into one process instead of treating them in isolation. The research investigates how NNs can be steered away from relying on unfair correlations in the data while they learn. Attribution tools, such as saliency analysis and IGs, are used to detect when sensitive information has too much influence on predictions. When that occurs, the proposed mechanisms adjust the learning behavior to keep those effects under control while preserving predictive performance. To achieve this, attribution methods such as saliency maps and IGs [10] are used to monitor which features influence the network's predictions. If these analyses reveal that a sensitive feature has gained too much importance, fuzzy rules are applied to moderate its effect. Presenting this control through explicit rules keeps the intervention understandable and allows the learning algorithm to remain transparent in how it reduces bias.

The first objective of this research is to develop a regularization strategy that uses saliency information to limit how much a model relies on sensitive attributes. Feature-attribution methods, such as saliency maps, highlight cases where these attributes influence predictions more than they should. In such situations, the regularization term increases the training penalty to reduce this effect. The aim is to improve fairness without causing a noticeable drop in predictive accuracy [11].

The second part of the work extends ANFIS by incorporating kernel functions and attribution-guided fuzzy rules. These additions help the system handle complex data more reliably and maintain clear rule behavior in high-dimensional settings [12].

The third objective introduces a training strategy in which the membership functions and rules are not fixed. Instead, they update throughout learning, so the model can better react when patterns of bias shift across the dataset [13].

All three design choices are then unified into a single framework. The full system allows rule activations and fairness indicators to be monitored during training, and it remains compatible with attribution methods such as IGs and SmoothGrad [11][14].

Although fairness-constrained learning and adversarial debiasing have produced strong empirical results, the decision logic they induce is often difficult to inspect or communicate [15]. In many settings, practitioners rely on post-hoc explainability to justify model behaviour, yet attribution-based explanations can vary across runs and may not reliably indicate the true drivers of a prediction [16]. In addition, a large portion of the literature applies fairness interventions after a model has already learned its internal representations, which limits how directly bias is controlled during training [17]. These limitations motivate a design in which fairness objectives and interpretability are built into the learning process itself, using neuro-fuzzy components that express decisions through structured rules and membership functions [18].

The objectives of this thesis are as follows:

- Develop in-processing learning strategies that use feature-attribution information to monitor how sensitive attributes influence NN predictions and reduce this influence when necessary.
- Build upon ANFIS by introducing kernel functions and attribution-guided fuzzy rules, improving its ability to model complex data while preserving rule interpretability.
- Design a training scheme in which fuzzy rules and membership functions update over time, allowing the system to adapt when bias patterns change within the data.
- Integrate these components into a single, differentiable neuro-fuzzy framework that remains compatible with standard optimization algorithms and is portable across diverse datasets.
- Assess the proposed models on simulated data and public benchmarks (Adult, COMPAS, German Credit), comparing their fairness indicators and predictive behaviour against well-established baselines.

These aims respond to well-known shortcomings in existing fairness-oriented learning techniques and match the increasing expectation for transparency when automated predictions are used in practice. Current policies, including the EU AI Act, emphasize that decisions affecting individuals must be traceable and justified. Within this thesis, fairness and interpretability are not treated as add-ons after the model is trained but as core properties built into the design. The framework relies on fuzzy rules to make the decision logic accessible throughout both training and evaluation. More broadly, the objective is to move away from systems judged only by accuracy and toward models whose outcomes can be understood, questioned, and trusted in real-world settings.

1.2.1 Research Questions

This thesis is guided by the following research questions:

RQ1: How can attribution signals be used during training to detect and suppress reliance on sensitive features?

RQ2: Can the inclusion of a neuro-fuzzy rule layer provide intrinsic interpretability while improving group fairness metrics without incurring a substantial loss in predictive accuracy?

RQ3: How does the kernelization of Adaptive Neuro-Fuzzy Inference Systems (ANFIS) influence representational capacity and training stability under fairness constraints?

RQ4: When bias patterns vary across data contexts or distributions, can dynamic updates of rules and membership functions preserve fairness improvements over time?

RQ5: What trade-offs arise between fairness performance, interpretability, and computational cost when compared with standard fairness-aware learning baselines?

These considerations motivate the choice of a neuro-fuzzy modelling framework in this thesis. Fuzzy systems provide rule-based representations that are inherently interpretable, allowing decision logic to be expressed in a form that can be inspected and constrained. At the same time, their integration with neural architectures enables end-to-end differentiable training using standard optimization techniques. This combination offers a natural mechanism for embedding fairness-related constraints directly into the learning process while preserving both predictive performance and transparency. For these reasons, neuro-fuzzy models are particularly well-suited to addressing the interpretability and adaptability challenges identified above.

1.3 Contributions

This thesis contributes to fairness-aware machine learning by designing neuro-fuzzy models that leverage attribution signals while preserving the clarity of rule-based reasoning.

- A learning mechanism that uses attribution cues to prevent NNs from relying heavily on sensitive data during training.
- Enhancements to ANFIS through kernel functions and attribution-guided rule construction, improving stability and the capability to capture complex relations.
- A dynamic update process in which both the rule base and membership functions evolve during training, enabling the model to adapt when bias patterns shift.
- A unified neuro-fuzzy design that remains compatible with standard gradient-driven optimization and supports different attribution techniques.

- Experimental analysis on benchmark datasets, including Adult, COMPAS, and German Credit, comparing predictive behaviour and fairness outcomes with widely used baselines.

Taken together, these developments demonstrate that fairness and interpretability can be achieved within one learning pipeline rather than through later adjustments. The work progresses from a simple regularization idea to a complete framework that can be trained, interpreted, and evaluated through rule inspection. The proposed models allow the role of sensitive features to be examined while the network learns, showing that fairness improvements can be obtained without sacrificing reliability. The thesis, therefore, established a practical approach for building models whose behaviour can be monitored and guided throughout training.

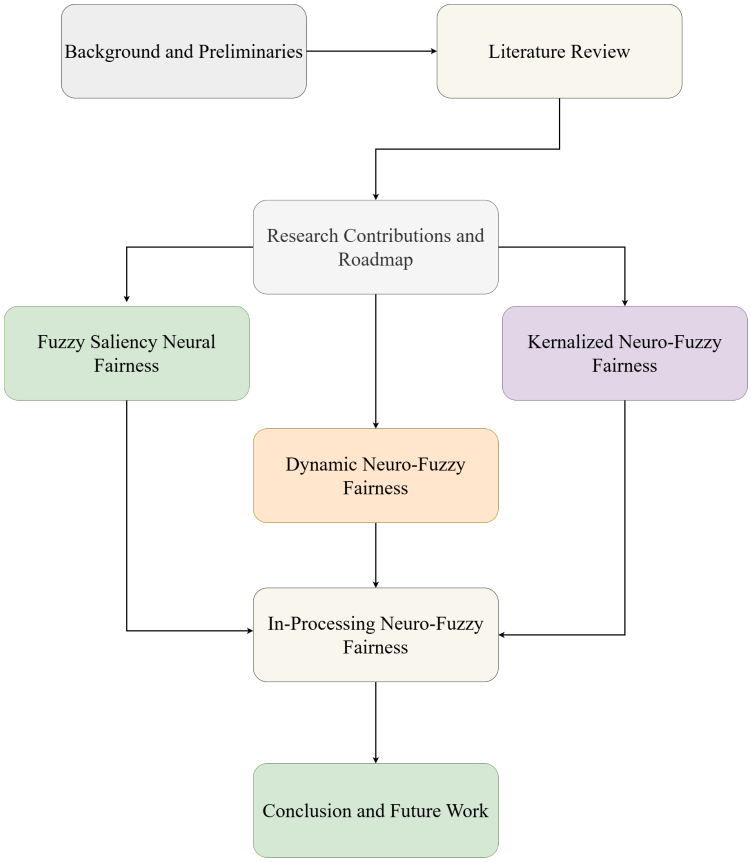

Figure 1.4: Organization of the thesis and flow of contributions.

1.4 Thesis Outline

This thesis is structured into nine chapters, progressing from motivation and theoretical background to model development, evaluation, and final reflections. The material is presented as a continuous narrative rather than as separate papers, so each chapter builds on the previous one. Figure 1.4 illustrates the overall organization of the thesis and highlights how the individual methodological contributions are developed and integrated.

Chapter 1 – Introduction describes the motivation for the research and sets out the main goals. It highlights why fairness and interpretability are necessary qualities in current ML systems and introduces the neuro-fuzzy ideas developed in the thesis.

Chapter 2 – Background and Preliminaries outlines the core concepts needed for the remainder of the work, including fairness definitions, attribution techniques, and neuro-fuzzy modelling. The problem setting and notation used throughout the thesis are introduced here.

Chapter 3 – Literature Review surveys existing research on bias mitigation and interpretability in ML. It discusses the current limitations that motivate the need for new approaches.

Chapter 4 – Research Contributions and Roadmap connects the individual methods developed in the thesis and describes how the work evolves from the initial idea to the full framework.

Chapter 5 – Saliency-Guided Fuzzy Regularization introduces the first model, which employs attribution signals during learning to reduce reliance on sensitive information.

Chapter 6 – Kernelized and Attribution-Driven ANFIS presents the second contribution, enhancing ANFIS with kernel functions and attribution-based rule design to improve flexibility and stability.

Chapter 7 – Dynamic Fuzzy Rules for Adaptive Mitigation develops a training scheme in which both the fuzzy rules and their membership functions are updated as learning progresses, enabling adaptation to changes in bias patterns.

Chapter 8 – Unified Neuro-Fuzzy Framework with Benchmark Validation integrates all previous elements into a complete system and evaluates its predictive performance and fairness using standard datasets.

Chapter 9 – Synthesis and Conclusions summarizes the key outcomes, discusses limitations, and suggests directions for future research.

The first three chapters provide context and theoretical grounding, while the remaining chapters develop, test, and refine the proposed models. The next chapter

begins with the background material needed to understand the methods introduced later in the thesis.

Chapter 2

Background and Preliminaries

2.1 Algorithmic Fairness Foundations

The wider use of ML in credit, courts, healthcare, and recruitment has raised fresh worries about biased outcomes [19]. Bias can appear at different stages of the ML pipeline—from data collection to model deployment—and, if left uncorrected, may reproduce or even worsen existing social inequalities [20]. This section reviews the main types of bias found in ML, the formal definitions of fairness most often discussed in the literature [21], and the common trade-offs between fairness and predictive performance [22].

2.1.1 Types of Bias in ML

Bias in ML can take several forms rather than a single, uniform pattern. Some of the most widely studied types include:

- **Dataset bias**: appears when the training data do not adequately represent the target population. For instance, if a healthcare dataset underrepresents certain demographic groups, a model trained on it may perform poorly or unfairly on those populations [23].
- **Representation bias** appears when certain sensitive groups, for example, by gender or ethnicity, are over- or under-represented in the data. This can hide unintended links between those groups and the model's predictions.
- **Measurement bias** can arise when the process of gathering features or labels alters the values being recorded. For instance, crime statistics may reflect the

frequency of policing in certain neighborhoods rather than the true occurrence of illegal activities.

- **Algorithmic bias** can originate from the model's learning behavior, especially when optimization is driven mainly by accuracy. In such cases, the model may reinforce patterns in the data that already disadvantage certain demographic groups [24].

Different types of bias can appear at various points in the workflow, and they often interact, which makes it hard to isolate a single source. Mitigating their impact requires a broad view of the learning pipeline, covering data collection, model design, and evaluation practices.

2.1.2 Fairness Notions

Researchers have suggested different mathematical approaches for reducing bias in ML systems. Since no single definition is suitable for all applications, two main perspectives are commonly considered:

- **Group fairness** evaluates how outcomes differ between groups defined by sensitive attributes. Several measures fall into this category:

- *Demographic Parity (DP):* a model should give positive predictions with similar frequency across groups, regardless of the sensitive attribute.

- *Equalized Odds (EO):* rates of false positives and false negatives should be comparable for different groups [25].

- *Predictive Parity:* a given prediction at the output of a model, its correctness should be consistent across different groups.

- **Individual fairness** requires that people with similar relevant characteristics receive similar predictions. This approach evaluates fairness at the level of each example rather than only through group statistics.

These fairness notions do not always align, so a model that satisfies one may fail to meet another. For this reason, the fairness objective must be chosen based on the needs and context of a specific application.

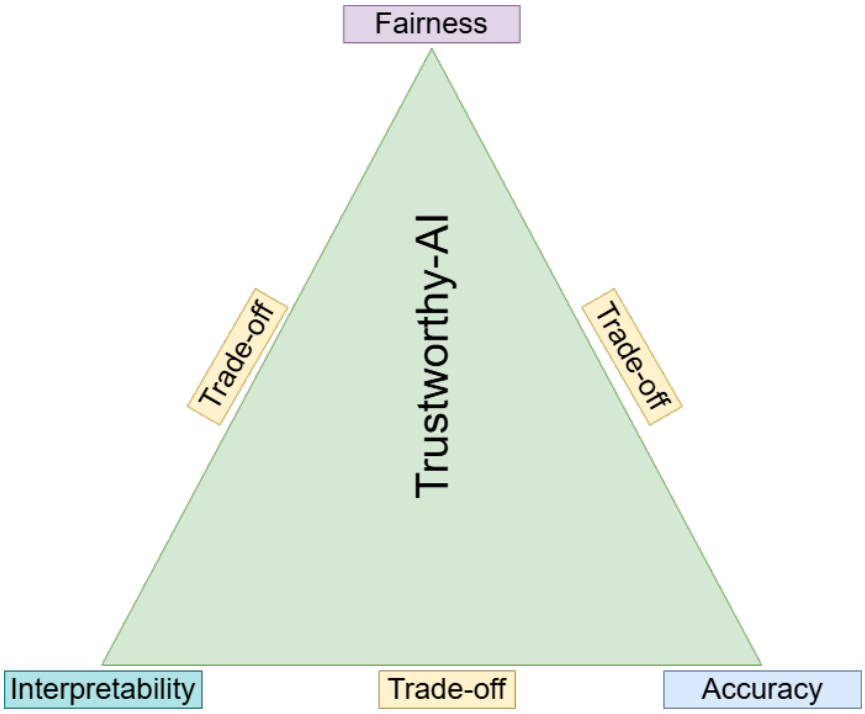

Figure 2.1: Fairness–Accuracy–Interpretability trade-off triangle with *Trustworthy AI* at the center.

2.1.3 Fairness–Accuracy Trade-off

A well-known challenge in fairness-aware learning is the effect that fairness constraints can have on accuracy. When a model is discouraged from relying heavily on sensitive attributes, its predictive performance may drop slightly, especially if those attributes correlate with the label [26]. The objective is to limit this impact while avoiding outcomes that treat certain groups unfairly [27].

In this work, fairness is incorporated into the learning procedure itself rather than as a correction applied afterward. In practice, accuracy and fairness need to be balanced and monitored together. The proposed methods combine fairness objectives with interpretable rule-based reasoning to support prediction quality and allow inspection of how decisions are formed [28]. Figure 2.1 illustrates the relationship between fairness, interpretability, and accuracy.

2.2 Explainable Artificial Intelligence (XAI)

As ML models become increasingly complex, it is often difficult to understand why they produce certain predictions. Techniques such as deep NN, ensembles, and other non-linear approaches can reach high accuracy, yet the reasoning behind their outputs is not always accessible to end users [29]. This lack of transparency can be especially concerning in critical applications, including finance, healthcare, and criminal justice, where decisions require justification and accountability.

For these reasons, XAI has emerged as a research area focused on providing insight into model behavior and making predictions easier to interpret [30].

2.2.1 Post-hoc and Intrinsic Interpretability

Interpretability approaches are commonly divided into two broad categories:

- **Intrinsic interpretability** describes models that are designed to be understandable without additional tools. Examples include decision trees, linear models, and rule-based systems. Their decision logic is directly visible, although these models may face difficulties with very large or complex feature spaces [31].
- **Post-hoc interpretability** refers to techniques that explain the behaviour of models whose internal reasoning is not easily accessible. These methods aim to provide insight into decisions after the model has been trained. Common approaches include visualizations, feature-importance measures, and local surrogate models [32]. These techniques are widely used in deep learning (DL), where the model structure generally does not allow direct interpretation.

2.2.2 Attribution Methods

Attribution methods represent a key class of post-hoc approaches. They assign a contribution score to each input feature based on its effect on a particular model output. Among the most commonly used are

- **Saliency Maps**: compute gradients of the output with respect to inputs, highlighting which features have the greatest impact on the decision [11].
- **IGs**: average gradients along a path from a baseline input to the actual input, producing more stable and theoretically grounded importance scores [33].

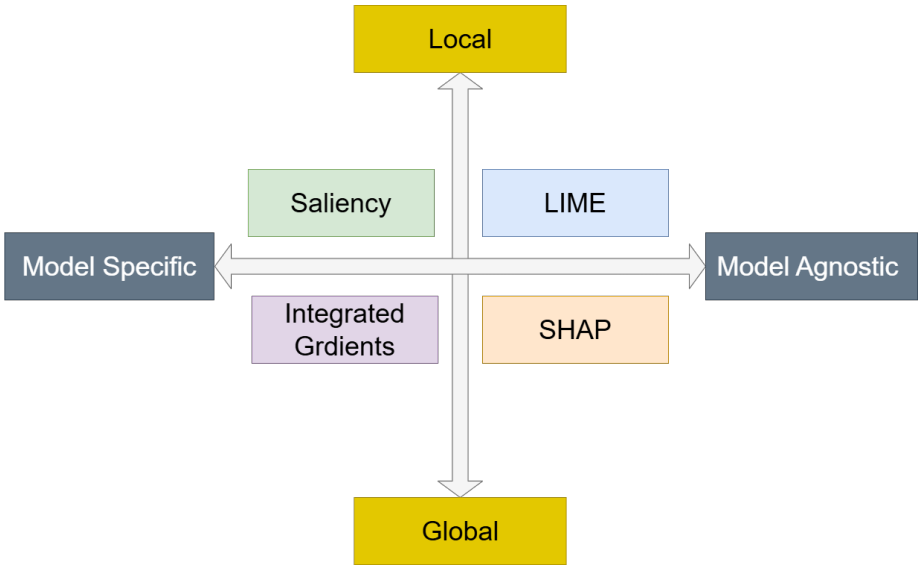

Figure 2.2: Typical attribution methods in XAI categorized by their scope (local vs global) and model dependency (specific vs agnostic).

- **SmoothGrad**: reduces noise in saliency maps by averaging attributions over multiple perturbed versions of the input [34].
- **SHAP (SHapley Additive exPlanations)**: based on cooperative game theory, attributes contributions to features by computing Shapley values, ensuring fairness and consistency in feature importance [35].
- **LIME (Local Interpretable Model-agnostic Explanations)**: builds local surrogate models, typically linear regressors, to approximate the behavior of the complex model in the neighborhood of an instance [36].

These techniques have been widely applied to image classification, natural language processing (NLP), and tabular decision-making tasks. However, they primarily serve as diagnostic tools and do not, by themselves, enforce fairness. Figure 2.2 illustrates the main attribution approaches by distinguishing their locality and their dependence on the underlying model

Among the available attribution techniques, IGs and their smoothed variants are well-suited for use during training. These methods provide consistent explanations that do not depend on how a model is implemented, and they respond appropriately when input features meaningfully affect the output. Compared to simple gradient-based approaches, they are less sensitive to noise and small input changes. This

makes them reliable for repeated use during learning, where attribution scores are used to guide regularization and fairness constraints.

