

Uncertainty Quantification with Machine Learning and Reinforcement Learning for Business Decision-Making

A Dissertation submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics and Management

by

Akash SINGH

Thesis Jury Members:

- Prof. Michael SCHYNS President of the Jury (HEC Liège, Management School of the University of Liège)
- Prof. Ashwin ITTOO, Supervisor (HEC Liège, Management School of the University of Liège)
- Prof. Elise VANDOMME, Supervisor (HEC Liège, Management School of the University of Liège)
- Prof. Julien HAMBUCKERS (HEC Liège, Management School of the University of Liège)
- Prof. Kristof COUSSEMENT (IÉSEG School of Management)
- Dr. Pierre ARS (Qualified Actuary)
- Prof. Siamak KHAYYATI (HEC Liège, Management School of the University of Liège)

Declaration of Authorship

I, Akash SINGH, declare that this thesis titled, “Uncertainty Quantification with Machine Learning and Reinforcement Learning for Business Decision-Making” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- No part of this thesis has previously been submitted for a degree or any other qualification at this University or at any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“All models are wrong, but some are useful”

George E. P. Box

UNIVERSITY OF LIEGE

Abstract

HEC

Doctor of Philosophy

**Uncertainty Quantification with Machine Learning and Reinforcement Learning
for Business Decision-Making**

by Akash SINGH

In recent years, the deployment of artificial intelligence (AI) in high-stakes domains such as law, medicine, finance, and customer profiling has intensified the need for *trustworthy AI*; systems whose outputs remain safe, reliable, and actionable under real-world constraints. In such settings, predictive performance on historical data is not sufficient: models can still produce *incorrect* predictions with high apparent confidence, particularly under label noise, limited evidence, severe class imbalance, and distribution shift between training and deployment data. This miscalibration and overconfidence are operationally costly: they can trigger unwarranted automated actions, hide failure modes that emerge out-of-distribution, and create a false sense of high performance. Decision-makers therefore need signals of when a prediction is trustworthy and when it should be treated cautiously, deferred, or supported by additional information. This thesis therefore focuses on *uncertainty quantification* (UQ) as a practical mechanism for trustworthy decision-making by making explicit when model outputs are reliable enough and when caution or human oversight is warranted.

The thesis investigates uncertainty across both supervised machine learning and reinforcement learning. In supervised learning, we study uncertainty estimation in operational prediction problems where data are noisy and imbalanced. We examine distribution-free and evidence-theoretic approaches, including conformal prediction and belief functions, and identify conditions under which they lose actionability (e.g., inflated prediction sets under weak separability or high residual ignorance under conflicting evidence). Motivated by these limitations, we develop an ensemble-based approach tailored to real-world decision support: a heterogeneous ensemble framework that combines diverse learning paradigms to improve calibration and to operationalize epistemic uncertainty as disagreement across heterogeneous hypotheses, while separately characterizing aleatoric uncertainty arising from irreducible data noise. We further analyze how feature transformations and model composition affect both predictive performance and the stability of uncertainty estimates.

Beyond supervised prediction, the thesis extends uncertainty-aware learning to sequential decision-making. We propose a lightweight uncertainty-aware reinforcement learning framework that leverages masksembles to obtain epistemic uncertainty estimates without requiring large ensembles. By incorporating uncertainty into temporal-difference learning, the resulting agents improve training stability and robustness under extreme observation noise and perturbations, addressing a key gap in uncertainty-aware reinforcement learning where uncertainty is often used primarily for exploration rather than for stabilizing value-function learning.

From a managerial perspective, the thesis demonstrates how calibrated uncertainty enables decision policies that are more dependable than score-based automation alone. Uncertainty quantification supports selective prediction and human-in-loop workflows, risk-aware thresholding, and resource allocation strategies such as sub-portfolio selection, where limited operational capacity requires prioritizing actions on cases that are both high-impact and sufficiently reliable. Overall, the thesis provides a cross-paradigm foundation for uncertainty-aware learning systems that translate predictive uncertainty into safer, more robust, and operationally actionable decision support.

Acknowledgements

I would first like to express my sincere gratitude to my supervisors, Professor Ashwin Ittoo and Professor Élise Vandomme. Throughout this journey, they were guiding lights. Their support, feedback, and steady guidance shaped both my work and my growth as a researcher.

I would also like to thank my dear friend Rishabh Shukla, whom I consider a brother. He helped me through countless late nights on calls, supported me with technical advice, and kept my spirits up when the work felt heavy.