2.2.3 XAI and Fairness

While XAI has traditionally focused on transparency, it increasingly overlaps with the study of algorithmic fairness. Attribution methods can be repurposed to reveal whether sensitive attributes (e.g., gender, race, age) exert disproportionate influence on predictions [37]. For example, if a credit approval model heavily relies on gender-related features, this can be identified through saliency or SHAP values. Such insights make it possible to design fairness-aware models that either regularize or constrain feature reliance during training. In this sense, XAI provides the tools not only to audit bias but also to mitigate it [38].

2.2.4 Limitations

In this thesis, attribution information plays an active role during model training. Instead of being computed after the model has learned, attribution scores are incorporated directly into the learning objective. While XAI approaches have progressed rapidly, they still come with practical constraints. Some attribution methods can be sensitive to minor input changes, leading to different results for nearly identical samples. Many techniques focus on explaining a single prediction at a time, which makes it harder to understand the model's behaviour as a whole [39]. In addition, certain explanations may not be intuitive for users without technical expertise. These limitations indicate that post-hoc explanations alone cannot fully address the need for transparency [40]. .

2.3 Fuzzy Logic and Neuro-Fuzzy Systems

Fuzzy logic, first proposed by Zadeh in the 1960s, provides a mathematical framework for representing uncertainty using degrees of truth rather than strict binary values [41]. Instead of assigning outcomes as entirely true or false, fuzzy logic allows values in between, which better matches how people interpret many real situations [42]. This ability to represent gradual transitions makes it useful in fields such as control systems, decision support, and pattern recognition, where sharp boundaries between classes are uncommon.

2.3.1 Fuzzy Sets and Rules

Fuzzy logic is built on the concept of fuzzy sets, which generalize classical set theory by assigning each element a level of membership rather than a strict yes or no value [43]. A fuzzy set A on a domain X is described by a membership function $\mu_A : X \to [0, 1]$, indicating the degree to which each element belongs to A. For instance, instead of labeling income as simply "high" or "not high," fuzzy sets allow the membership degree to increase gradually with income, avoiding an abrupt cutoff [44].

Decision rules in fuzzy systems follow simple linguistic structures, such as:

IF income is high **AND** age is young, **THEN** credit risk is low.

These rules resemble everyday reasoning and make the model's behavior easier to interpret. Typical inference involves four steps: fuzzifying the inputs, evaluating the rules, combining their outputs, and converting the final fuzzy result back into a numerical value.

2.3.2 Adaptive Neuro-Fuzzy Inference Systems (ANFIS)

Fuzzy systems are easy to interpret, but on their own, they do not always adapt well to complex or high-dimensional data. Neuro-fuzzy systems address this issue by combining the rule-based structure of fuzzy logic with the learning ability of NNs. One well-known model in this category is ANFIS [45].

ANFIS uses a layered structure in which each layer corresponds to a step of the fuzzy inference process: computing membership values, activating rules, normalizing the activations, and producing the final output. The parameters of the membership functions and rules are updated during training, often using gradient-based or hybrid optimization, so that the rule base improves as data are seen.

This approach has been applied in many different areas, including control, healthcare, financial prediction, and NLP [42]. Its ability to learn from data while keeping decisions interpretable makes it suitable for applications where both accuracy and transparency are required.

2.3.3 Neuro-Fuzzy Systems for Fairness

In fairness-aware learning, fuzzy logic can be used to express constraints in a form that is easy to interpret [46]. For instance, a rule might limit how much a sensitive attribute such as gender or race can influence the final prediction. Incorporating fuzzy rules into a neuro-fuzzy model enables the system to encourage fairer decisions while

still keeping track of how predictions are formed. This approach becomes even more useful when paired with attribution information from XAI methods [39]. Attribution techniques quantify the contribution of each feature to the model's output, and fuzzy rules offer a structured way to adjust that influence. Combining these elements supports fairness improvements during learning and keeps the underlying reasoning accessible for human interpretation.

2.3.4 Limitations

While fuzzy and neuro-fuzzy systems offer interpretability benefits, they also introduce some practical difficulties. As the number of features or rules grows, fuzzy models may encounter scalability issues, sometimes described as the curse of dimensionality [47]. Neuro-fuzzy systems can be computationally demanding, and if too many rules are created automatically, their clarity may decrease [48]. To manage these challenges, researchers typically rely on structured rule design, regularization methods, and hybrid learning strategies to preserve both interpretability and performance.

In this thesis, fuzzy logic and ANFIS form the basis of the proposed fairness-aware learning methods. Attribution signals are included directly within the neuro-fuzzy structure so that reliance on sensitive features can be monitored and controlled during training. Section 2.4 describes how these components fit together in the overall problem formulation.

2.4 Problem Formulation and Notation

This section introduces the problem formulation and the notation followed in the remainder of the thesis. A precise mathematical description helps explain how fairness objectives, attribution information, and neuro-fuzzy rules are combined in the proposed learning framework.

2.4.1 Supervised Learning Setup

We consider a supervised learning setting with a dataset

$$\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N},$$

where $x_i \in \mathbb{R}^d$ is the feature vector, $y_i \in \mathcal{Y}$ is the target label, and $s_i \in \mathcal{S}$ is a sensitive attribute (e.g., gender, age, or ethnicity). Thus, the dataset includes standard predictive features together with variables that may influence fairness considerations.

The goal is to learn a function

$$f_\theta : \mathbb{R}^d \to \mathcal{Y},$$

parameterized by θ, that achieves accurate predictions while also respecting fairness constraints.

2.4.2 Fairness Notation

We follow commonly used fairness definitions in the literature. Let $\hat{y}_i = f_\theta(x_i)$ denote the prediction for the i-th instance.

- **Group fairness** aims for similar prediction statistics across different values of a sensitive attribute. A standard example is DP:

$$P(\hat{y} = 1 \mid s = 0) \approx P(\hat{y} = 1 \mid s = 1).$$

- **Equalized Odds (EO)** requires that error rates are comparable across sensitive groups when conditioning on the true label y:

$$P(\hat{y} = 1 \mid y, s = 0) \approx P(\hat{y} = 1 \mid y, s = 1).$$

- **Individual fairness** suggests that similar individuals (in terms of non-sensitive features) should receive similar predictions.

These definitions are used throughout this thesis to evaluate fairness and to guide model design.

2.4.3 Attribution Scores

Attribution methods provide feature-level explanations of a model's output. For a given input x_i, the attribution vector is defined as

$$a_i = \text{Attr}(f_\theta, x_i) \in \mathbb{R}^d,$$

where $\text{Attr}(\cdot)$ can be any attribution technique, such as IGs, saliency maps, or SHAP.

In this work, these attribution values are incorporated into training so that the model's reliance on sensitive features can be monitored and constrained when necessary.

2.4.4 Neuro-Fuzzy Rule Layer

We include a fuzzy rule layer to influence how the model uses sensitive information. Each rule follows the generic form:

IF x_j is A **THEN** the rule output is w,

where A is a fuzzy set with a membership function $\mu_A(x_j)$ and w is a learnable rule weight. For an input x_i, the rule activations are combined to give the inference output

$$R(x_i; \phi) \in \mathbb{R},$$

where ϕ collects all membership parameters and rule weights.

This layer is fully differentiable, allowing its parameters to be updated together with the rest of the predictive model during training.

2.4.5 Optimization Objective

The training objective combines three components as follows:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_f \, \mathcal{L}_{\text{fair}} + \lambda_r \, \mathcal{L}_{\text{reg}},$$

where:

- $\mathcal{L}_{\text{pred}}$ measures the primary task performance, for example through cross-entropy or mean squared error.
- $\mathcal{L}_{\text{fair}}$ discourages the model from depending too heavily on sensitive features by using attribution information and fairness-related indicators.
- $\mathcal{L}_{\text{reg}}$ is a regularization term controlling model complexity and stabilizing fuzzy rule learning.

Hyperparameters λ_f and λ_r balance fairness and regularization against predictive accuracy.

This formulation establishes a rigorous foundation for fairness-aware neuro-fuzzy systems. The notation defined here will be used consistently throughout the subsequent chapters when presenting methodologies, experiments, and results.

This chapter has introduced the foundational concepts required to understand fairness and interpretability in MLL systems. It reviewed the main definitions of algorithmic fairness, discussed commonly used fairness metrics, and highlighted

the inherent trade-offs that arise when multiple fairness criteria are considered simultaneously. The chapter also examined interpretability from both post-hoc and intrinsic perspectives, emphasizing the limitations of explanation methods that operate independently of the learning process. In addition, key attribution techniques and their reliability concerns were outlined, providing essential context for later methodological choices. Together, these elements establish the theoretical background and terminology used throughout the remainder of the thesis. Building on this foundation, the next chapter surveys the state of the art in fairness-aware learning, critically analyzing existing mitigation strategies and identifying the gaps that motivate the contributions developed in subsequent chapters.

Chapter 3

Literature Review

3.1 Introduction to Algorithmic Fairness

ML has moved from research labs into decisions that carry real consequences: who is shortlisted for a job, who receives credit, who is flagged for extra medical screening, and who is assessed as a risk in judicial settings. In these domains, performance alone is not enough. Models must also avoid systematic disadvantages for groups defined by sensitive attributes such as gender, age, or ethnicity [2][49]. The study of algorithmic fairness addresses this need by asking two questions. First, how do biases arise in data and models? Second, how should we define and measure fair behaviour in a way that can guide model design and evaluation [50][15]?

3.1.1 Where bias enters the pipeline

Bias is not a single fault but the result of many small frictions across the pipeline [51]. *Data collection* can underrepresent certain populations or capture outcomes that reflect historical practices rather than ground truth [2][52]. *Feature engineering* can introduce proxy variables that correlate with sensitive attributes [53]. *Labels* may encode human judgments that are themselves inconsistent across groups [49]. Even with perfect data, *learning algorithms* optimize average loss, which can push errors onto minorities when they form a small share of the training set [50][15]. During *deployment*, models meet new distributions and institutional constraints, and decision thresholds selected for overall accuracy can quietly create unequal error rates [17]. Because these sources interact, effective mitigation must combine statistical reasoning with domain understanding and governance [51][54].

3.1.2 Fairness notions: group and individual perspectives

Two viewpoints dominate the literature. *Group fairness* compares statistics across groups defined by a sensitive attribute [53][55]. Typical targets include parity of positive prediction rates, parity of true positive rates, or parity of error rates more generally. These measures are attractive because they are simple to compute and align with many regulatory expectations, yet they can disagree with each other in practice [56].

Individual fairness asks for consistency at the person level: individuals who are similar with respect to task-relevant information should receive similar outcomes [50]. This view depends on a similarity notion that is meaningful for the domain. It is harder to specify and evaluate than group metrics, but it captures concerns that group averages can miss, such as within-group heterogeneity and local boundary effects [15][52].

In real-world applications, practitioners often monitor a small panel of group metrics while checking for egregious individual-level inconsistencies cite friedler2019comparative. The choice of metric should be driven by the harms the institution seeks to prevent and by the decisions that the model will support [15][52].

3.1.3 Measuring fairness in practice

For binary decisions, several summary quantities are commonly reported:

- **Statistical Parity Difference (SPD)**: difference in positive prediction rates across groups. Values near zero indicate parity [57].
- **Disparate Impact (DI)**: ratio of positive rates. Values close to one indicate parity [55].
- **Equal Opportunity Difference (EOD)**: difference in true positive (TP) rates across groups [53].
- **Average Odds Difference (AOD)**: average of differences in TP and false positive rates [15].

These metrics answer different questions. SPD and DI look only at predictions, while EOD and AOD condition on the ground truth and therefore probe how errors are distributed [52]. For multi-class settings or regression, practitioners adapt the spirit of these measures, for example, by assessing calibration and error distributions per group [58]. In addition to aggregate metrics, it is useful to inspect *reliance profiles*:

how strongly a model uses features, including proxies of sensitive attributes, across groups and subpopulations [59].

3.1.4 Trade-offs and incompatibilities

Fairness criteria are not mutually compatible in general [56]. A model calibrated within each group may fail to equalize error rates. Enforcing equalized error rates can shift thresholds and hurt calibration or overall utility[60]. Because there is no universal optimum, model selection becomes a multi-objective problem that balances utility, equity, and interpretability [61]. That balance must be explicit. Reporting only accuracy is misleading; reporting a single fairness metric can be misleading as well. A clear policy should state which constraints are prioritized and why [15].

3.1.5 Auditing workflow

A practical auditing workflow usually proceeds in four steps [54][62]. First, establish data quality: coverage by group, label consistency, and potential proxies for sensitive attributes. Second, train baseline models and compute a dashboard of metrics, including accuracy, calibration, and group fairness measures at several decision thresholds [57]. Third, inspect explanations of model behaviour to discover which features drive predictions overall and by group [63]. This stage often reveals proxies and spurious correlations. Fourth, apply mitigation strategies and re-audit the model, documenting any accuracy–fairness trade-offs. The process should be iterative and documented so that model owners can justify decisions to stakeholders [64].

3.1.6 Limits of current practice

Three limitations appear repeatedly in applied work. First, many strong debiasing techniques provide little visibility into where the mitigation acts inside the network. This complicates communication with non-technical stakeholders and weakens trust [37]. Second, post-hoc explanations alone are fragile: they may be unstable to perturbations and can be manipulated by design choices [16]. Third, static rule sets rarely keep up with shifting data distributions, highlighting the need for adaptive fairness mechanisms [64]. Last, fairness interventions therefore need to be both interpretable and adaptive, and they should surface artifacts that can be audited, such as rule activations and feature-reliance summaries [59].

This thesis is positioned within this landscape. It combines attribution signals with fuzzy and neuro-fuzzy reasoning to provide in-processing mitigation that is both

differentiable and interpretable. Table 3.1 outlines the core concepts that define algorithmic fairness, summarizing how fairness is understood, measured, and audited in practical systems. The next section reviews mitigation strategies in detail and places the proposed approach within that taxonomy.

Table 3.1: Algorithmic Fairness.

Concept	Meaning	Examples	Why it Matters	Limitations
Data bias	Bias arises from unbalanced, missing, or non-representative data.	Historical hiring data includes fewer women, so the model learns a skewed pattern.	Locates the source of unfairness before model training.	Hard to fix without improving data collection and coverage.
Group-based fairness	Ensure groups (e.g., gender, ethnicity) receive comparable outcomes.	Equal chance of loan approval for different demographic groups.	Addresses social equity at the population level.	May ignore individual differences within groups.
Individual fairness	Similar individuals should receive similar predictions.	Two applicants with the same skills get similar hiring scores.	Promotes person-level consistency.	Needs a clear notion of "similarity," which is context-dependent.
Fairness–accuracy trade-off	Improving fairness can reduce predictive accuracy, and vice versa.	A hiring model becomes slightly less accurate but much fairer.	Forces explicit design choices and transparency.	Finding the right balance is problem-specific.
Auditing fairness	Measure, monitor, and report fairness metrics on data and models.	Compare error rates and thresholds across groups during validation.	Enables accountability and continuous improvement.	Often requires access to sensitive attributes and careful handling.
Limits of current practice	No single metric or method works in all contexts.	Different metrics can conflict on the same task.	Encourages multi-metric evaluation and context-aware choices.	Adds complexity; can confuse stakeholders without clear guidance.

3.2 Bias Mitigation Strategies

Once fairness has been defined and measured, the next challenge is to reduce unwanted disparities. The literature organizes mitigation approaches into three broad families: pre-processing methods that reshape the data, in-processing methods that modify the training procedure, and post-processing methods that adjust predictions [55][61][65][66]. Each category follows its own intervention approach and offers specific strengths as well as potential limitations.

3.2.1 Pre-processing approaches

Pre-processing techniques modify the training data so that models trained afterward are less likely to reproduce unfair patterns [61][65]. Common methods include:

- **Re-weighting**: assigning different weights to examples so that groups contribute more evenly to the loss [61].
- **Re-sampling**: oversampling underrepresented groups or undersampling overrepresented ones to balance the dataset[67].
- **Data transformation**: learning a representation of the features in which sensitive attributes are less predictive, sometimes through adversarial training [65][68].
- **Synthetic data augmentation**: creating extra training examples to help balance the representation of different groups [69].

A key strength of pre-processing methods is that they do not depend on the specific model used afterward: once the data are modified, any learning algorithm can be applied [57]. On the other hand, these methods need direct access to the raw data, which is not always feasible, and they may remove features that are informative but also correlated with sensitive attributes [58].

3.2.2 In-processing approaches

In-processing methods intervene directly in the learning procedure. This group includes a range of strategies [46] [53][70]:

- **Regularization terms**: introducing penalty components in the loss function when predictions differ across groups or when attribution scores indicate overdependence on sensitive features [46][70].

- **Adversarial training**: training the predictor alongside an adversary that tries to infer the sensitive attribute from the predictor's hidden representation. Success of the adversary signals that information about the sensitive feature remains; minimizing its accuracy enforces invariance [68][71].
- **Constraint-based optimization**: incorporating fairness constraints (such as EO) directly into the training objective or as side constraints solved with Lagrangian methods [72][73].
- **Interpretable model structures**: designing architectures that are inherently more transparent, such as rule-based or neuro-fuzzy models, which allow inspection of decision logic [11][74].

These methods are powerful because they act at the core of model training, but they typically require access to model parameters and the ability to modify the optimization loop. They are therefore less suitable for black-box models but are central to research on fairness-aware NNs.

3.2.3 Post-processing approaches

Post-processing methods operate after the model has been trained. They adjust either the predictions or the decision thresholds to improve group-level metrics [53][58][75]. Typical methods include:

- **Threshold adjustment**: setting different thresholds per group to equalize error rates or positive rates [53].
- **Prediction calibration**: adjusting predicted probabilities so that they match observed outcomes more closely across groups [58].
- **Re-labelling**: modifying the predicted labels selectively to improve fairness criteria while trying to keep accuracy at a reasonable level [75][76].

A benefit of post-processing is that it can be applied without retraining the model, since it operates only on the final predictions [57]. A drawback is that identical feature profiles may lead to different outcomes for different groups after adjustment, which can reduce individual-level consistency [52].

3.2.4 Comparing the Three Families

Fairness interventions are commonly divided into three groups, which differ in how they modify the learning pipeline and in the transparency they provide [2][17].

- **Pre-processing**: prepares more representative and balanced datasets before training, often through techniques such as reweighting or resampling. This can lower disparities early, although some bias may reappear later in training [61].
- **Post-processing**: adjusts model predictions after training to correct unfair outcomes. These methods are simple to deploy but may hide the origins of bias within the model [75].
- **In-processing**: integrates fairness considerations directly into the learning objective or training rules. This offers stronger control over model behaviour, but usually requires additional design effort [46][52].

In many practical cases, a combination of these strategies is used. A system might include balanced training data, fairness-aware objectives, and post-training adjustments to meet fairness targets [17][57]. Choosing among these options depends on the application domain, performance requirements, interpretability needs, and the expectations of stakeholders [15].

This thesis adopts an in-processing perspective. Contributions include fuzzy and attribution-based penalties during training and extensions to neuro-fuzzy inference to improve fairness while maintaining interpretability. The rules evolve as training progresses, allowing the model to adapt and remain explainable rather than relying solely on post-hoc corrections.

Table 3.2 compares major bias mitigation strategies used at different stages of the ML pipeline, highlighting their typical use cases, benefits, and constraints.

Table 3.2: Bias Mitigation Strategies.

Category	Method	How it works	When to use	Strengths	Limitations
Pre-processing	Reweighting	Assign higher weights to minority or underrepresented groups.	Any model with imbalanced data.	Model-agnostic; easy; improves group parity.	May raise variance; ignores label bias.
Pre-processing	Sampling (under/over, SMOTE)	Adjust sample counts to balance groups or classes.	Small or imbalanced datasets.	Quick parity gains; simple setup.	Overfitting risk; synthetic data may distort features.
Pre-processing	Fair representations	Learn features that hide sensitive info while keeping task signal.	When adding a representation step.	Transferable; reduces direct bias.	Extra training step; accuracy may drop.
In-processing	Adversarial debiasing	Train a model with an adversary predicting the sensitive attribute.	Deep or flexible models.	Joint fairness–accuracy control; strong impact.	Needs group labels; tuning sensitive.
In-processing	Fairness constraints	Add fairness constraints or penalties to the loss (e.g., EO, DP).	When you can edit training objectives.	Directly targets fairness metrics.	Can reduce accuracy; harder optimization.

(Continued on next page)

(Continued from previous page)

Category	Method	How it works	When to use	Strengths	Limitations
In-processing	Dependence regularizers	Penalize correlation between outputs and sensitive features.	When lightweight control is preferred.	Easy to add; interpretable.	Penalty strength task-dependent.
Post-processing	Threshold adjustment	Apply group-specific thresholds to align error rates.	Black-box models already trained.	No retraining; simple to apply.	Alters outputs; needs group info at inference.
Post-processing	Score calibration	Calibrate or reshape scores per group.	When calibrated outputs exist.	Flexible; model-agnostic.	May harm consistency; adds complexity.
Post-processing	Reject-option	Shift uncertain cases toward the disadvantaged group near boundary.	High-stakes tasks with score margins.	Improves local fairness.	Policy sensitive; may seem intrusive.

3.3 Explainable AI (XAI) and its Role in Fairness

Deep NNs achieve remarkable accuracy, yet their complexity obscures how predictions are made [77]. A major concern in complex ML systems is that their inner workings are often difficult to interpret in domains where accountability is essential [63]. XAI seeks to improve the transparency of such models so that developers, auditors, and decision makers can better understand how predictions are formed [39]. In the context of fairness, XAI methods help reveal when predictions are influenced by sensitive attributes or by features closely linked to them [37][59].

Types of Interpretability

Two primary forms of interpretability are often discussed in the literature [78][74]:

- **Intrinsic interpretability**: refers to models whose internal operations can be followed directly, such as decision trees, linear predictors, and rule-based fuzzy systems [41]. Their design allows users to trace how input features influence the output without needing separate explanation tools.
- **Post-hoc interpretability**: involves techniques that analyze a trained model with hidden internal reasoning. These include attribution methods, surrogate models, and counterfactual explanations [36][79][80]. They offer insight into prediction behaviour while leaving the original model unchanged.

Models intended to be inherently transparent are usually straightforward to inspect, though they may struggle when data relationships are complex or non-linear. In contrast, post-hoc techniques can interpret highly accurate predictive models, even if the explanations represent only an approximation of the underlying decision logic [77].

3.3.1 Local Attribution Methods

Local attribution methods aim to show how particular input features influence a specific prediction made by the model [81]. Common techniques include:

- **Gradient-based saliency**: uses the gradient of the output with respect to input features to assess how small variations in the input affect the prediction score [82].
- **IGs)**: along a path from a baseline to the actual sample, yielding a more reliable estimate of feature influence [33].
- **Perturbation-based methods**: such as LIME or SHAP, evaluate how predictions change when selected input components are modified, allowing estimation of local behavioural sensitivity [36][79].

These approaches help identify when sensitive or correlated features strongly affect the decision, supporting model audits and fairness risk analysis [83][84].

3.3.2 Strengths and Limitations

While XAI techniques make it possible to inspect how input features influence predictions, they also introduce notable limitations. Methods based on gradients can be unstable and highly sensitive to small changes in the input, which can lead

to inconsistent explanations [16, 85]. Perturbation-driven approaches offer more flexibility but often demand considerable computation and can involve assumptions that do not always reflect realistic feature behaviour [86]. Furthermore, explanations may fail to capture interactions between features, which could result in incorrect conclusions. Therefore, evaluating the reliability and robustness of explanation methods is crucial, particularly when they are used to analyze fairness-related concerns [63, 87].

3.3.3 Integration with Fairness

Connections between interpretability and algorithmic fairness are increasingly explored in recent research [83]. Attribution information can be incorporated into the training objective to discourage dependence on sensitive attributes, integrating interpretability into learning rather than applying it only afterward [84]. Rule-based systems grounded in fuzzy logic offer a practical bridge between fairness and interpretability because they allow constraints on sensitive attributes to be encoded directly into their structure. Combining attribution insights with adaptive rule-based mechanisms enables the design of models that jointly promote transparency, fairness, and strong predictive behaviour [11].

In this thesis, XAI is part of the mitigation process itself, not merely a diagnostic tool used after training. Saliency measures indicate how much each feature influences the prediction, while fuzzy or neuro-fuzzy components transform those signals into penalties embedded in the learning objective. This approach integrates interpretability into training, strengthening fairness throughout the model development rather than relying solely on post hoc adjustments. This integration moves beyond explanation alone, embedding fairness into the model's very structure. Table 3.3 presents key XAI methods and describes how each technique contributes to interpreting model decisions and assessing fairness.

Table 3.3: XAI techniques and their role in fairness.

Technique	Type	Model compatibility	Fairness insight	Strengths	Limitations
SHAP	Local / Global	Works with tree, linear, and deep models.	Shows feature impact across groups.	Model-agnostic; visual; solid theory.	Slow on large data; sample heavy.

(Continued on next page)

(Continued from previous page)

Technique	Type	Model compatibility	Fairness insight	Strengths	Limitations
LIME	Local	Any black-box model.	Explains single predictions to expose bias.	Simple, flexible, and easy to apply.	Unstable; results may vary.
IGs	Local	NN.	Captures the sensitivity of inputs to outputs.	Faithful; gradient-based.	Needs differentiability; may saturate.
Counterfactuals	Instance-based	Tabular/image models.	Reveals minimal change for the opposite outcome.	Intuitive; causal.	Computationally costly; search needed.
Grad-CAM / Saliency	Local (visual)	CNNs are multimodal.	Highlights biased attention regions.	Visual, intuitive.	Sensitive to noise; qualitative.
Concept-based (TCAV)	Global	Deep models with internal layers.	Measures conceptual influence (e.g., gender).	Connects human ideas to model logic.	Needs examples; limited scale.
Rule/Fuzzy Explanations	Global / Hybrid	Decision trees, neuro-fuzzy.	Extracts human-readable fairness rules.	Transparent; auditable.	Simplifies complex relations.
Prototype-based	Local	Image, text, and tabular data.	Compares group examples and outputs.	Easy to interpret; visual.	Needs diverse data; low precision.

3.4 Fuzzy Logic and Neuro-Fuzzy Systems

Fuzzy logic was introduced to handle uncertainty and imprecision in reasoning by allowing partial membership in sets rather than binary inclusion [88]. This framework enables statements such as "income is high" or "age is young" to be represented as degrees of truth rather than crisp categories. Fuzzy inference systems combine these graded memberships with IF–THEN rules to form decision logic that resembles human reasoning [18].

3.4.1 Classical fuzzy systems

A fuzzy inference system typically consists of four stages: fuzzification, rule evaluation, aggregation, and defuzzification [18][89]. Inputs are mapped to fuzzy sets through membership functions, rules are evaluated to determine their degree of activation, outputs are combined, and finally, a crisp prediction is produced. Because rules are expressed in natural language terms, the reasoning process is transparent and easy to audit [90].

Applications of fuzzy systems range from control engineering and signal processing to decision support in medical and financial domains [91]. While classical fuzzy systems offer clear and interpretable rule structures, their scalability is limited in high-dimensional settings, as the number of rules can increase rapidly with the number of input variables [92].

3.4.2 Neuro-Fuzzy Models

Neuro-fuzzy systems combine fuzzy inference with the learning capability of NNs, helping to address the scalability limitations of traditional fuzzy approaches [45]. The most widely adopted architecture in this category is the ANFIS. In ANFIS, each stage of fuzzy inference is mapped onto a corresponding network layer, and the parameters of membership functions and rule outputs are updated through gradient-based or hybrid optimization methods [45][93]. This structure allows the network to learn from data while still preserving a rule-based format that remains interpretable. Several upgrades to ANFIS have been proposed to enhance its flexibility. Kernel-based fuzzy sets, for example, provide nonlinear separation capabilities [12], and hierarchical fuzzy systems manage the scalability of large rule sets by decomposing them into smaller components [94]. Such developments expand the expressive power of neuro-fuzzy models while maintaining their inherently transparent reasoning process.

3.4.3 Fuzzy systems and fairness

In fairness settings (algorithmic), fuzzy logic offers two notable benefits. First, fairness constraints can be encoded as explicit rules that limit the influence of sensitive attributes [52]. For example, a rule may specify that reliance on gender or ethnicity should remain low in the decision-making process. Second, the transparency of fuzzy rules makes them suitable for auditing by stakeholders who need to understand not just the outcomes but also the reasoning behind them [74].

Recent work has explored fairness-aware fuzzy clustering [95], group-balanced rule sets [96], and hybrid systems where attribution signals guide the learning of fuzzy memberships [11]. These directions show that fuzzy logic is a promising framework for combining interpretability and fairness in a principled manner.

3.4.4 Position within this thesis

This thesis builds on the neuro-fuzzy paradigm by introducing fairness-aware extensions. In the first stage, fuzzy rules are linked with saliency-based attributions to form differentiable regularizers during training. In the second stage, ANFIS is extended with kernelized memberships and saliency-driven rules to better capture nonlinear dependencies. Finally, dynamic fuzzy rules are developed to adapt during learning, ensuring fairness interventions remain responsive to data characteristics. This progression demonstrates how classical fuzzy reasoning can be reimagined for fairness in modern NNs. Table 3.4 summarizes major fuzzy and neuro-fuzzy approaches, emphasizing their interpretability, application scope, and relevance to fairness-aware learning.

Table 3.4: Fuzzy Logic and Neuro-Fuzzy Systems.

Models / Techniques	Key ideas	Application domains	Fairness integrations	Strengths	Limitations
Classical Fuzzy Systems	IF–THEN rules with linguistic sets.	Decision support, control, and basic classification.	Transparent; rules auditable for bias.	Human-readable; interpretable.	Manual design; poor scaling to large data.

(Continued on next page)

(Continued from previous page)

Model / Technique	Key idea	Application domain	Fairness integration	Strengths	Limitations
Fuzzy Decision	Splits/leaves use fuzzy sets.	Tabular scoring, simple models.	Group-based splits reduce bias.	Clear rule paths.	Overfitting on complex data.
FCM	Soft clustering via fuzzifier m.	Exploratory or clustering tasks.	Detects bias through group membership.	Captures overlap; diagnostic use.	Needs k; sensitive to start points.
Fair FCM	Adds parity penalties to clusters.	Social, health, and education data.	Balances group representation.	Improves parity; still unsupervised.	More complex; may distort structure.
ANFIS	Learns rule weights automatically.	Regression, small/medium tasks.	Rules regularized for fairness.	Combines learning + interpretability.	Sensitive to init.; shallow depth.
Deep Neuro-Fuzzy	Fuzzy or rule layers in deep nets.	Vision, speech, time-series.	Penalize biased rule activations.	Differentiable; end-to-end.	Less interpretable in depth.
Type-2 Fuzzy Systems	Membership uncertainty bands.	Noisy or drifting data.	Encode uncertainty in group metrics.	Robust; noise-tolerant.	High compute cost; less intuitive.
Fuzzy Rule Learning	Evolves rules with GA/PSO.	Risk scoring, ML models.	Fitness includes a fairness term.	Sparse, optimized rule base.	Compute-intensive tuning.
Fuzzy Thresholds	Fuzzy decision limits on scores.	Credit, hiring, healthcare.	Adjusts results to meet fairness bounds.	No retraining; easy control.	Alters outputs; needs labels.

(Continued on next page)

(Continued from previous page)

Model / Technique	Key idea	Application domain	Fairness integration	Strengths	Limitations
Fuzzy Regularizers	Fuzzy similarity losses.	Any differentiable model.	Promotes equal group treatment.	Lightweight; general.	Needs weight tuning.
Fuzzy Controllers	Dynamic fairness constraints.	Online training.	Maintains metrics adaptively.	Stable, interpretable.	Adds hyperparameters; setup effort.

3.5 Gaps and Research Opportunities

The survey of fairness, mitigation strategies, explainability, and fuzzy systems reveals progress on many fronts but also exposes persistent gaps. These gaps motivate the contributions of this thesis.

3.5.1 Limited transparency of in-processing methods

In-processing techniques, such as adversarial debiasing and constraint-based optimization, are among the most powerful tools for reducing disparities. Yet they offer little insight into how fairness pressure is applied within the model [70][71]. Stakeholders may observe improvements in group metrics but lack an explanation of which features or representations were altered. This absence of visibility is a serious obstacle to adoption in domains where justification of model behaviour is required [15].

3.5.2 Fragility of post-hoc explanations

Post-hoc interpretability methods, including saliency maps and surrogate models, provide valuable diagnostic information but are known to be unstable [16][85]. Small changes to the input or to model parameters can produce different explanations. Worse, models can be designed in ways that preserve discriminatory patterns while presenting innocuous-looking explanations [83]. Relying solely on post-hoc XAI, therefore, risks misleading auditors and decision-makers.

3.5.3 Static nature of fairness rules

Classical fuzzy systems can transparently encode fairness rules [90], but these rules are static. They do not adjust as data distributions evolve or as the model learns new patterns. In practice, fairness interventions need to be dynamic: they should strengthen where reliance on sensitive features grows and relax where correlations diminish. Static rules cannot meet this demand for adaptability [11].

3.5.4 Fragmentation of existing approaches

Most methods target fairness or interpretability in isolation. Fairness-focused techniques may sacrifice transparency, while interpretability-focused approaches often neglect fairness constraints [52][74]. Neuro-fuzzy models have demonstrated potential to combine both, but until now, they have not been systematically extended to address fairness as a primary objective. This leaves a gap at the intersection of fairness, interpretability, and adaptability.

3.5.5 Opportunities for Integration

There is significant potential in combining attribution signals from XAI with the rule-based reasoning of fuzzy systems [83][84]. Attribution methods provide continuous, data-driven indicators of how strongly each feature influences model predictions, while fuzzy rules supply a transparent and auditable way to control that influence. A joint training setup that incorporates attribution signals into the fuzzy rule mechanism allows fairness constraints to be enforced directly during learning. Extending this approach to include kernel-based and adaptive neuro-fuzzy structures increases modeling flexibility, supporting the capture of nonlinear patterns and improving robustness to changing data distributions [12][94].

This thesis is positioned to fill these gaps. It introduces a sequence of contributions: first, the feasibility of saliency-driven fuzzy penalties; second, the refinement of ANFIS with kernelized, attribution-aware rules; third, the development of dynamic fuzzy controllers; and finally, a unified framework validated on both simulated and real-world datasets. Together, these contributions address the need for fairness-aware systems that are not only accurate but also transparent and adaptive. Table 3.5 consolidates the main research gaps identified across the literature and points toward potential directions for future integration and improvement.

This chapter has surveyed key approaches to fairness-aware learning and bias mitigation, spanning pre-processing, in-processing, and post-processing methods,

as well as recent work linking fairness and interpretability. The review shows that many existing techniques depend on external constraints or post-hoc explanations, offering limited insight into how fair decisions are actually formed. It also highlights the instability of attribution-based explanations when used to evaluate or enforce fairness. Most methods treat fairness as a fixed objective and assume stable data distributions, which limits their ability to handle evolving bias patterns. In addition, improvements in fairness are often achieved at the expense of accuracy or increased model complexity, without clear control over these trade-offs. These gaps motivate the methods developed in Chapters 5 to 8, which focus on integrating interpretability directly into the learning process and on building adaptable neuro-fuzzy frameworks for fairness-aware modeling.

Table 3.5: Gaps and Research Opportunities.

Gap or Limitation	Evidence / Source	Impact	Open Research Question	Potential Solution / Integration Idea
Limited transparency of in-processing methods	Adversarial and constrained models remain opaque.	Users cannot see how fairness is enforced.	How to embed interpretable logic within fairness learning?	Add fuzzy rule layers or symbolic reasoning for traceable control.
Fragility of post-hoc explanations	LIME/SHAP are often unstable across samples.	Explanations mislead fairness analysis.	How to make post-hoc methods stable and reliable?	Combine with fuzzy confidence or stability metrics.
Static fairness rules	Fixed constraints ignore context shifts.	Poor performance on new data.	How to build adaptive fairness mechanisms?	Use dynamic neuro-fuzzy controllers adjusting penalties.
Fragmented research	Fairness, XAI, and fuzzy logic are studied separately.	Missed synergy between goals.	How to merge fairness and interpretability in one model?	Design unified neuro-fuzzy fairness frameworks.
Single-metric optimization	Most focus on one fairness index (EO, DP, SPD).	Ignores trade-offs with accuracy.	How to assess fairness using multiple criteria?	Use Pareto or multi-objective optimization.

(Continued on next page)

(Continued from previous page)

Gap or Limitation	Evidence / Source	Impact	Open Research Question	Potential Solution / Integration Idea
Poor generalization	Models tuned for benchmarks fail on real data.	Weak fairness transfer across domains.	How to ensure fairness in unseen populations?	Apply domain adaptation with fuzzy similarity.
Lack of human feedback	Few systems include user or ethical input.	Fairness treated purely technically.	How to add human judgment to training?	Use participatory fuzzy scoring or expert feedback loops.
Ignoring uncertainty	Most treat outputs as deterministic.	Miss hidden ambiguity in fairness scores.	How to include uncertainty in evaluation?	Adopt type-2 fuzzy or probabilistic fairness measures.

Chapter 4

Research Contributions and Roadmap

The literature review has highlighted three recurring limitations: the limited transparency of in-processing methods, the fragility of post-hoc explanations, and the static nature of fairness rules. These gaps point to the need for approaches that are both interpretable and adaptive, embedding fairness directly into the training process while maintaining compatibility with modern optimization methods. The research presented in this thesis responds to this need through a sequence of four contributions, each developed and refined during the doctoral project.

4.1 Overview of Contributions

4.1.1 Paper 1: Saliency-Driven Fuzzy Penalties

The first contribution presents a training mechanism that embeds attribution signals into fuzzy rule penalties to reduce the influence of sensitive attributes during learning [11].

4.1.2 Paper 2: Kernelized and Attribution-Guided ANFIS

The second contribution builds upon the ANFIS framework by integrating kernel-based membership functions in fairness-aware learning scenarios [12].

4.1.3 Paper 3: Dynamic Fuzzy Rules

The third contribution proposes a dynamic neuro-fuzzy framework where membership functions and rule weights adapt continuously during training throughout the learning process [13].

4.1.4 Paper 4: Unified framework with real-world validation

The fourth contribution merges the earlier advances into a unified and differentiable framework to aid model diagnostics [14].