I am deeply thankful to my parents, Dr. A. P. Singh and Mrs. Madhu Singh, for their constant support over the years. My mother's prayers gave me strength and calm when I needed it most. My father's stories about his own doctorate, and his reminder that perseverance pays in the end, kept me focused and resilient.

I would like to thank my brother, Ambar, who motivated me during times when I felt alone. He was my biggest cheerleader, and his own hard work and discipline inspired me to keep pursuing my goals. In India, we often say that a brother is your right arm: a pillar of strength, protection, and dependability. Ambar has been exactly that for me.

Finally, I want to thank the love of my life, my beautiful wife Hajnalka Frenyo Singh, who came into my life during my doctorate. When I was sad or overwhelmed, she was the first to hold me and bring me back to hope. She supported me through sleepless nights, cared for me with patience, and advised me with the honesty of a true friend. To me, she is like Goddess Lakshmi, bringing warmth and abundance into our home. Above all, she made this journey meaningful, and I am grateful for her love every day.

Contents

Declaration of Authorship	iii
Abstract	vii
Acknowledgements	ix
List of Abbreviations	xix
1 Introduction	1
2 Uncertainty Quantification with Belief functions and Conformal Predictions	11
2.1 Introduction	12
2.2 Literature study	13
2.2.1 Belief Functions	13
2.2.2 Conformal Predictions	14
2.3 Background	16
2.3.1 Belief Functions	18
2.3.2 Conformal Prediction	22
2.4 Methodology	23
2.4.1 Belief Functions	23
2.4.2 Conformal Prediction	25
2.5 Data	26
2.5.1 Proprietary churn dataset (Version 1)	26
2.5.2 Proprietary churn dataset (Version 2)	28
2.5.3 Open credit card fraud dataset	29
2.6 Results	30
2.6.1 Belief Functions	30
2.6.2 Conformal Prediction	31
Proprietary churn dataset (Version 2)	31
Open credit card fraud dataset	32
2.7 Discussion	33
3 Heterogeneous Ensemble framework for Uncertainty Quantification	35
3.1 Introduction	36
3.2 Literature review	39

3.2.1	Uncertainty Estimation in Machine Learning	39
3.2.2	Uncertainty in Machine Learning for Operations Research . . .	41
3.2.3	Use case: Churn prediction	41
3.2.4	Use case: Customer conversion	42
3.2.5	Literature review findings	43
3.3	Methodology	43
3.3.1	Uncertainty in machine learning	43
3.3.2	Uncertainty in HEUQ: a <i>Basic Composition Rule (BCR)</i> approach	45
	Total uncertainty for a given model/hypothesis	46
	Total uncertainty for an ensemble model	46
	Decomposition of total uncertainty for an ensemble of hypothe-	
	ses	47
3.3.3	Heterogeneous Ensemble for Uncertainty Prediction: HEUQ . .	49
3.4	Data	50
3.4.1	Proprietary churn dataset (Version 2)	52
	Data challenges and mitigation strategy	52
	Data transformation for churn experiments	53
3.4.2	Proprietary Customer conversion dataset description	54
3.5	Experimental Setup	55
3.5.1	Performance measure	56
3.5.2	Training and implementation details	57
3.6	Results	59
3.6.1	Experiments: Churn Prediction	59
3.6.2	Experiments: Uncertainty quantification churn prediction . . .	61
3.6.3	Ablation study	62
	Heterogeneous strong learners	62
3.6.4	Customer conversion (Case study)	65
	Web Dataset Performance	65
	Physical Dataset Performance	65
3.7	Discussion and Conclusion	66
4	Uncertainty Aware Reinforcement Learning	69
4.1	Introduction	69
4.2	Literature study	72
4.3	Methodology	73
4.3.1	Masksembles	73
4.3.2	Q-Learning and Contextual Bandits	74
4.3.3	MASURE: Masksembles for Stable and Uncertainty-aware RL	
	Environments	75
4.3.4	Experimental Setup	76
4.4	Results	79
4.5	Conclusion	82

5 Conclusion & Future works	85
5.1 Introduction	85
5.2 Answers to the Research Questions	86
5.3 Human-in-the-loop workflow for uncertainty-aware decision support	87
5.3.1 Sub-portfolio optimization: Churn management	88
5.4 Synthesis of Key Empirical Insights	92
5.5 Limitations	93
5.6 Future Work	93
5.6.1 From prediction to “predict-then-optimize” with uncertainty	94
A Basic vs convex combination rule in HEUQ	97
A.1 Convex Combination Rule	97
Adaptation of Uncertainties	97
A.1.1 Comparison of BCR results with some specific $CCR(\alpha)$	98
A.2 HEUQ and Data transformations	99
A.3 Churn literature study	100
Bibliography	103
Index	119