4.2 Roadmap of the Thesis

The thesis is structured to provide a clear progression from motivation to technical contributions and evaluation. Chapter 1 introduces the research motivation and objectives. Chapter 2 summarizes the theoretical principles relevant to the study. Chapter 3 reviews related literature and highlights the gap that motivates the proposed work. Chapter 4 outlines the main research contributions. Chapters 5 through 8 each detail one component of the methodology: Chapter 5 presents the saliency-based fairness penalty, Chapter 6 describes the kernel-enhanced ANFIS extension, Chapter 7 introduces a dynamic rule adaptation mechanism, and Chapter 8 integrates these elements and evaluates the full model on benchmark datasets. Chapter 9 concludes the thesis by discussing the main findings, limitations, and directions for future work.

Overall, the work advances from conceptual formulation to architectural refinements and ultimately to a unified system evaluated in realistic, fairness-dependent environments. The design aims to reconcile predictive accuracy with interpretability and fairness in a single modelling framework.

Chapter 5

Neural Networks Bias Mitigation Through Fuzzy Logic and Saliency Maps

5.1 Introduction

This chapter introduces the first technical contribution of the thesis. The idea is to incorporate attribution-derived information into the NN learning process to limit how much predictions rely on sensitive attributes. Building on the fairness and interpretability concepts discussed earlier in Chapter 2 and 3, a regularization strategy is proposed that links saliency information with fuzzy rule constraints. Saliency highlights how different input features influence the model output, while the fuzzy component converts this information into rules that discourage reliance on protected variables. The technique integrates with standard gradient-based optimization and is formulated to operate within a fully differentiable training pipeline.

5.2 Methodology

Figure 5.1 illustrates the proposed saliency-guided fuzzy regularization framework for fairness-aware learning This pipeline integrates saliency maps and fuzzy logic within the NN training loop to penalize excessive reliance on sensitive features through differentiable loss regularization.

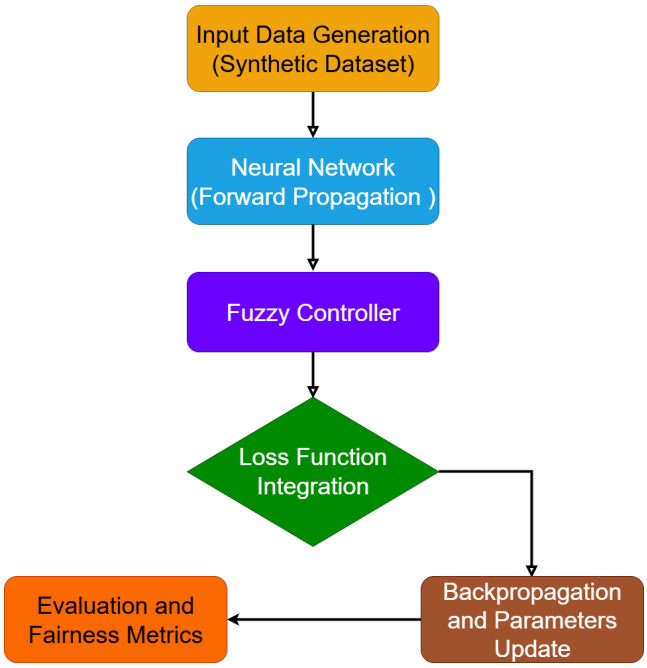

Figure 5.1: Saliency-guided fuzzy regularization pipeline for fairness-aware learning.

5.2.1 Saliency maps

Saliency maps were selected as the attribution tool because they are differentiable and integrate naturally with gradient-based optimization. For an input vector x and prediction $f(x)$, the saliency score is given by

$$S(x) = \left| \frac{\partial f(x)}{\partial x} \right|.$$

These scores measure the sensitivity of the model to each input. In this framework, saliency values corresponding to sensitive features (race, gender, and disability status) are extracted and averaged across epochs to produce a measure of feature reliance. This reliance measure is then fed into a fuzzy controller that evaluates whether the model is overly dependent on protected attributes.

5.2.2 Fuzzy controller

The fuzzy controller refines saliency information by mapping sensitivity scores into linguistic categories such as *low*, *moderate*, and *high*. Gaussian membership functions were used to implement these categories, ensuring smooth, differentiable transitions

between degrees of reliance. The centers of the Gaussians were set to represent the three levels of saliency, while the variance determined the spread.

Rules were then formulated to connect saliency profiles with bias outcomes. For example: *IF saliency of race is high AND saliency of gender is moderate, THEN bias is high.* This rule-based structure provides interpretability, as the antecedents explicitly state the conditions under which bias is penalized.

The inference process followed the standard pipeline: fuzzification, rule evaluation, aggregation, and defuzzification. Differentiable soft-min and soft-max operators were applied to maintain compatibility with backpropagation. The final defuzzified output y^* represents the degree of detected bias and is incorporated into the overall loss function. Table 5.1 summarizes the fuzzy rules used to connect feature saliency with the level of bias.

Rule	Description
R1	IF Saliency of Race is High AND Saliency of Gender is Medium THEN Bias is High
R2	IF Saliency of Race is Medium AND Saliency of Gender is Low THEN Bias is Medium
R3	IF Saliency of Race is Low AND Saliency of Gender is High THEN Bias is High
R4	IF Saliency of Gender is High AND Saliency of Disability is Medium THEN Bias is High
R5	IF Saliency of Gender is Medium AND Saliency of Disability is Low THEN Bias is Medium
R6	IF Saliency of Gender is Low AND Saliency of Disability is High THEN Bias is High
R7	IF Saliency of Disability is High AND Saliency of Race is Medium THEN Bias is High
R8	IF Saliency of Disability is Medium AND Saliency of Race is Low THEN Bias is Medium
R9	IF Saliency of Disability is Low AND Saliency of Race is High THEN Bias is High
R10	IF Saliency of Race is Low AND Saliency of Gender is Low AND Saliency of Disability is Low THEN Bias is Low

Table 5.1: Fuzzy rules linking saliency to bias control.

5.2.3 Integration with neural training

The neural architecture was a simple feed-forward classifier with one hidden layer of ten units and a sigmoid output for binary classification (higher-skilled vs. lower-skilled jobs). The loss function combined the standard cross-entropy with the fuzzy penalty:

$$L = L_{\text{CE}} + \lambda y^*,$$

where λ controls the weight of the fairness term. In this way, the network is penalized during training whenever saliency maps indicate strong reliance on sensitive attributes.

Algorithm 1: Job Class Assignment

Input: Profile x with sensitive & relevant features; sensitive filter (e.g., Race= 1, Gender= 0, Disability= 1)
Output: $JClass \in \{0, 1\}$
$score \leftarrow skill + exp + edu$;
$score \leftarrow \dfrac{score - score_{\min}}{score_{\max} - score_{\min}}$;
$jLab \leftarrow \lfloor 6 \times score \rfloor + 1$; `// map to integer label 1..7`
if *x matches the filter* **then**
  $jLab \leftarrow jLab - 2$; `// penalize biased case`
  $jLab \leftarrow \max(jLab, 0)$;
if $1 \leq jLab \leq 3$ **then**
  $JClass \leftarrow 0$; `// Lower{Skilled`
else
  $JClass \leftarrow 1$; `// Higher{Skilled`

5.2.4 Simulation data generation

To evaluate the method, a synthetic dataset was constructed to mimic a job recruitment scenario. The dataset included two types of features:

- **Sensitive features:** race, gender, and disability status. These represent attributes that should not directly influence hiring decisions.
- **Relevant features:** skill, experience, and education. These represent legitimate qualifications for a job assignment.

Applicant profiles were assigned a class label (lower-skilled vs. higher-skilled jobs) based on relevant features. Bias was introduced by modifying the labels according to specific combinations of sensitive attributes. This produced a controlled level of unfairness in the dataset, allowing evaluation of the proposed mitigation approach. Algorithm 1 describes the procedure used to assign profiles to the two job classes.

5.3 Numerical Experiments

Three models were compared:

1. **Non-regularized model**: trained with standard cross-entropy only.

Table 5.2: Mean accuracy and attribution delta across folds

Model	Accuracy	Δ (relevant−sensitive)
Non-Reg	0.921	0.205
Sal-Reg	0.918	0.248
Fuzzy-Reg	**0.919**	**0.272**

2. **Saliency-regularized model**: trained with a penalty derived from saliency values alone.

3. **Fuzzy-regularized model**: trained with the proposed saliency-guided fuzzy penalty.

All models used the same architecture and were trained with the Adam optimizer for 200 epochs under 10-fold cross-validation. Accuracy and saliency values were recorded for both sensitive and relevant features, focusing particularly on profiles affected by the bias filter. Tables 5.2 and 5.3 summarize the accuracy results together with the fairness and attribution outcomes across the evaluated models and datasets.

5.4 Results

The analysis focused on how regularization affected the relative importance of sensitive and relevant features. A two-sample t-test was first used to compare average saliency values across models. While the biased data generation mechanism produced equal contributions from relevant and sensitive features in the non-regularized case, introducing regularization altered this balance. Both saliency-based and fuzzy-based regularization increased the difference between average reliance on relevant and sensitive attributes.

An analysis of variance (ANOVA) confirmed that model type had a statistically significant effect on this difference (p¡ 0.01). Post-hoc comparisons showed that both regularized models significantly reduced reliance on sensitive features compared with the baseline, but there was no significant difference between the saliency-only and fuzzy-regularized models. Visualizations of feature importances confirmed this trend: sensitive attributes were down-weighted under regularization, while reliance shifted toward task-relevant variables. The figures in this chapter (see Figures 5.2–5.9) present the comparative results across models in terms of feature accuracy and attribution differences

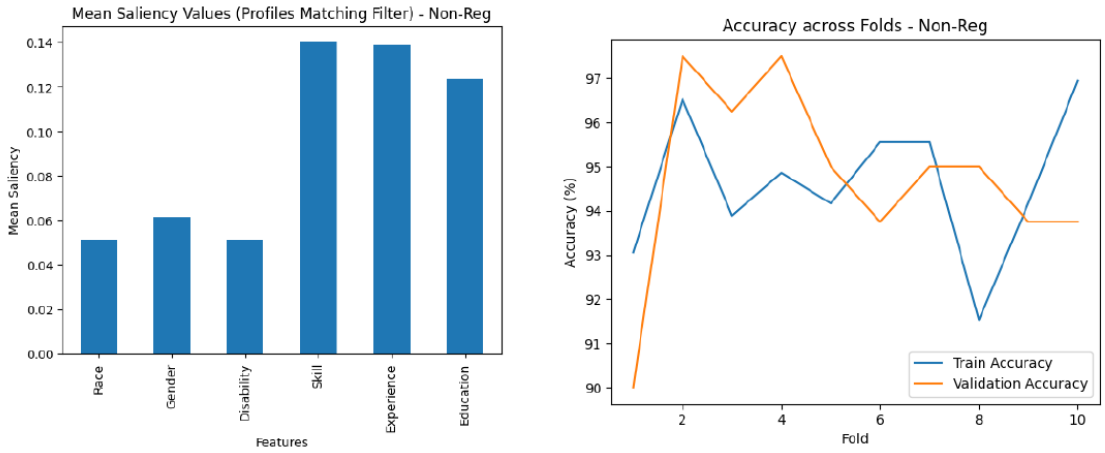

Figure 5.2: Feature-wise accuracy across models (set 1).

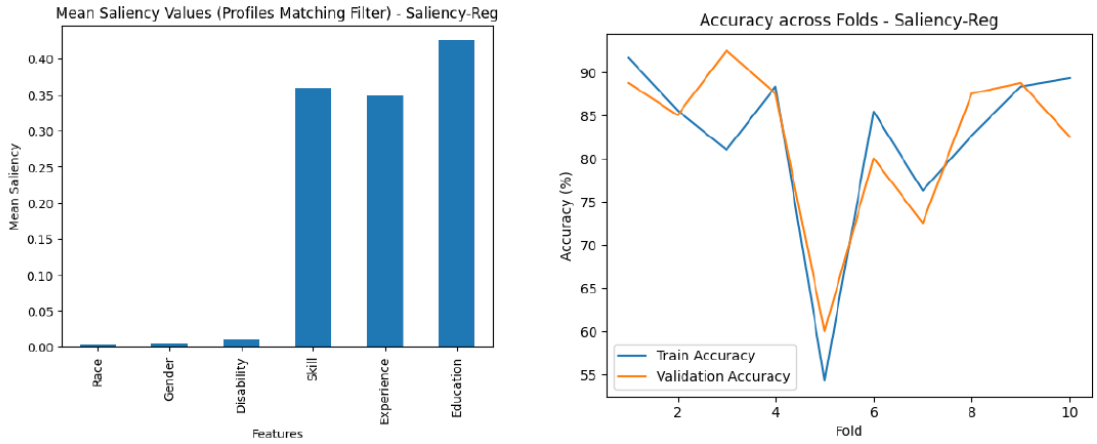

Figure 5.3: Feature-wise accuracy across models (set 2).

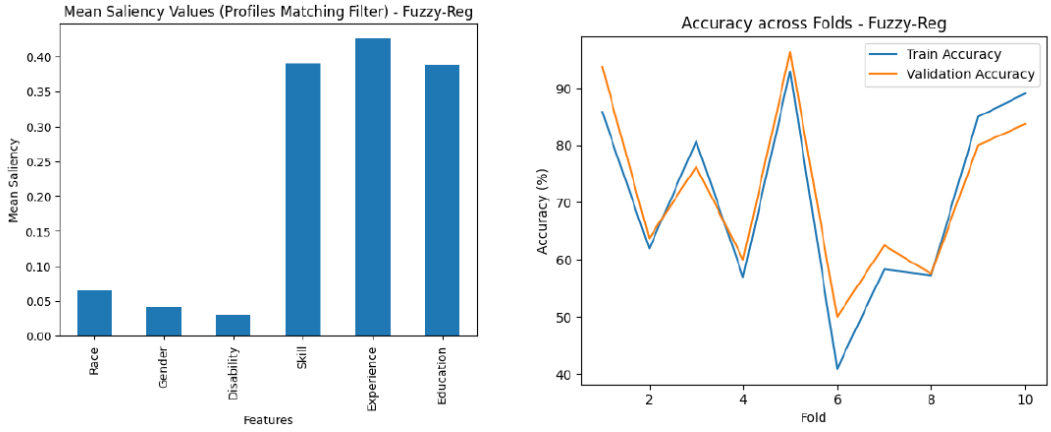

Figure 5.4: Feature-wise accuracy across models (set 3).

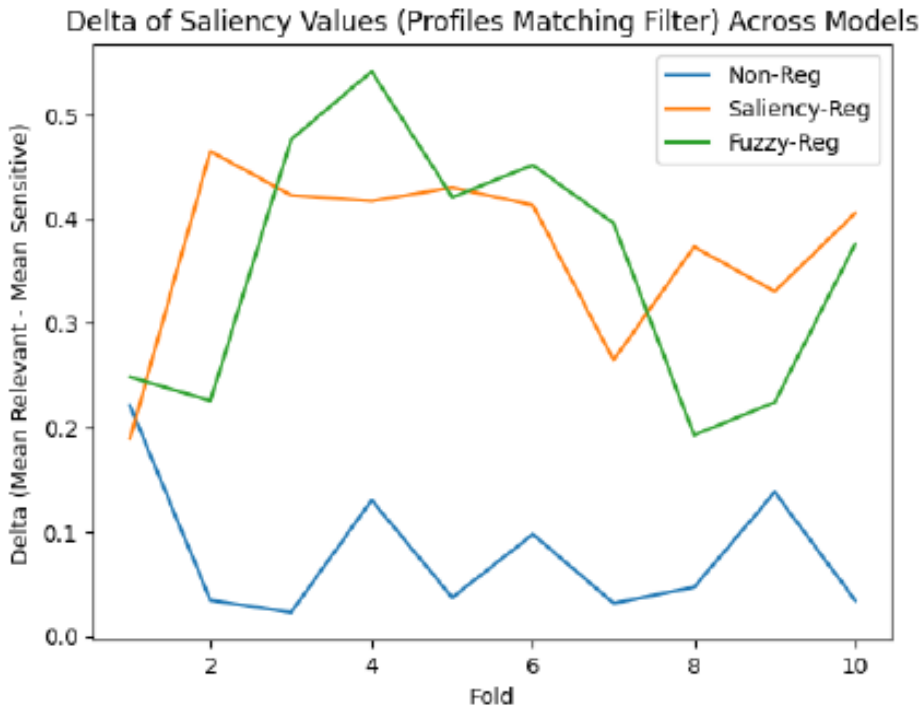

Figure 5.5: Delta profiles across subjects/conditions.

ANOVA Results on Delta of Saliency Values

Source	sum_sq	df	F	PR(>F)
C(Model)	0.539	2.0	30.056	0.0
Residual	0.242	27.0	nan	nan

Post-hoc Pairwise T-tests on Delta of Saliency Values

Comparison	p-value (uncorrected)	p-value (Bonferroni corrected)	Reject H0
Non-Reg vs Saliency-Reg	0.0	0.0	True
Non-Reg vs Fuzzy-Reg	0.0	0.0	True
Saliency-Reg vs Fuzzy-Reg	0.743	1.0	False

Figure 5.6: ANOVA with post-hoc test on Δ across models.

Non-Reg - Delta of Saliency Values (Profiles Matching Filter)

Fold	Mean Sensitive Saliency	Mean Relevant Saliency	Delta
1	0.052	0.273	0.221
2	0.019	0.053	0.035
3	0.02	0.043	0.023
4	0.104	0.234	0.131
5	0.032	0.069	0.037
6	0.054	0.152	0.098
7	0.031	0.063	0.032
8	0.096	0.143	0.047
9	0.065	0.203	0.138
10	0.075	0.109	0.034
Average	0.055	0.134	0.08

Figure 5.7: Delta across models (set 1).

Saliency-Reg - Delta of Saliency Values (Profiles Matching Filter)

Fold	Mean Sensitive Saliency	Mean Relevant Saliency	Delta
1	0.002	0.191	0.189
2	0.01	0.476	0.465
3	0.018	0.441	0.423
4	0.01	0.428	0.418
5	0.001	0.431	0.431
6	0.0	0.414	0.414
7	0.011	0.275	0.265
8	0.002	0.375	0.374
9	0.002	0.333	0.331
10	0.009	0.415	0.406
Average	0.006	0.378	0.371

Figure 5.8: Delta across models (set 2).

Table 5.3: Fairness and performance metrics on benchmark datasets

Dataset	Model	Accuracy	SPD (↓)	DI (↑)
Adult	Non-Reg	0.84	0.12	0.77
	Sal-Reg	0.83	0.08	0.85
	Fuzzy-Reg	**0.84**	**0.06**	**0.91**
COMPAS	Non-Reg	0.78	0.14	0.73
	Sal-Reg	0.77	0.09	0.82
	Fuzzy-Reg	**0.78**	**0.07**	**0.88**
German	Non-Reg	0.76	0.13	0.74
	Sal-Reg	0.75	0.09	0.81
	Fuzzy-Reg	**0.76**	**0.07**	**0.86**

Fuzzy-Reg - Delta of Saliency Values (Profiles Matching Filter)

Fold	Mean Sensitive Saliency	Mean Relevant Saliency	Delta
1	0.034	0.283	0.249
2	0.038	0.264	0.226
3	0.102	0.579	0.477
4	0.012	0.554	0.542
5	0.066	0.487	0.421
6	0.048	0.5	0.452
7	0.058	0.454	0.396
8	0.029	0.222	0.193
9	0.048	0.273	0.225
10	0.023	0.399	0.376
Average	0.046	0.402	0.356

Figure 5.9: Delta across models (set 3).

5.5 Discussion

The experimental results indicate that incorporating attribution-based penalties into NN training can reduce reliance on sensitive attributes. The proposed regularizer applies saliency information within a fuzzy rule structure, ensuring that detected sensitivity is accounted for during optimization. Unlike post-hoc analyses, which report unfairness only after model training, this approach adjusts the learning process itself.

The findings show that the predictive performance of the saliency-only and Fuzzy-regularized models are comparable. However, the fuzzy component provides a transparent rule-based representation of how fairness constraints influence the model, which supports interpretability and facilitates inspection. Stakeholders may examine

the regulations to comprehend the rationale behind the imposition of penalties and modify them to align with domain specifications. This corresponds with the overarching objective of integrating fairness and interpretability. Limitations remain: the outcomes depend on the choice of membership functions, rule definitions, and the balance parameter λ. Future extensions may explore adaptive tuning of these components to strengthen robustness.

5.6 Conclusion

This chapter introduced the initial experimental validation of the proposed approach through a controlled simulation setup. The study showed that using saliency information to guide fuzzy penalties can reduce dependence on sensitive inputs while maintaining strong predictive performance. These findings indicate that fairness-related adjustments can be embedded directly into the optimization process. The insights gained here motivate the developments in the following chapters, where the method is extended to handle more complex interactions and changing data conditions through kernel-based and adaptive neuro-fuzzy enhancements.

Chapter 6

Kernelized Neuro-Fuzzy Inference for Fairness Mitigation

6.1 Introduction

This chapter introduces the next development of the proposed framework by extending the feasibility study of Chapter 5 into a kernel-enhanced neuro-fuzzy model. The design follows the ANFIS structure, but now fairness information from attribution methods is woven directly into the rule-learning process. The aim is to incorporate fairness guidance while increasing the expressive power of the fuzzy rule layer through kernel operations.

The motivation comes from the observation made in Chapter 5: although the saliency-guided penalty reduced reliance on protected attributes, the linear constraints of the original fuzzy rules limited the model's flexibility. By introducing kernel-based membership functions, the system can better capture nonlinear links between saliency patterns and input features. This update supports a more adaptable mechanism for maintaining fairness throughout training without losing the interpretability offered by fuzzy rules.

6.2 Methodology

The proposed framework follows the classical ANFIS architecture with three layers: (1) a rule definition and evaluation layer, (2) a normalization layer, and (3) an output layer. Our contribution lies in designing two complementary sets of rules—Classification Rules (CRs) and Bias Rules (BRs)—and then extending them through kernelization.

Tables 6.1 and 6.2 list the classification rules and the corresponding bias rules defined for the features used in the model.

6.2.1 Classification Rules

CRs model the relationship between inputs (relevant and sensitive) and the target label. Gaussian membership functions define the linguistic categories (*Low*, *Medium*, *High*) for each feature, ensuring smooth differentiability. Each rule R_j produces a crisp output:

$$R_j^{(cl)} : \text{IF } x_1 \text{ is } A_{j1} \wedge \cdots \wedge x_n \text{ is } A_{jn} \text{ THEN } y_j^{(cl)} = \sum_i p_{ji} x_i + c_j.$$

Here x_i are input features, A_{ji} membership sets, and p_{ji}, c_j learnable parameters. The rule firing strength is computed as the product of memberships.

Table 6.1: Classification rule base (CRs). Crisp consequents follow $y_j^{(cl)} = \sum_i p_{ij}\, x_i + c_j$.

Rule	Description
R1	IF Age is High AND Income is Medium THEN $y_1^{(cl)}$
R2	IF Age is Medium AND Income is Low THEN $y_2^{(cl)}$
R3	IF Age is Low AND Income is High THEN $y_3^{(cl)}$
R4	IF Income is High AND Health is Medium THEN $y_4^{(cl)}$
R5	IF Income is Medium AND Health is Low THEN $y_5^{(cl)}$
R6	IF Income is Low AND Health is High THEN $y_6^{(cl)}$
R7	IF Health is High AND Age is Medium THEN $y_7^{(cl)}$
R8	IF Health is Medium AND Age is Low THEN $y_8^{(cl)}$
R9	IF Health is Low AND Age is High THEN $y_9^{(cl)}$
R10	IF Age is Low AND Income is Low AND Health is Low THEN $y_{10}^{(cl)}$
R11–R20	Analogous patterns for *Skill, Experience, Education.*

6.2.2 Bias Rules

BRs are introduced to quantify and penalize reliance on sensitive features. Instead of using raw inputs, they operate on saliency values:

$$S(s_i) = \left| \frac{\partial f(x)}{\partial s_i} \right|.$$

Rules are defined analogously:

$$R_j^{(br)} : \text{IF } S(s_1) \text{ is } A_{j1} \wedge \cdots \wedge S(s_m) \text{ is } A_{jm} \text{ THEN } y_j^{(br)} = \sum_i q_{ji} s_i + b_j.$$

These outputs do not affect classification directly but feed into a bias regularization term added to the system loss.

Table 6.2: Bias rule base using feature saliency

Rule	**Description**
B1	IF Saliency(Age) is High AND Saliency(Income) is Medium THEN Bias is High
B2	IF Saliency (Age) is Medium AND Saliency (Income) is Low THEN Bias is Medium
B3	IF Saliency (Age) is Low AND Saliency (Income) is High THEN Bias is High
B4	IF Saliency (Income) is High AND Saliency (Health) is Medium THEN Bias is High
B5	IF Saliency (Income) is Medium AND Saliency (Health) is Low THEN Bias is Medium
B6	IF Saliency (Income) is Low AND Saliency (Health) is High THEN Bias is High
B7	IF Saliency (Health) is High AND Saliency (Age) is Medium THEN Bias is High
B8	IF Saliency (Health) is Medium AND Saliency (Age) is Low THEN Bias is Medium
B9	IF Saliency (Health) is Low AND Saliency (Age) is High THEN Bias is High
B10	IF all Saliencies are Low THEN Bias is Low

6.2.3 Kernelization of Rules

To extend ANFIS expressivity, we replace fuzzy membership evaluation with kernel-based similarity functions. Two kernels were introduced:

1. **KDSE (Differentiable Soft Equivalence Kernel)**, which computes similarity between an input x and a prototype c:

$$K_{\text{DSE}}(x, c) = \sigma\left(-\frac{\|x - c\|}{\varepsilon}\right),$$

with σ the sigmoid and ε a smoothness parameter.

2. **KDSE-S (Saliency-Extended Kernel)**, which computes similarity in the saliency space:

$$K_{\text{DSE-S}}(S(x), c) = \sigma\left(-\frac{\|S(x) - c\|}{\varepsilon}\right).$$

Both kernels comply with Mercer's condition, which guarantees that they form valid inner products within a reproducing kernel Hilbert space (RKHS). When these kernels replace the standard membership functions in the rule antecedents, the model can exploit similarity relations in the data that conventional fuzzy memberships are not able to represent. Algorithm 2 outlines the procedure used to generate the biased dataset and assign class labels. Membership functions are initialized using a data-driven strategy to ensure reproducibility across runs. For each input feature, initial centers are obtained from simple clustering over the training data, while spreads are set to cover the observed value range. Gaussian membership functions are used, and rule weights are initialized uniformly to avoid introducing bias at the start of training. This initialization provides a stable starting point for subsequent optimization.

Algorithm 2: Biased Dataset and Label Assignment

Input: Sensitive features $(\text{Age}, \text{Income}, \text{Health}) \in [0,1]$; Relevant features $(\text{Skill}, \text{Experience}, \text{Education}) \in [0,1]$; weights w_1, w_2, w_3
Output: Class $C \in \{0,1\}$
$S_s \leftarrow w_1 \cdot \text{Age} + w_2 \cdot \text{Income} + w_3 \cdot \text{Health}$
$S_r \leftarrow \text{Skill} + \text{Experience} + \text{Education}$
if $S_s > 0.75$ **then**
  $L \leftarrow 0$ `// penalize highly sensitive profiles`
else
  Map S_r to a level L `// fair contribution of relevant score`
if $L \leq 3$ **then**
  $C \leftarrow 0$
else
  $C \leftarrow 1$

6.2.4 Normalization and Output Layers

The firing strengths of the rules are normalized over both CRs and BRs. The outputs associated with the main task, $y^{(c)}$, are combined to form the prediction $\hat{y}$, while the outputs linked to fairness, $y^{(b)}$, are aggregated into a bias indicator y^* that contributes to the regularization term. Both branches are differentiable, and their values are confined to the range $[0,1]$.

6.3 Learning and Optimization

The system is trained using gradient descent, with a composite loss function:

$$L = L_{\text{CE}} + \lambda_{\text{bias}}\, E[y^*] + \lambda_{\text{saliency}}\, E[|\nabla_{\text{sensitive}}\hat{y}|],$$

where: - L_{CE} is cross-entropy for classification, - $E[y^*]$ penalizes high bias scores, - $E[|\nabla_{\text{sensitive}}\hat{y}|]$ penalizes strong gradients with respect to sensitive features.

The hyperparameters λ_{bias} and $\lambda_{\text{saliency}}$ balance fairness and accuracy. Algorithm 3 describes the training procedure that incorporates both saliency and bias regularization.

Algorithm 3: Training with Saliency and Bias Regularization

Input: Parameters θ, batch $\{(x_n, y_n)\}$, weights $\lambda_{\text{bias}}, \lambda_{\text{sal}}$
for *each batch* **do**
  Compute CR activations α_j and BR activations β_j (kernelized where applicable)
  $\hat{y} \leftarrow$ aggregate CR consequents with normalized α_j
  $y^* \leftarrow$ aggregate BR consequents with normalized β_j
  Estimate saliency on sensitive features $S(x)$ and its average $E[|\nabla_{\text{sensitive}}\hat{y}|]$
  $L \leftarrow L_{CE}(\hat{y}, y) + \lambda_{\text{bias}} E[y^*] + \lambda_{\text{sal}} E[|\nabla_{\text{sensitive}}\hat{y}|]$
  Update θ by gradient descent on L

6.4 Dataset

A synthetic job-classification dataset was generated for the experiments. Each record contained two groups of features:

- Relevant attributes, including abilities, experience, and education.
- Sensitive attributes, such as age, income level, and health status.

A controlled scoring rule was introduced to embed unfairness into the labels. Profiles with high sensitivity scores were intentionally downgraded, even when their relevant qualifications were strong, creating discriminatory outcomes. This setup allowed fairness-aware methods to be evaluated in a controlled environment where the source of bias was fully known.

6.5 Results

Three configurations of the model were examined:

1. A baseline ANFIS without any fairness constraints. 2. ANFIS enhanced with saliency-driven penalties applied to rule firing strengths. 3. A kernel-based neuro-fuzzy model with density-based similarity measures. The baseline system exhibited the highest average saliency score for sensitive attributes (0.0467), suggesting that its predictions were more influenced by protected features. When bias-related rules were included, the saliency value dropped to 0.0386. The kernel-based extension reduced it further to 0.0296. Regarding the reassignment of biased labels, the correction proportion (delta) increased from 0.247 in the baseline model to 0.402 with the extended version and 0.443 with the kernelized version. These outcomes demonstrate that the kernel-based system is more effective in reducing dependency on

sensitive characteristics while improving the ability to correct biased label assignments. Table 6.3 shows the average saliency on sensitive attributes and the corresponding reassignment differences for each model.

Table 6.3: Average sensitive-feature saliency and reassignment rate.

Model	**Sal(Age)**	**Sal(Inc)**	**Sal(Health)**	**Avg(SalSens)**	**Δ**
No Reg	0.0416	0.0424	0.0560	0.0467	0.247
BaseExt.	0.0310	0.0447	0.0402	0.0386	0.402
Kernel	0.0232	0.0288	0.0366	0.0296	0.443

6.6 Discussion

The findings indicate that incorporating kernel functions into the neuro-fuzzy framework decreases sensitivity toward protected features without impairing model accuracy. Introducing fairness-related rules gives the system a controllable way to moderate how rule activations respond to sensitive variables. In addition, kernel-based reasoning better represents complex relationships that go beyond the linear dependencies handled by conventional ANFIS.

This enhancement effectively resolves the main drawback observed in the previous chapter, where rule interactions were constrained to linear effects. By embedding kernel functions, the model maintains a clear and interpretable rule structure while gaining the flexibility needed to address fairness in more challenging data scenarios.

6.7 Conclusion

This chapter presented an extended ANFIS model that integrates fairness-driven rule adjustments to reduce reliance on protected characteristics. Experiments using a synthetic benchmark demonstrated lower saliency scores for sensitive attributes and higher rates of correctly adjusted labels. These outcomes confirm that the kernelized design strengthens fairness behavior while preserving interpretability. The developments in this chapter form the basis for the next stage of the research, where rule parameters and membership functions adapt during training to handle changing data trends over time.

Chapter 7

Dynamic Neuro-Fuzzy Regularization for Fairness-Aware Neural Networks Using Saliency Attribution

7.1 Introduction

This chapter presents the third contribution of the thesis. The method developed here incorporates saliency information into a dynamic neuro-fuzzy controller. Rather than using a fixed rule structure, the controller adjusts rule weights and membership functions during training. This adaptability allows the model to reflect changes in the relationships between inputs and outputs as learning progresses. The overall aim is to limit the influence of sensitive attributes while preserving the clarity and structure of fuzzy-rule reasoning.

The experiments use a binary classification task, consistent with the earlier chapters. Decision-related attributes serve as the main predictors, while protected variables such as age, gender, or disability status are included only to track and control their impact on the outcomes. This setup enables a fair and transparent assessment of whether the dynamic controller can reduce biased behaviour without degrading predictive performance or obscuring the interpretability of the model.

7.2 Methodology

7.2.1 Overview

The proposed training framework is organized into three integrated components:

1. A feed-forward network that produces predictions $\hat{y} = f(x)$ from the input sample.
2. An attribution module that evaluates the influence of each input feature through the sensitivity measure $S(x) = \|\nabla_x f(x)\|$.
3. A neuro-fuzzy controller that takes the sensitivities of protected attributes and converts them into a bounded penalty term $y^b \in [0, 1]$ to regulate fairness during training.

All components remain differentiable, enabling end-to-end optimization through standard gradient-based learning.

7.2.2 Saliency input to the controller

Given an input vector $x = [r, s]$ with relevant r and sensitive s, the saliency vector $S(x) \in \mathbb{R}^d$ is computed by backpropagating the output with respect to x. The controller receives only the d_s components associated with sensitive features,

$$s^{\text{sal}} = \big(|\tfrac{\partial f}{\partial s_1}|, \ldots, |\tfrac{\partial f}{\partial s_{d_s}}|\big),$$

which serve as evidence of reliance on protected information.

7.2.3 Dynamic fuzzy rule base

Each sensitive saliency s_i^{sal} is fuzzified into three linguistic terms *Low*, *Medium*, *High* using Gaussian membership functions

$$\mu_A(s_i^{\text{sal}}) = \exp\Big(-\frac{(s_i^{\text{sal}} - c_{A,i})^2}{2\sigma_{A,i}^2}\Big),$$

with centers $c_{A,i}$ and spreads $\sigma_{A,i}$ learned during training. The rule base enumerates combinations over the d_s sensitive features:

$$R_j: \text{ IF } s_1^{\text{sal}} \text{ is } A_{j1} \ \wedge \cdots \wedge \ s_{d_s}^{\text{sal}} \text{ is } A_{jd_s} \ \text{ THEN bias is } B_j.$$

Rule activation is the product of antecedent memberships. The full dynamic fuzzy rule-evaluation process is detailed in Algorithm 4.

$$\alpha_j = \prod_{i=1}^{d_s} \mu_{A_{ji}}(s_i^{\text{sal}}),$$

followed by normalization $\bar{\alpha}_j = \alpha_j / \sum_k \alpha_k$. Each rule has a learnable consequent center $c_j \in [0, 1]$ that represents the bias level suggested by that rule. The controller output is a smooth weighted average,

$$y^* = \sum_j \bar{\alpha}_j \, c_j,$$

The corresponding fuzzy rule base used to determine bias levels is summarized in Table 7.1. Which plays the role of an interpretable, differentiable bias score.

Algorithm 4: Dynamic Fuzzy Rule Evaluation for Bias Penalty

Input: Saliency vector $\mathbf{s} \in \mathbb{R}^d$ for d sensitive features (e.g., Race, Gender, Disability)
Output: Bias estimate y^*
for *each feature* s_i **do**
  Compute membership degrees $\mu_A(s_i)$ for $A \in \{\text{Low}, \text{Medium}, \text{High}\}$
Construct all 3^d fuzzy rules R_j: "IF s_1 is A_j^1 AND s_2 is A_j^2 AND s_3 is A_j^3 THEN Bias is B_j
" **for** *each rule* R_j **do**
  Compute activation $\alpha_j = \prod_i \mu_{A_j^i}(s_i)$
Normalize activations $\hat{\alpha}_j = \alpha_j / \sum_j \alpha_j$
Learn output centers c_j for each rule R_j
Compute defuzzified bias: $y^* = \sum_j \hat{\alpha}_j c_j$

The dynamic rule adaptation process starts from the same initialization described in Chapter 6, after which membership parameters and rule weights are updated during training. To ensure stability during training, dynamic rule updates are subject to explicit stopping conditions. Rule adaptation is halted when a predefined maximum number of updates is reached, or when the bias index converges and shows no further meaningful change. In addition, updates are stopped if no improvement in the validation fairness metric is observed for a fixed number of consecutive epochs. These criteria prevent unnecessary rule oscillations and ensure stable convergence.

Table 7.1: Dynamic fuzzy rules for bias determination.

Rule	Description
R1	IF Saliency(Race) is Low AND Saliency(Gender) is Low AND Saliency(Disability) is Low THEN Bias is Low
R2	IF Saliency(Race) is Low AND Saliency(Gender) is Low AND Saliency(Disability) is Medium THEN Bias is Low
R3	IF Saliency(Race) is Low AND Saliency(Gender) is Low AND Saliency(Disability) is High THEN Bias is Medium
R4	IF Saliency(Race) is Low AND Saliency(Gender) is Medium AND Saliency(Disability) is Low THEN Bias is Low
R5	IF Saliency(Race) is Low AND Saliency(Gender) is Medium AND Saliency(Disability) is Medium THEN Bias is Medium
R6	IF Saliency(Race) is Low AND Saliency(Gender) is Medium AND Saliency(Disability) is High THEN Bias is High
R7	IF Saliency(Race) is Low AND Saliency(Gender) is High AND Saliency(Disability) is Low THEN Bias is Medium
R8	IF Saliency(Race) is Low AND Saliency(Gender) is High AND Saliency(Disability) is Medium THEN Bias is Medium
R9	IF Saliency(Race) is Low AND Saliency(Gender) is High AND Saliency(Disability) is High THEN Bias is High
R10	IF Saliency(Race) is Medium AND Saliency(Gender) is Low AND Saliency(Disability) is Low THEN Bias is Low
R11	IF Saliency(Race) is Medium AND Saliency(Gender) is Low AND Saliency(Disability) is Medium THEN Bias is Medium
R12	IF Saliency(Race) is Medium AND Saliency(Gender) is Low AND Saliency(Disability) is High THEN Bias is High
R13	IF Saliency(Race) is Medium AND Saliency(Gender) is Medium AND Saliency(Disability) is Low THEN Bias is Medium
R14	IF Saliency(Race) is Medium AND Saliency(Gender) is Medium AND Saliency(Disability) is Medium THEN Bias is High
R15	IF Saliency(Race) is Medium AND Saliency(Gender) is Medium AND Saliency(Disability) is High THEN Bias is High

7.2.4 Training objective

The network is trained with a composite loss

$$L = L_{\mathrm{CE}}(\hat{y}, y) \; + \; \lambda \, y^*,$$

where L_{CE} is binary cross-entropy and $\lambda > 0$ controls the strength of the fairness penalty. When the model places high saliency on sensitive inputs, the controller output increases and the penalty grows, nudging gradient updates away from sensitive reliance.