List of Figures

2.1	Data distribution of Iris features values by species.	17
2.2	Histogram of petal length features by species. It is noticeable how Versicolor and Virginica overlap.	19
2.3	Illustration of petal-length overlap between Versicolor and Virginica. .	21
2.4	Feature Distribution for Classes Churn/Not Churn. The complete overlap across all features indicates a severe lack of separability between the classes. This absence of clear, distinctive evidence prevents the belief function (mass function) from assigning mass with certainty, leading to high ambiguity and an inability to accurately classify data points.	27
3.1	An example of the binary classification task. Three different optimal hypotheses are represented as three separating lines that divide the data into two classifying regions. Although these hypotheses are learned differently and correspond to distinct points in hypothesis space, they can yield similar class distributions and comparable classification performance.	44
3.2	HEUQ architecture, heterogeneous ensemble approach offers the advantage of combining various independent models/hypotheses. This also allows the flexibility in the rule of combination, such as convex or basic combination rule.	51
3.3	Balanced accuracy (BA) across uncertainty quartiles on the test set. BA decreases from 82.53% in the lowest-uncertainty quartile (Q1) to 50.70% in the highest-uncertainty quartile (Q4), confirming the negative relation between uncertainty and performance.	64
4.1	The CartPole environment used in our experiments. The agent must balance a pole atop a moving cart by applying left or right forces. . . .	77
4.2	The MountainCar environment used in our experiments. An agent must learn to build momentum in an underpowered car by driving back and forth to reach the hilltop goal.	77
4.3	The LunarLander-v2 environment, where the agent learns to control a lander using thrusters to touch down safely on a designated pad. . .	78
4.4	Comparison of SOTA IVDQN and MASURE in noisy environments, showing improved stability of MASURE.	80

4.5	Comparison of performance of baseline agents and MASURE in noisy environments, showing improved stability of MASURE.	81
5.1	Human-in-the-loop workflow using uncertainty estimates (illustrated with Heterogeneous Ensemble for Uncertainty Quantification (HEUQ)).	88
5.2	Precision score, in function of classes defined from crossing of quantile bins of the private data.	89

List of Tables

1.1	Overview of datasets and their connection to chapters and methodologies.	9
2.1	Key feature distributions in the proprietary churn dataset version 1 . . .	26
2.2	Key feature distributions in the proprietary dataset version 2	29
2.3	Prediction vs Actual Percentages (Belief Function Classifier) Proprietary dataset version 1	30
2.4	Conformal Prediction results for LR and HEUQ at $\alpha = 0.05$ and $\alpha = 0.02$ (percentages of actual class)	31
2.5	Conformal Prediction results for Logistic Regression on the Credit Card Fraud dataset	32
3.1	Overview and characteristics of Churn Data set	51
3.2	Data snapshot	52
3.3	Churn prediction performance across datasets, models, and data transformations (Balanced Accuracy)	60
3.4	Friedman test ranking comparing all baseline learners for all datasets and different data transformation (p -value:0.0051)	60
3.5	Uncertainty quantification baseline. p -value:0.007 for Friedman test ranking	62
3.6	HEUQ ablation with heterogeneous strong learners	63
3.7	Performance metrics of machine learning models on Web and Physical Datasets	65
4.1	Compute efficiency comparison of MASURE and ensemble-based baselines on LunarLander-v2.	74
4.2	Experimental setup, noise schedule, and hyperparameters.	76
4.3	Balanced accuracy (BA) (%) of RL agents for churn prediction (offline learning).	79
4.4	Average ranks of RL agents under the Friedman test (lower is better). . .	81
5.1	Balanced accuracy (BA) computed on sub-portfolios obtained by crossing quantile bins of epistemic uncertainty u_e (8 bins) and total uncertainty u_t (5 bins) on the proprietary churn dataset (Version 2.5.2). Each cell reports the BA for the subset of customers whose (u_e, u_t) values fall in the corresponding bin ranges.	91

5.2	Evolution of the BA score with increasing level of uncertainty (u_t, u_e, u_a)	92
A.1	Performance comparison of different convex combination rule with HEUQ	99
A.2	Performance HEUQ with RG	99
A.3	Existing literature explores ensemble methods in the use case of churn prediction, yet none address uncertainty estimation.	101