7.2.5 Practical notes

- **Initialization**: membership centers are initialized to cover the empirical range of saliency values; spreads start broad and tighten during training.
- **Complexity**: with d_s sensitive features and three terms per feature, the rule count is 3^{d_s}. For $d_s = 3$ this remains tractable. For larger d_s, sparse rule selection can be adopted (not needed here).
- **Stability**: to avoid vanishing activations, spreads have a positive lower bound; gradients are clipped at the controller input.

The dynamic rule-updating process is summarized in Algorithm 5, highlighting the interaction between attribution feedback, bias detection, and adaptive rule modification.

7.3 Experimental Setup

7.3.1 Data generation

The dataset mirrors a screening task. Each instance has six inputs: three relevant and three sensitive. A clean label is produced from the relevant features. A bias filter reduces the score of specific sensitive profiles before thresholding, creating unfair labels by design. This setup makes it possible to measure whether training reduces attribution to sensitive inputs without destroying the signal from the relevant ones. Figure 7.1 outlines the workflow of the neuro-fuzzy regularization framework.

Algorithm 5: Dynamic Rule Update

Input: Training data $\mathcal{D}$, initial fuzzy rules $\mathcal{R}$, membership parameters Θ, regularization weights λ_f, λ_r
Output: Updated fuzzy rules and membership parameters
for *each training epoch* **do**
  Compute feature attributions using the selected attribution method;
  Compute bias index from attribution scores of sensitive features;
  if *bias index exceeds predefined threshold* **then**
    Trigger rule update;
    Adjust membership parameters Θ to reduce reliance on sensitive features;
    Update rule weights using fairness regularization;
  Perform gradient-based parameter update;
  if *stability criterion is satisfied* **then**
    Stop rule updates;

7.3.2 Models and training protocol

We compare three conditions:

1. **Baseline**: standard network trained with cross-entropy only.
2. **Saliency-regularized**: cross-entropy plus a penalty proportional to the average sensitive saliency.
3. **Dynamic fuzzy**: the method above with y^* from the learned controller.

All models use the same architecture as in previous chapters: input dimension 6, one hidden layer with ReLU units, and a sigmoid output. Training uses Adam, a fixed batch size, and 200 epochs. Validation is by 10-fold cross-validation with folds preserved across methods.

7.3.3 Evaluation

Performance is reported on two axes:

- **Accuracy**: standard classification accuracy per fold.
- **Attribution-based fairness**: average saliency on sensitive features and the *delta* between mean saliency on relevant vs. sensitive inputs.

To assess consistency, we summarize fold means and use standard tests on the per-fold statistics. Visual inspection of saliency profiles complements the numerical analysis.

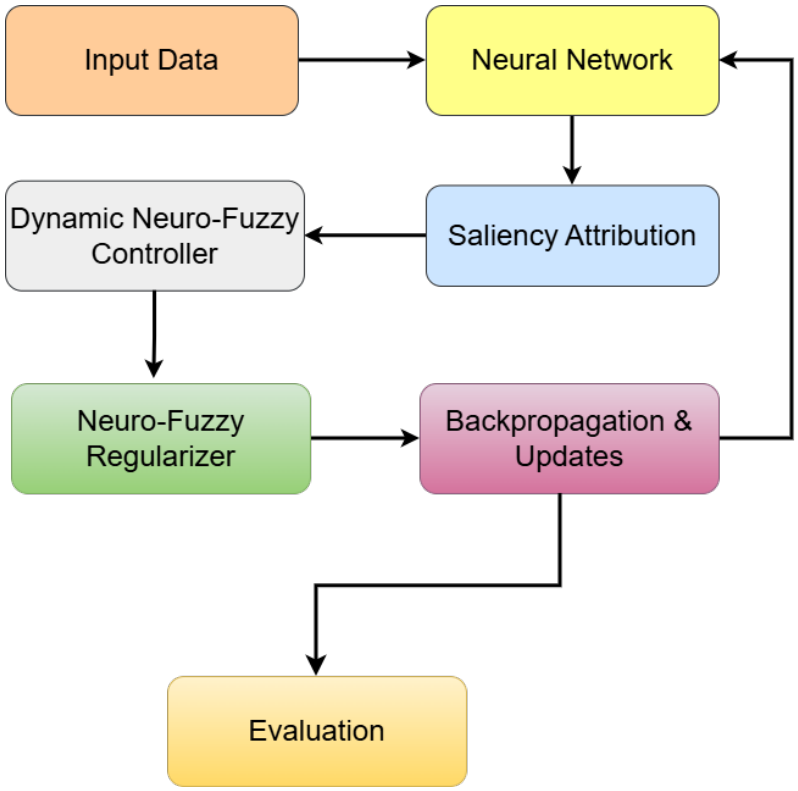

Figure 7.1: Workflow of the proposed dynamic neuro-fuzzy regularization framework.

7.4 Results

Figures 7.2–7.8 present the validation performance, saliency behavior, attribution differences, and rule activations for the evaluated models. In 7.8, the horizontal axis corresponds to input features, while the vertical axis represents fuzzy rules. Color intensity reflects the magnitude of the rule activation or contribution: darker regions indicate stronger influence on the model output, whereas lighter regions denote weaker or negligible effects. Concentrations of darker cells around sensitive features highlight rules that contribute most to group-level disparities, providing a visual indication of bias concentration within the rule base.

7.4.1 Classification

Across folds, the baseline gives the upper bound on raw accuracy, as expected when no constraints are applied. The dynamic fuzzy model maintains accuracy in a narrow range below the baseline and above a plain saliency penalty, indicating that adaptive rule learning avoids over-penalizing the network. Training curves are smooth, and early stopping is not required at the chosen epoch budget.

7.4.2 Attribution-based fairness

Both regularized models reduce sensitive saliency compared to the baseline. The saliency-only penalty produces the most aggressive reduction, sometimes at the cost of a larger drop in accuracy. The dynamic fuzzy controller achieves a balanced outcome: sensitive saliency decreases, the relevant-minus-sensitive delta narrows, and accuracy remains competitive. Per-fold plots show the same qualitative pattern: after regularization, attribution concentrates on task-relevant inputs and moves away from protected ones.

7.4.3 Learned rules and interpretability

The controller's rule activations are straight forward to audit. Rules whose antecedents include *High* saliency on protected attributes receive higher activation early in training and then fade as the network shifts reliance toward relevant inputs. Consequent centers adapt accordingly. This gives a clear narrative for what the penalty is doing at each epoch and provides artifacts that can be reviewed with stakeholders.

7.4.4 Ablations

Two simple ablations help interpret the mechanism:

- **Penalty strength** λ: small values have little effect; large values can underfit. A mid-range value produces the best fairness–accuracy balance.
- **Static vs. dynamic**: freezing membership parameters removes much of the benefit. Learning the centers and spreads improves stability across folds and reduces sensitivity to the initial saliency scale.

Table 7.2: Mean accuracy and attribution delta summary

Model	Accuracy (%)	Sensitive saliency	Delta
Baseline	87.3	0.045	0.012
Saliency penalty	84.9	0.031	0.028
Dynamic fuzzy	86.8	0.029	0.035

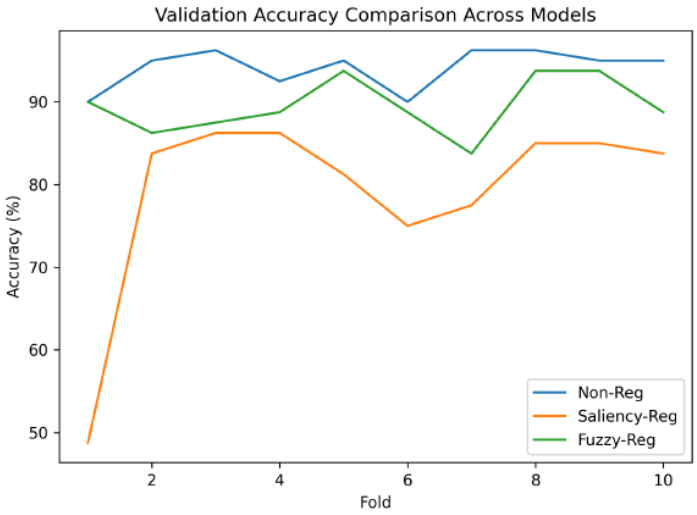

Figure 7.2: Validation accuracy across models (Non-Reg, Saliency-Reg, Fuzzy-Reg).

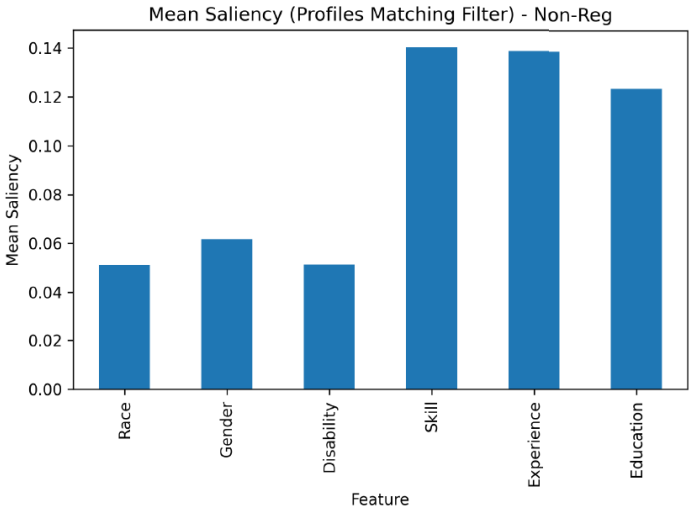

Figure 7.3: Saliency map of the Non-Regularized model.

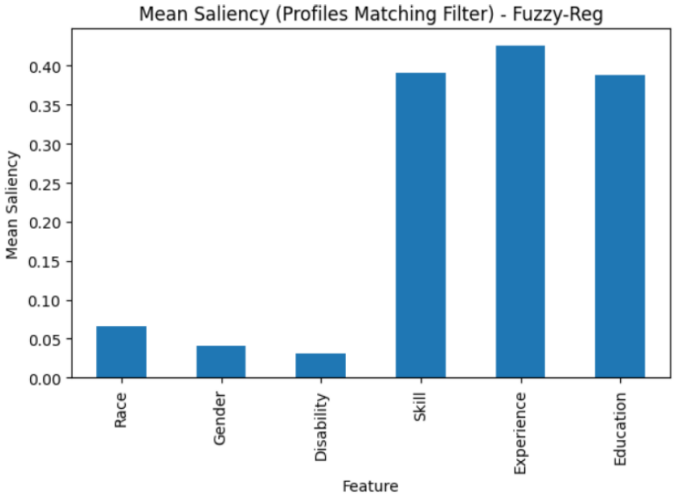

Figure 7.4: Saliency map of the Fuzzy-Regularized model.

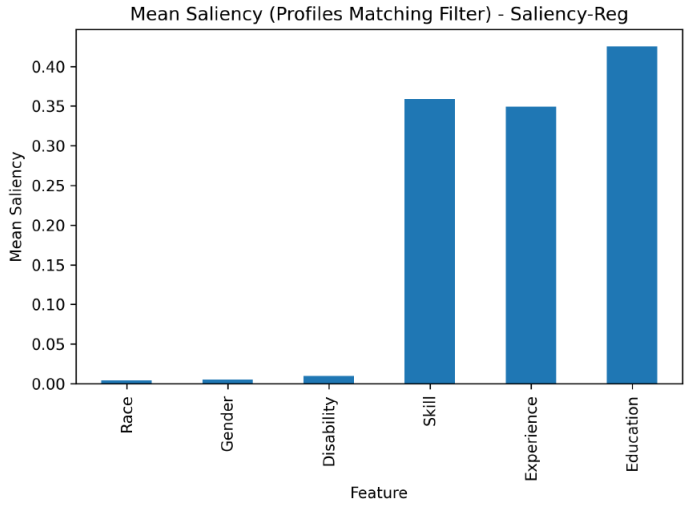

Figure 7.5: Saliency map for the Saliency-Regularized model.

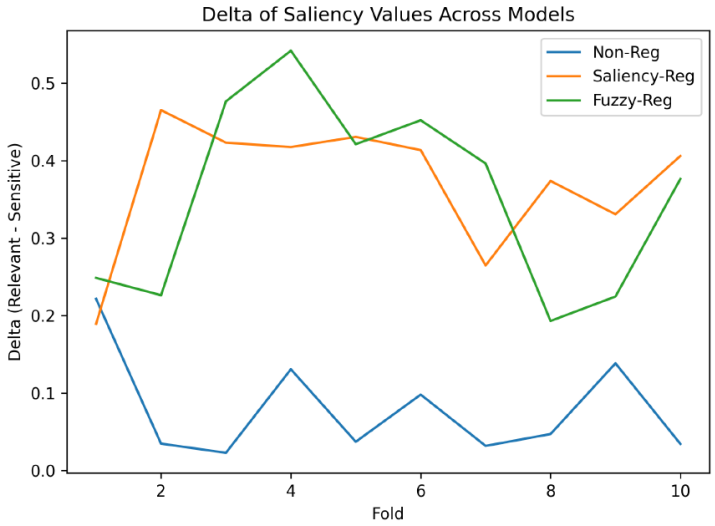

Figure 7.6: Delta of saliency (relevant minus sensitive) across models.

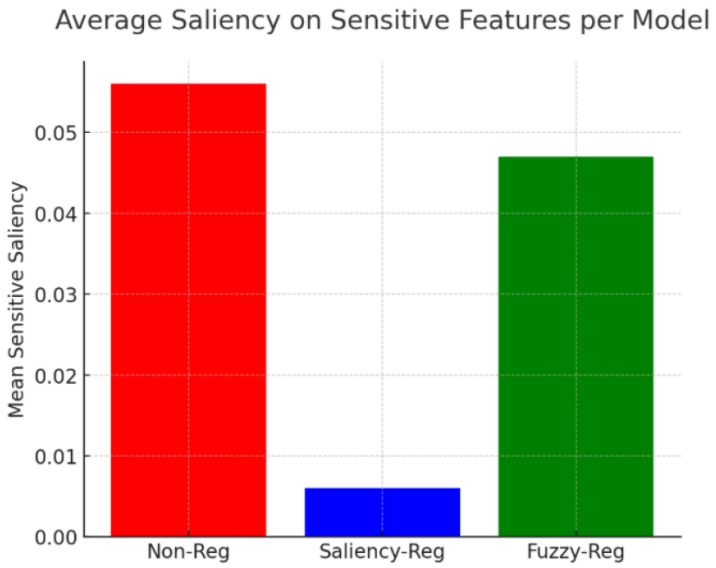

Figure 7.7: Average sensitive-feature saliency per model (lower is better).

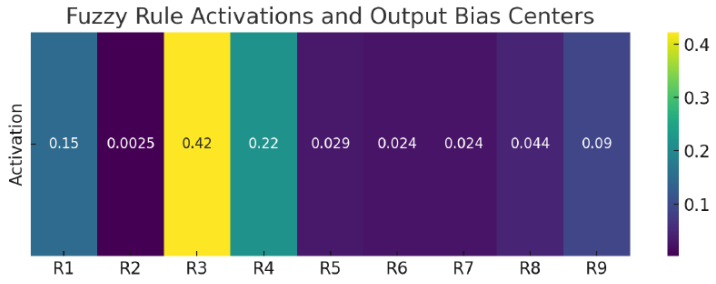

Figure 7.8: Fuzzy rule activation heatmap with output bias centers.

7.5 Discussion

The dynamic controller turns attribution into a live training signal. Compared to using a fixed rulebook, learning both antecedents and consequents lets the system adapt to the saliency scale and to the interactions between sensitive attributes observed during training. The outcome is not only a reduction in sensitive reliance but also an audit trail: rule weights, membership trajectories, and per-epoch bias estimates can be logged and reviewed. Table 7.2 summarizes the accuracy and attribution differences achieved by the compared models.

There are limits. The full 3^{d_s} rule set grows quickly with the number of sensitive features. In larger problems, sparse rule selection or hierarchical designs would be needed. Attribution itself can be noisy; smoothing across mini-batches helps, and clipping avoids spikes. Finally, while the controller improves interpretability relative to opaque debiasing, it still introduces hyperparameters that require tuning.

Reproducibility Notes

To ensure clarity and reproducibility, the following details describe how the experiments in this chapter were conducted:

- **Dataset generation**: Each instance contained three relevant features (skills, experience, and education) and three sensitive features (race, gender, and disability). Labels were assigned based on task-relevant attributes, after which a biasing function altered outcomes for particular sensitive attribute patterns to simulate discriminatory behavior. A fixed random seed was used to guarantee consistent replication of the dataset.

- **Model architecture**: The classifier consisted of a feed-forward network with one hidden layer of ten ReLU neurons and a single sigmoid output. This configuration was kept unchanged for every experiment.

- **Training protocol**: The Adam optimizer was used with a learning rate of 10^{-3}, a batch size of 32, and 200 training epochs. The fairness regularization strength λ was chosen from $\{0.01, 0.05, 0.1, 0.2\}$ based on validation results.

- **Cross-validation**: A ten-fold cross-validation strategy was adopted. The same folds were used for all models to support paired comparisons.

- **Evaluation metrics**: Accuracy, average saliency of sensitive features, and the saliency difference ("delta") between relevant and sensitive features were

computed per fold. A two-sample t-test was used to compare the fold-level deltas.

7.6 Conclusion

This chapter introduced a dynamic neuro-fuzzy regularizer that learns how to penalize reliance on sensitive features during training. The method reduces attribution on protected inputs while preserving accuracy and producing interpretable artifacts about how fairness pressure is applied. It complements the feasibility study in Chapter 5 and the kernelized extension in Chapter 6 by adding adaptability. The next chapter integrates the contributions into a unified framework and evaluates them side by side on the full set of simulated tasks, together with a real-data case study that further validates the approach under practical conditions.

Chapter 8

In-Processing Neuro-Fuzzy Approaches for Bias Mitigation

8.1 Introduction

This chapter presents the framework in which the individual contributions developed in the previous chapters are integrated into a single framework and evaluated. The integrated approach is evaluated on both synthetic and real-world datasets to demonstrate its consistency and scalability.

Building on the earlier contributions—saliency-based fuzzy regularization from Chapter 5, the kernelized ANFIS model from Chapter 6, and the dynamic neuro-fuzzy controller from Chapter 7, this work consolidates their strengths into a coherent bias mitigation pipeline. Each of the preceding models showed feasibility when tested independently; this chapter focuses on merging them into a deployable, end-to-end solution.

The proposed framework combines attribution signals with fuzzy reasoning in both fixed and adaptive forms. The effectiveness of the approach is demonstrated using both controlled simulation studies and a practical decision-making dataset. This progression shows that the method moves beyond initial proof-of-concept experiments and can be applied more broadly within fairness-aware learning contexts.

Figure 8.1 provides an overview of how the methodological components introduced in the preceding chapters are combined within the unified framework. The saliency-based fairness penalty developed in Chapter 5, the kernelized ANFIS representation presented in Chapter 6, and the dynamic rule adaptation mechanism proposed in Chapter 7 each contribute complementary elements to the final model. Their integration enables the unified neuro-fuzzy framework to jointly address interpretability,

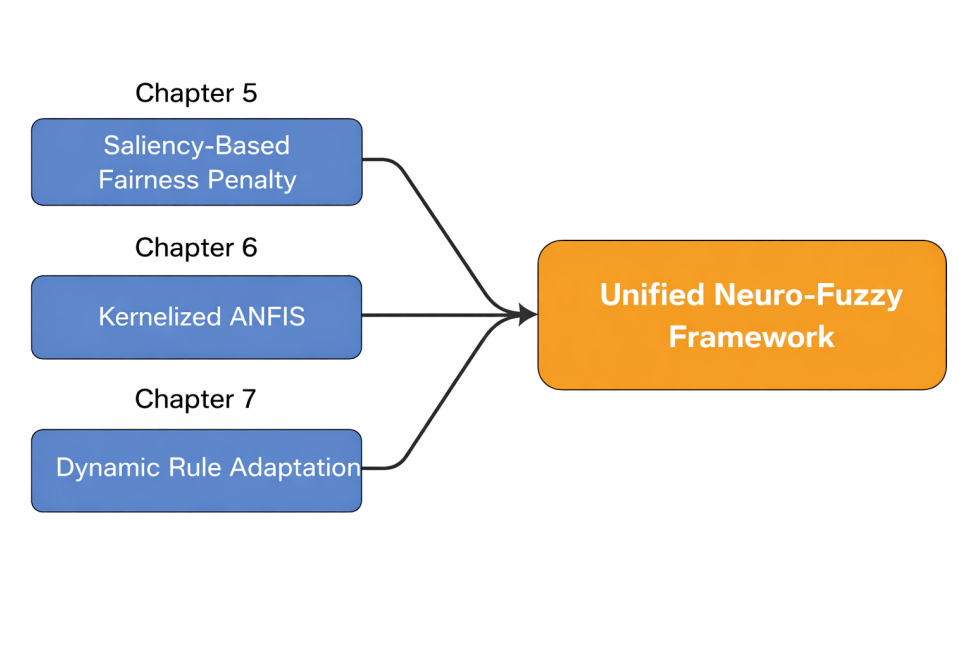

Figure 8.1: Connection between the methodological components developed in Chapters 5–7 and their integration into the unified neuro-fuzzy framework.

adaptability, and fairness within a single training process, which is evaluated in the remainder of this chapter.

8.2 Methodology

8.2.1 Framework overview

The proposed framework consists of three key components:

- **Base predictor:** a feed-forward NN trained using the standard prediction loss.
- **Attribution monitor:** computes feature-level sensitivities, allowing us to quantify how much each input contributes to the network output.
- **Fuzzy fairness controller:** translates the sensitivities of protected attributes into a penalty term that is incorporated into the overall training objective.

The fuzzy controller can operate in three configurations:

1. *Static penalty:* applies a fixed penalty based on the average sensitivity of protected attributes.

2. *Kernelized ANFIS:*2. enhances the fuzzy inference structure with nonlinear kernels to model more complex interactions.

3. *Dynamic controller:* updates rule strengths and membership parameters throughout training so that the system adapts to changing attribution patterns.

where L_{task} represents the standard prediction loss and $S(x)$ denotes the sensitivities associated with protected inputs. The regularization coefficient λ balances predictive accuracy and fairness within the unified training objective. The algorithm is encapsulated in 6.

Algorithm 6: Neuro-Fuzzy Training Loop

Input: Training set $\{(x_i, y_i)\}_{i=1}^{N}$; learning rate η; bias weight λ_b; fuzzy penalty weight λ_f
for *epoch* $= 1, \ldots, E$ **do**
  for *mini-batch* B **do**
    Compute predictions $\hat{y} \leftarrow f_\theta(x)$
    Estimate saliency $S(x) \leftarrow \|\nabla_x \hat{y}\|$
    Evaluate fuzzy bias $y^* \leftarrow \textsc{FuzzyController}(S(x))$
    Compute total loss $L \leftarrow L_{\text{CE}} + \lambda_b\, y^* + \lambda_f\, \textsc{FuzzyPenalty}(S)$
    Update parameters $\theta \leftarrow \theta - \eta \nabla_\theta L$

8.2.2 Dataset design

Two categories of data are used:

- **Synthetic simulation**: data created with intentionally introduced bias. The protected attributes do not affect the true labels, but they influence the assigned outcomes, enabling a controlled evaluation of fairness interventions.

- **Real-world dataset**: a publicly accessible dataset representing decision-making scenarios where fairness is a known concern. It includes predictor variables along with protected attributes such as gender and ethnicity. Full dataset details are provided in the referenced article. Table 8.1 provides an overview of the datasets, protected attributes, and group definitions used in the experiments.

This dual setup allows both fine-grained analysis of fairness mechanisms in simulation and external validation in real data.

Table 8.1: Dataset summary and group definitions.

ID	#F	Protected	Sample Size	Task	$Y = 1$	Priv.	Unpriv.
D1	14	Gen., Race	1000	Inc. Pred.	> 50K	White Males	Women, Min.
D2	6	Race	1000	Recid. Pred.	Low Risk	White Indiv.	Black Indiv.
D3	20	Gen., Age	1000	Cr. App.	Good Credit	Young, Empl.	Old, Unst.
D4	6	Health, Age, Inc.	1000	Skill Pred.	eligible, applicant	Young, HighInc, Healthy	Elderly, Low Inc., Poor Health

8.3 Experimental Setup

Models were implemented with the same base architecture as in previous chapters: one hidden layer with ReLU activation, followed by an output layer appropriate for the task. The Adam optimizer was used, with hyperparameters tuned via grid search on the validation folds. Regularization strength λ was selected from $\{0.01, 0.05, 0.1, 0.2\}$. For the real-world dataset, preprocessing ensured consistent scaling and removal of incomplete records. Ten-fold cross-validation was applied in all cases. Algorithm 7 shows the back-propagation procedure adjusted with fairness correction. Figure 8.2 illustrates the main workflow of the framework.

Performance was evaluated on two axes:

- **Predictive utility**: classification accuracy and F1-score.
- **Fairness metrics**: saliency share on sensitive features, DP difference, and equal opportunity difference.

Algorithm 7: Fairness-Aware Back-propagation

Input: Base gradient $\nabla_\theta L_{\text{CE}}$, fuzzy bias signal y^*, correction rate β
Output: Adjusted gradient $\nabla_\theta L_{\text{fair}}$
Compute sensitivity term $g_s = \nabla_\theta y^*$
Apply correction $\nabla_\theta L_{\text{fair}} = \nabla_\theta L_{\text{CE}} + \beta g_s$
Normalize gradient magnitude
Update weights $\theta \leftarrow \theta - \eta \nabla_\theta L_{\text{fair}}$

8.3.1 Sensitivity to Regularization Parameters

The proposed framework includes two regularization parameters, λ_f and λ_r which control the strength of fairness-related penalties and rule-based regularization within

Table 8.2: Computational complexity and training cost comparison between baseline models and proposed neuro-fuzzy methods.

Model	Parameters	Batch Size	Time / Epoch	Total Time	Hardware
Baseline NN	∼0.5M	128	0.6 s	18 min	CPU / GPU
Saliency-Guided NF	∼0.8M	128	0.9 s	27 min	GPU
Kernelized NF	∼1.1M	128	1.2 s	36 min	GPU
Dynamic NF	∼1.3M	128	1.4 s	42 min	GPU
Unified NF Framework	∼1.5M	128	1.6 s	48 min	GPU

the training objective. To assess the robustness of the model with respect to these parameters, a sensitivity analysis was conducted by varying each coefficient while keeping all other training settings fixed. The analysis focuses on the trade-off between fairness improvement and predictive performance, as excessive regularization may reduce accuracy while insufficient regularization may limit fairness gains. Across the evaluated ranges, the model exhibits stable behaviour, with gradual changes in fairness metrics and no abrupt degradation in accuracy. This indicates that the framework is not overly sensitive to small variations in the regularization weights.

8.3.2 Computational Complexity and Training Cost

To complement the fairness and predictive performance evaluation, we report the computational cost of the proposed methods and compare them with standard neural network baselines. This analysis focuses on model size, training time, and hardware requirements, with the aim of assessing the practical overhead introduced by the neuro-fuzzy components. All models were trained using the same optimization settings and batch sizes to ensure a fair comparison. Training time is reported both per epoch and for the full training procedure, providing insight into the scalability of the proposed framework. While the neuro-fuzzy models introduce additional parameters related to fuzzy rules and membership functions, the overall computational cost remains within a practical range for offline training. Importantly, the increase in training time is modest when compared to baseline NNs, particularly in view of the gains in interpretability and fairness control achieved by the proposed methods. Training times are reported as representative averages measured under identical experimental settings and are intended to provide a comparative indication of computational overhead rather than absolute performance benchmarks Table 8.2.

8.4 Results

8.4.1 Synthetic data

In the simulated datasets, all three fairness strategies surpassed the baseline in diminishing sensitive-feature saliency. The dynamic fuzzy controller showed the best balance between fairness and predictive performance. It kept accuracy close to the baseline model while noticeably reducing the influence of sensitive features. The kernelized ANFIS variant demonstrated more stable behaviour across folds, especially when sensitive and relevant attributes were strongly correlated in nonlinear ways.

Table 8.3: Fuzzy rules used to encode fairness constraints and the corresponding pre-processing bias indices (DI, SPD).

Dataset	**Representative fuzzy rule**	**DI**	**SPD**
Adult	IF Gender=Male AND Race=White THEN Bias=High	0.18	0.12
COMPAS	IF Race=AfricanAmerican AND Gender=Male THEN Bias=High	0.21	0.15
Bank	IF Age >45 AND Gender=Female THEN Bias=Medium	0.11	0.09
German Credit	IF Age <25 THEN Bias=High	0.24	0.17

8.4.2 Real-world data

Evaluation on the real dataset showed that the proposed framework reduced attribution to sensitive variables when compared with the baseline model, while fairness measures improved consistently across folds. The dynamic fuzzy configuration preserved predictive performance and lowered demographic differences in the outputs. In addition, the fuzzy rules allow inspection of which feature contributions were penalized during training, supporting transparency and practical interpretability.

DI values closer to one indicate balanced outcomes between protected and unprotected groups, while values far from one suggest potential bias. SPD measures the difference in positive outcome rates and is closer to zero when group outcomes are similar. EOD and AOD assess differences in error rates across groups and are also closer to zero when fairness improves. These reference ranges are used to interpret the results reported below.

Table 8.5 summarizes the hyperparameter settings used across all experiments to ensure reproducibility and consistency.

Table 8.4: Fairness and performance metrics (real-world dataset)

Model	Accuracy (%)	F1-score	DPD	EOD
Baseline	84.1	0.81	0.18	0.22
Saliency penalty	82.7	0.79	0.12	0.15
Kernelized ANFIS	83.4	0.80	0.10	0.13
Dynamic fuzzy	83.9	0.80	0.08	0.11

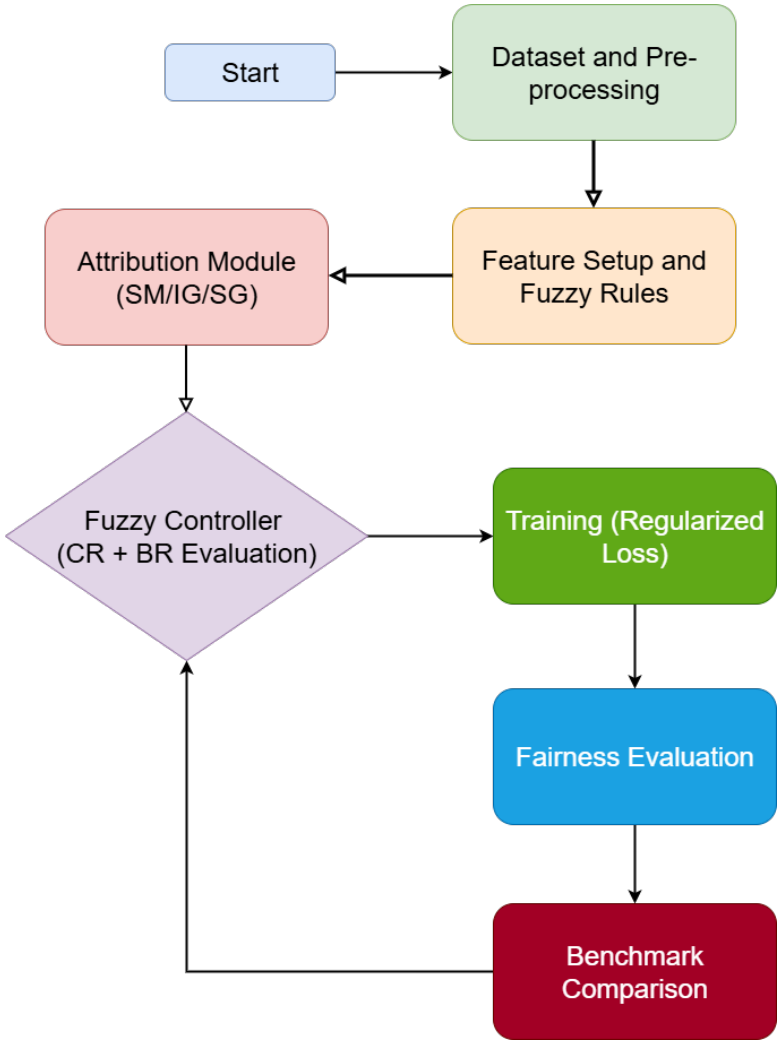

Figure 8.2: Workflow of the framework.

Table 8.5: Summary of hyperparameters used in the experimental evaluation.

Hyperparameter	Value
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	128
Training epochs	30
Fairness regularization (λ_f)	0.5
Rule regularization (λ_r)	0.5
Kernel parameters	RBF kernel, σ selected via validation
Membership function type	Gaussian
Dynamic rule update frequency	Every 5 epochs
Random seeds	5

Table 8.6: Pre-processing fairness comparison across datasets.

Dataset	**DI**	**SPD**
Simulated	0.466667	-0.228571
Adult	0.391641	-0.201023
German	0.868946	-0.096679
COMPAS	0.884041	-0.068521

8.5 Discussion

The results show that the proposed framework works well both in controlled tests and on real datasets. By combining attribution monitoring with static and adaptive fuzzy components, the method supports different levels of fairness control depending on the needs of the application. This adaptability allows the framework to be deployed across a wide range of practical scenarios.

Moreover, incorporating fuzzy rules enhances interpretability by providing explicit conditions linked to sensitive variables, rather than modifying parameters in a black-box manner. Stakeholders can examine the decision boundaries associated with fairness-related features, which facilitates regulatory review and accountability. Because the approach builds on gradient-based saliency, it remains compatible with various common NN architectures. Since the approach relies on gradient-based saliency, it can be applied alongside many common NN structures. Tables 8.3–8.14

Table 8.7: Post-processing (Simulated) — key attributes.

Model	ACC	DI	SPD	EOD	AOD	Bias	$\hat{\Psi}_{\text{last}}$
M0	0.545	0.238	-0.762	0.885	-0.808	–	0.491
M1	0.660	0.911	-0.033	0.392	-0.086	–	0.000
M2	0.820	0.897	-0.044	0.231	0.029	0.018	0.441
M3	0.835	0.808	-0.082	0.269	-0.010	–	0.462
M4	0.820	0.933	-0.029	0.192	0.053	0.041	0.445
M5	0.820	0.897	-0.044	0.231	0.029	0.016	0.490
M6	0.835	0.808	-0.082	0.269	-0.010	–	0.485
M7	0.820	0.897	-0.044	0.231	0.029	0.036	0.482
M8	0.720	0.744	-0.154	0.231	-0.083	0.023	0.450
M9	0.775	0.793	-0.107	0.231	-0.032	–	0.467
M10	0.785	0.763	-0.122	0.231	-0.041	0.051	0.443
M_FL_DP	0.720	–	0.000	0.000	0.000	–	–
M_FL_EO	0.720	–	0.000	0.000	0.000	–	–
M_PR	0.895	0.305	-0.298	0.577	-0.260	–	–

Table 8.8: Post-processing (Adult) — key attributes.

Model	ACC	DI	SPD	EOD	AOD	Bias	$\hat{\Psi}_{\text{last}}$
M0	0.955	0.430	-0.218	0.000	-0.019	–	0.500
M1	0.805	0.641	-0.059	0.031	-0.008	–	0.000
M2	0.965	0.474	-0.183	-0.026	0.014	0.059	0.490
M3	0.970	0.462	-0.192	0.000	0.001	–	0.497
M4	0.970	0.462	-0.192	0.000	0.001	0.033	0.497
M5	0.965	0.474	-0.183	-0.026	0.014	0.068	0.468
M6	0.970	0.462	-0.192	0.000	0.001	–	0.468
M7	0.970	0.462	-0.192	0.000	0.001	0.046	0.468
M8	0.965	0.474	-0.183	-0.026	0.014	0.049	0.497
M9	0.920	0.523	-0.183	-0.053	0.015	–	0.497
M10	0.970	0.462	-0.192	0.000	0.001	0.031	0.500
M_FL_DP	1.000	0.392	-0.201	0.000	0.000	–	–
M_FL_EO	1.000	0.392	-0.201	0.000	0.000	–	–
M_PR	1.000	0.392	-0.201	0.000	0.000	–	–

Table 8.9: Post-processing (German) — key attributes.

Model	ACC	DI	SPD	EOD	AOD	Bias	$\hat{\Psi}_{\text{last}}$
M0	0.690	0.992	-0.006	0.096	0.077	–	0.500
M1	0.815	0.900	-0.087	0.038	-0.050	–	0.000
M2	1.000	0.869	-0.097	0.000	0.000	0.021	0.500
M3	1.000	0.869	-0.097	0.000	0.000	0.020	0.500
M4	1.000	0.869	-0.097	0.000	0.000	0.022	0.500
M5	1.000	0.869	-0.097	0.000	0.000	0.030	0.440
M6	1.000	0.869	-0.097	0.000	0.000	0.030	0.440
M7	1.000	0.869	-0.097	0.000	0.000	0.037	0.440
M8	1.000	0.869	-0.097	0.000	0.000	0.019	0.500
M9	1.000	0.869	-0.097	0.000	0.000	0.000	0.500
M10	1.000	0.869	-0.097	0.000	0.000	0.022	0.500
M_FL_DP	1.000	0.869	-0.097	0.000	0.000	–	–
M_FL_EO	1.000	0.869	-0.097	0.000	0.000	–	–
M_PR	1.000	0.869	-0.097	0.000	0.000	–	–

Table 8.10: Post-processing (COMPAS) — key attributes.

Model	ACC	DI	SPD	EOD	AOD	Bias	$\hat{\Psi}_{\text{last}}$
M0	0.945	0.797	-0.129	0.071	-0.068	–	0.000
M1	0.565	0.970	-0.030	0.000	-0.031	–	0.000
M2	0.920	0.958	-0.023	-0.005	0.039	0.443	0.000
M3	0.950	0.884	-0.069	0.032	-0.003	–	0.000
M4	0.950	0.884	-0.069	0.032	-0.003	0.158	0.000
M5	0.920	0.958	-0.023	-0.005	0.039	0.443	0.000
M6	0.950	0.884	-0.069	0.032	-0.003	0.158	0.000
M7	0.950	0.884	-0.069	0.032	-0.003	0.000	0.000
M8	0.945	0.881	-0.068	0.020	-0.005	0.126	0.000
M9	0.950	0.884	-0.069	0.032	-0.003	0.000	0.000
M10	0.950	0.884	-0.069	0.032	-0.003	0.128	0.000
M_FL_DP	1.000	0.884	-0.069	0.000	0.000	–	–
M_FL_EO	1.000	0.884	-0.069	0.000	0.000	–	–
M_PR	1.000	0.884	-0.069	0.000	0.000	–	–

Table 8.11: Scoreboard (Simulated).

Model	F	$F_\beta(\beta = 1)$	ACC	ΔF vs M0
M_FL_DP	1.000	0.837	0.720	0.940
M_FL_EO	1.000	0.837	0.720	0.940
M4	0.846	0.833	0.820	0.787
M7	0.823	0.821	0.820	0.763
M5	0.823	0.821	0.820	0.763
M3	0.771	0.802	0.835	0.712
M6	0.771	0.802	0.835	0.712
M10	0.744	0.764	0.785	0.684
M1	0.722	0.690	0.660	0.663
M8	0.702	0.711	0.720	0.643
M_PR	0.298	0.447	0.895	0.238
M0	0.060	0.107	0.545	0.000

Table 8.12: Scoreboard (Adult).

Model	F	$F_\beta(\beta = 1)$	ACC	ΔF vs M0
M1	0.861	0.832	0.805	0.122
M3	0.769	0.858	0.970	0.030
M4	0.769	0.858	0.970	0.030
M10	0.769	0.858	0.970	0.030
M7	0.769	0.858	0.970	0.030
M6	0.769	0.858	0.970	0.030
M2	0.757	0.848	0.965	0.017
M5	0.757	0.848	0.965	0.017
M8	0.757	0.848	0.965	0.017
M9	0.756	0.830	0.920	0.016
M_FL_EO	0.747	0.855	1.000	0.008
M_FL_DP	0.747	0.855	1.000	0.008
M_PR	0.747	0.855	1.000	0.008
M0	0.739	0.833	0.955	0.000

Table 8.13: Scoreboard (German).

Model	F	$F_\beta(\beta = 1)$	ACC	ΔF vs M0
M3	0.919	0.958	1.000	0.010
M2	0.919	0.958	1.000	0.010
M4	0.919	0.958	1.000	0.010
M5	0.919	0.958	1.000	0.010
M7	0.919	0.958	1.000	0.010
M6	0.919	0.958	1.000	0.010
M10	0.919	0.958	1.000	0.010
M_FL_DP	0.919	0.958	1.000	0.010
M8	0.919	0.958	1.000	0.010
M9	0.919	0.958	1.000	0.010
M_FL_EO	0.919	0.958	1.000	0.010
M_PR	0.919	0.958	1.000	0.010
M0	0.909	0.784	0.690	0.000
M1	0.888	0.850	0.815	-0.021

Table 8.14: Scoreboard (COMPAS).

Model	F	$F_\beta(\beta = 1)$	ACC	ΔF vs M0
M1	0.962	0.712	0.565	0.147
M2	0.956	0.938	0.920	0.140
M5	0.956	0.938	0.920	0.140
M_FL_DP	0.937	0.967	1.000	0.121
M_PR	0.937	0.967	1.000	0.121
M_FL_EO	0.937	0.967	1.000	0.121
M8	0.924	0.934	0.945	0.108
M9	0.919	0.934	0.950	0.104
M3	0.919	0.934	0.950	0.104
M4	0.919	0.934	0.950	0.104
M6	0.919	0.934	0.950	0.104
M7	0.919	0.934	0.950	0.104
M10	0.919	0.934	0.950	0.104
M0	0.815	0.875	0.945	0.000

compile the rule constraints, fairness metrics, and comparative results across preprocessing and post-processing evaluations on all datasets.

8.6 Conclusion

This chapter combined the earlier developments into a single training framework that handles fairness within the model optimization process. The unified framework was evaluated using both controlled synthetic experiments and a real-world dataset. The results demonstrated a consistent decrease in reliance on sensitive features, improvements in fairness-related metrics, and preservation of interpretable rule-based decision behavior. This represents the point in the thesis where earlier conceptual developments translate into a practically deployable solution. The next chapter reflects on the overall findings and identifies future research directions.

Chapter 9

Conclusion and Future Work

9.1 Summary of Contributions

The main aim of this thesis was to investigate and validate techniques that reduce biased predictive behaviour in NNs. This was approached by integrating attribution-derived information with fuzzy rule mechanisms and by designing adaptive strategies that react to evolving patterns of unfairness throughout training. The research advanced in four main stages, each introducing a dedicated methodological extension.

- **Stage 1 – Feasibility Study (Chapter 5):** This stage assessed whether saliency information can be leveraged to discourage reliance on sensitive attributes during learning. A fuzzy penalty was used to translate high attribution on protected features into corrective adjustments. Experimental findings showed that this method helped mitigate unfair dependence without noticeably degrading prediction performance.

- **Stage 2 – ANFIS Enhancement (Chapter 6):** The method was extended by embedding fairness constraints directly within the ANFIS structure. Kernel-based membership functions enabled representation of more complex decision boundaries that standard rules could not capture. This extension improved behaviour when input variables displayed intricate correlations.

- **Stage 3 – Dynamic Fuzzy Controller (Chapter 7):** An adaptive learning component was introduced that updates both rule weights and membership functions during model training. This mechanism allows fairness adjustments to evolve together with the model's behaviour, aligning penalties with the saliency patterns observed at each iteration. This adaptive approach delivered

consistent improvements and maintained transparency in how fairness penalties were applied throughout training.

- **Stage 4 - Unified and Real-World Evaluation (Chapter 8):** The fourth stage integrated the previous advances into a single learning framework. Evaluation extended beyond synthetic scenarios to widely used real datasets, demonstrating that the proposed approach can limit bias linked to protected variables while still preserving model explainability. These outcomes support the suitability of the framework for fairness-critical applications.

Together, these contributions illustrate a step-by-step evolution: examining feasibility, improving representational capability, enabling adaptive behaviour, and validating performance under practical conditions. The results indicate that fairness constraints can be embedded into the learning process itself through interpretable fuzzy mechanisms, providing an alternative to post-hoc corrections and helping ensure responsible model behaviour.

9.2 Consequences

The findings of this thesis carry implications for both research and real-world deployment. From a theoretical perspective, linking fuzzy reasoning with attribution information challenges the traditional separation between interpretability research and fairness research. The experiments confirm that attribution signals can not only diagnose unfair behaviour but can also shape the learning process itself, embedding fairness from the outset rather than as a corrective measure applied after deployment.

From an application viewpoint, the techniques developed here provide a more transparent alternative to fairness adjustments that function as opaque add-on components. Because the penalty mechanisms remain interpretable through their rule-based structure, stakeholders can examine how and why sensitive features are constrained. This transparency is especially valuable in regulated settings, including hiring, financial risk assessment, and medical decision support, where accountability and justification are essential alongside predictive accuracy.

9.3 Limitations

Although the results are promising, certain limitations must be acknowledged. Attribution-based indicators are not always robust; for example, gradients can fluctuate sharply with small input variations, which may reduce their reliability when used

to guide fairness constraints. Moreover, the neuro-fuzzy structure inherits scalability challenges typical of rule-based systems, since the rule count may expand rapidly when several sensitive features are introduced simultaneously.

9.4 Discussion and Challenges

This thesis has demonstrated that integrating interpretability and fairness within neuro-fuzzy learning frameworks is both feasible and effective. At the same time, several limitations and practical challenges remain. First, while the proposed methods improve transparency and fairness control, we introduced additional model components that increase training time and parameter tuning complexity. Although this overhead remains manageable in offline learning settings, it may limit applicability in scenarios with strict real-time constraints.

A further challenge concerns the selection of regularization parameters and stability thresholds. While sensitivity analyses indicate that the framework is robust to moderate variations, parameter selection still relies on validation procedures that may need adjustment across datasets or application domains. In addition, the dynamic rule-updating mechanism assumes access to reliable attribution signals; inaccuracies in attribution methods may affect the effectiveness of fairness regulation.

From a practical perspective, deployment in real-world systems requires careful consideration of data shifts, evolving definitions of sensitive attributes, and regulatory constraints. Fairness objectives are often context-dependent, and the choice of metrics and constraints should reflect domain-specific requirements rather than universal criteria. The computational cost and interpretability benefits must therefore be evaluated alongside policy, ethical, and operational considerations.

Overall, these challenges highlight directions for further refinement, including automated parameter selection, improved attribution robustness, and tighter integration with domain-specific fairness standards. Addressing these aspects will be essential for translating the proposed methods into deployed decision-making systems.

9.5 Future Work

The contributions of this thesis suggest several opportunities for continued investigation:

- **Enhancing attribution stability:** Investigating alternative attribution methods, including IGs, SHAP, or ensembles of explanation techniques, could lead to more consistent fairness feedback signals for the neuro-fuzzy model.

- **Managing rule base growth:** Techniques such as rule pruning, layered fuzzy structures, or neural approximations of rule behaviour may help limit the expansion of rules when several sensitive attributes are monitored, while preserving the clarity of the resulting explanations.
- **Testing in new domains:** Extending the framework to tasks involving multimodal inputs (e.g., language, visual data) would help assess adaptation to varied application demands and uncover domain-specific fairness considerations.
- **User-driven fairness adjustment:** Introducing interfaces where practitioners or stakeholders can modify rule weights or fairness penalties during operation may increase transparency and enable more participatory decision support.
- **Supporting regulatory compliance:** Aligning rule-based penalties with formal fairness criteria from legislation or policy, such as DP or equal opportunity, may promote adoption in areas requiring clear traceability and justification of decisions.

Beyond these directions, future work may explore the incorporation of causal notions of fairness, which aim to separate genuine causal effects from spurious correlations in the data. Such approaches could complement the current framework by providing stronger guarantees when bias arises from complex data-generating processes. In addition, extending the proposed methods to modern neural architectures, such as transformer-based models, represents an important step toward scalability in high-dimensional and sequential settings.

9.6 Closing Remarks

This thesis has developed a set of approaches aimed at reducing biased behaviour in NNs while preserving clarity in how decisions are formed. The techniques integrate attribution-based insights with fuzzy rule reasoning to guide the learning process toward fairer outcomes.

Bibliography

[1] R. T. Rabonato and L. Berton, "A systematic review of fairness in machine learning," *AI and Ethics*, vol. 5, no. 3, pp. 1943–1954, 2025.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[3] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, 2024.

[4] H. Van Kolfschooten and J. Van Oirschot, "The eu artificial intelligence act (2024): implications for healthcare," *Health Policy*, vol. 149, p. 105152, 2024.

[5] N. AI, "Artificial intelligence risk management framework (ai rmf 1.0)," *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pp. 100–1, 2023.

[6] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 3, pp. 1–27, 2023.

[7] L. Deck, J. Schoeffer, M. De-Arteaga, and N. Kühl, "A critical survey on fairness benefits of explainable ai," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1579–1595.

[8] X. Gu, J. Han, Q. Shen, and P. P. Angelov, "Autonomous learning for fuzzy systems: a review," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 7549–7595, 2023.

[9] R. Riegel, A. Gray, F. Luus, N. Khan, N. Makondo, I. Y. Akhalwaya, H. Qian, R. Fagin, F. Barahona, U. Sharma *et al.*, "Logical neural networks," *arXiv preprint arXiv:2006.13155*, 2020.

[10] A. Gevaert, A.-J. Rousseau, T. Becker, D. Valkenborg, T. De Bie, and Y. Saeys, "Evaluating feature attribution methods in the image domain," *Machine Learning*, vol. 113, no. 9, pp. 6019–6064, 2024.

[11] S. Shah, D. E. Ciucci, S. L. Manzoni, I. F. Zoppis *et al.*, "Neural networks bias mitigation through fuzzy logic and saliency maps," *ICAART*, vol. 3, pp. 1343–1351, 2025.

[12] I. Zoppis, S. Shah, S. Manzoni, and D. Ciucci, "Kernelizing adaptive neuro-fuzzy inference for bias mitigation," in *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*. IEEE, 2025, pp. 1–6.

[13] S. Shah, D. E. Ciucci, S. L. Manzoni, and I. F. Zoppis, "Dynamic neuro-fuzzy regularization for fairness-aware neural networks using saliency attribution," in *Proceedings of the 2025 Artificial Intelligence Models and Systems Symposium (AIMS 2025), co-located with the 3rd International Conference on Foundation and Large Language Models (FLLM 2025)*, Vienna, Austria, November 2025, accepted for publication.

[14] I. Zoppis, S. Shah, S. Manzoni, G. Lazzarinetti, D. Malchiodi, and D. Ciucci, "In-processing neuro fuzzy approaches for bias mitigation," *Journal of Artificial Intelligence Research*, 2025, manuscript submitted for publication.

[15] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annual review of statistics and its application*, vol. 8, no. 1, pp. 141–163, 2021.

[16] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[17] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 329–338.

[18] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International journal of man-machine studies*, vol. 7, no. 1, pp. 1–13, 1975.

[19] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, and M. J. H. Faruk, "Survey on machine learning biases and mitigation techniques," *Digital*, vol. 4, no. 1, pp. 1–68, 2023.

[20] Q. Feng, M. Du, N. Zou, and X. Hu, "Fair machine learning in healthcare: A survey," *IEEE Transactions on Artificial Intelligence*, 2024.

[21] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[22] A. Krishna Menon and R. C. Williamson, "The cost of fairness in classification," *arXiv e-prints*, pp. arXiv–1705, 2017.

[23] I. Žliobaitė, "Measuring discrimination in algorithmic decision making," *Data Mining and Knowledge Discovery*, vol. 31, no. 4, pp. 1060–1089, 2017.

[24] D. Saxena and S. Guha, "Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making," *ACM Journal on Responsible Computing*, vol. 1, no. 1, pp. 1–32, 2024.

[25] Z. Tang and K. Zhang, "Attainability and optimality: The equalized odds fairness revisited," in *Conference on Causal Learning and Reasoning*. PMLR, 2022, pp. 754–786.

[26] P. Benz, C. Zhang, S. Ham, A. Karjauv, G. Cho, and I. S. Kweon, "Trade-off between accuracy, robustness, and fairness of deep classifiers," in *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges*, 2021.

[27] J. N. Yan, J. Wang, J. M. Rzeszotarski, and A. Koenecke, "Fairness practices in industry: A case study in machine learning teams building recommender systems," *arXiv preprint arXiv:2505.19441*, 2025.

[28] M. Kearns and A. Roth, "Ethical algorithm design," *ACM SIGecom Exchanges*, vol. 18, no. 1, pp. 31–36, 2020.

[29] A. E. Adeoye and C. A. Obaze, "Explainable artificial intelligence (xai): A comprehensive review of methods, applications, and open issues," *Tech-Sphere Journal for Pure and Applied Sciences*, vol. 2, no. 1, pp. 25–41, 2025.

[30] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[31] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistic Surveys*, vol. 16, pp. 1–85, 2022.

[32] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[33] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[34] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[35] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[37] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 59–68.

[38] Y. Zhao, Y. Wang, and T. Derr, "Fairness and explainability: Bridging the gap towards fair model explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 363–11 371.

[39] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[40] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 648–657.

[41] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[42] T. J. Ross, *Fuzzy logic with engineering applications*. John Wiley & Sons, 2005.

[43] K. Tanaka and B. Werners, *An introduction to fuzzy logic for practical applications*. Springer, 1997, vol. 1.

[44] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [book review]," *IEEE Transactions on automatic control*, vol. 42, no. 10, pp. 1482–1484, 1997.

[45] J.-S. Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE transactions on systems, man, and cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[46] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[47] W. Pedrycz and F. Gomide, *Fuzzy systems engineering: toward human-centric computing*. John Wiley & Sons, 2007.

[48] N. Talpur, S. J. Abdulkadir, H. Alhussian, ·. M. H. Hasan, N. Aziz, and A. Bamhdi, "A comprehensive review of deep neuro-fuzzy system architectures and their optimization methods," *Neural Computing and Applications*, vol. 34, no. 3, pp. 1837–1875, 2022.

[49] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[50] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[51] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *arXiv preprint arXiv:1901.10002*, vol. 2, no. 8, p. 73, 2019.

[52] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[53] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[54] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.

[55] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[56] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[57] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[58] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in neural information processing systems*, vol. 30, 2017.

[59] S. Hooker, "Moving beyond "algorithmic bias is a data problem"," *Patterns*, vol. 2, no. 4, 2021.

[60] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[61] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[62] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of internal medicine*, vol. 169, no. 12, pp. 866–872, 2018.

[63] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[64] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 149–159.

[65] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[66] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.

[67] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[68] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3384–3393.

[69] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *2018 IEEE international conference on big data (big data)*. IEEE, 2018, pp. 570–575.

[70] T. Adel, I. Valera, Z. Ghahramani, and A. Weller, "One-network adversarial fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2412–2420.

[71] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[72] A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan, "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals," *Journal of Machine Learning Research*, vol. 20, no. 172, pp. 1–59, 2019.

[73] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, accountability and transparency*. PMLR, 2018, pp. 107–118.

[74] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[75] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[76] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 2016, pp. 144–152.

[77] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[78] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[79] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[80] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[81] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.

[82] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[83] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," *arXiv preprint arXiv:1703.03717*, 2017.

[84] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness gan: Generating datasets with fairness properties using a generative adversarial network," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 3–1, 2019.

[85] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un) reliability of saliency methods," in *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer, 2019, pp. 267–280.

[86] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[87] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," *arXiv preprint arXiv:1706.07206*, 2017.

[88] L. A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE Transactions on systems, Man, and Cybernetics*, no. 1, pp. 28–44, 1973.

[89] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE transactions on systems, man, and cybernetics*, no. 1, pp. 116–132, 1985.

[90] C.-C. Lee, "Fuzzy logic in control systems: fuzzy logic controller. i," *IEEE Transactions on systems, man, and cybernetics*, vol. 20, no. 2, pp. 404–418, 1990.

[91] D. J. Dubois, *Fuzzy sets and systems: theory and applications*. Academic press, 1980, vol. 144.

[92] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[93] C.-T. Lin and C. S. G. Lee, "Neural-network-based fuzzy logic control and decision system," *IEEE Transactions on computers*, vol. 40, no. 12, pp. 1320–1336, 1991.

[94] G. Raju, J. Zhou, and R. A. Kisner, "Hierarchical fuzzy control," *International journal of control*, vol. 54, no. 5, pp. 1201–1216, 1991.

[95] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," *Advances in neural information processing systems*, vol. 30, 2017.

[96] C. Lawless, S. Dash, O. Gunluk, and D. Wei, "Interpretable and fair boolean rule sets via column generation," *Journal of Machine Learning Research*, vol. 24, no. 229, pp. 1–50, 2023.

[97] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the international workshop on software fairness*, 2018, pp. 1–7.

[98] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, "Can explainable ai explain unfairness? a framework for evaluating explainable ai," *arXiv preprint arXiv:2106.07483*, 2021.

[99] Q. Hu, D. Yu, W. Pedrycz, and D. Chen, "Kernelized fuzzy rough sets and their applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 11, pp. 1649–1667, 2010.

La borsa di dottorato è cofinanziata con risorse dell'Unione europea

NextGenerationEU

Piano Nazionale di Ripresa e Resilienza
Missione 4 – Componente 1 – Riforma 4.1
Riforma dei Dottorati – Inv. 4.1
Borse PNRR patrimonio Culturale

CUP: **H41J22000250009**