List of Abbreviations

AI Artificial Intelligence

BF Belief Functions

BG Bagging of decision trees

BS Brier Score

CB CatBoost

CNN Convolutional Neural Network

CP Conformal Prediction

DST Dempster–Shafer Theory

HEUQ Heterogeneous Ensemble for Uncertainty Quantification

ICP Inductive Conformal Prediction

LR Logistic Regression

MASURE Masksembles for Stable and Uncertainty-Aware Reinforcement Learning
Environments

ML Machine Learning

NLL Negative Log-Likelihood

NN Neural Network

PCA Principal Component Analysis

PLCP Partition Learning Conformal Prediction

RF Random Forest

RG Random Gaussian

RL Reinforcement Learning

UARL Uncertainty aware RL

UE Uncertainty Estimation

UQ Uncertainty Quantification

xx

XAI eXplainable Artificial Intelligence

XGB XGBoost

*To my parents, my brother Ambar, and my wife, Hajnalka
Frenyo Singh, for their unwavering support throughout this
doctoral journey.*

Chapter 1

Introduction

In recent years, the topic of trustworthy Artificial Intelligence (AI) has received increasing attention, driven by the widespread deployment of AI systems in domains such as law, medicine, finance, and customer profiling (Abdar et al., 2021). In these high-stakes settings, model predictions directly influence decisions with significant implications for individuals and organizations. Consequently, evaluating model performance solely on historical accuracy is insufficient. It is equally critical to understand how reliable and robust model outputs remain when confronted with noise, uncertainty, or distributional changes (Abdar et al., 2021; Hoffmann, Fortmeier, and Elster, 2021; Kline, 1985). Addressing these concerns requires moving beyond deterministic predictions and quantifying the degree of confidence associated with model decisions. Trustworthy AI refers to AI systems that can be used safely and reliably in real-life decisions (Chander et al., 2025). In real-world deployment, trustworthiness concerns whether model outputs remain dependable under real-world constraints such as data quality issues, distribution shifts, and the need for human oversight (Afroogh et al., 2024). In practice, trustworthiness is often applied via two complementary mechanisms: *eXplainable Artificial Intelligence* (XAI), which makes model reasoning inspectable, and *uncertainty quantification* (UQ), which makes model limitations explicit.

XAI answers the question “why did the model make this prediction?”, by highlighting factors that drove the output and enabling users to understand and challenge decisions. Methodologically, XAI spans attribution and feature-importance approaches, example- or prototype-based explanations, surrogate/interpretable models, and counterfactual explanations and recourse; these methods support debugging, compliance and auditing, and user-facing decision support (Schwalbe and Finzel, 2024; Tjoa and Guan, 2020; Danilevsky et al., 2020). However, many explanation interfaces implicitly condition on a single predicted outcome, which can create the impression that the prediction itself is reliable. A plausible explanation does not communicate *how confident* the model is; when evidence is weak, the case is unusual, or the model is operating outside what it learned, the decision may be unreliable, motivating UQ as the second branch of trustworthy AI.

Uncertainty quantification answers “how sure is the model?”, by warning when the evidence is weak, the case is unusual, or the model is operating outside what

it learned. Operationally, UQ supports calibrated decision thresholds, selective prediction (reject/defer), and risk-aware decision-making under distribution shift. Recent work increasingly treats explanation and uncertainty as two sides of the same coin for trustworthy decision support (Salvi et al., 2025). A key link is that uncertainty can inform when not to rely on an explanation: an explanation may appear clear while being unreliable if the model is highly uncertain (Bilkhoo et al., 2025). Conversely, explanations can make uncertainty actionable by helping users identify missing information or features that drive the model’s doubt. Several works explicitly position uncertainty estimation as part of the XAI toolbox. Bai et al. (2021) list uncertainty estimation as a distinct category in XAI, next to attribution-based and non-attribution-based methods, and connect uncertainty to desirable properties such as trust, actionability, and robustness (Doshi-Velez and Kim, 2017). More recently, Thuy and Benoit (2024) argue that uncertainty estimation can be viewed as a local, model-specific explanation, and they emphasize actionability via classification with rejection, where uncertain cases are deferred to a human expert. They also stress that operations-research literature often ignores distribution shifts when discussing actionability, motivating uncertainty-aware decision support under changing data.

Although several strands of work connect XAI and Uncertainty Quantification (UQ), a fully principled and widely adopted integration remains an open methodological challenge (Mehdiyev, Majlatow, and Fettke, 2023). Recent studies have nevertheless started to link uncertainty-aware reasoning with explanation and recourse, for example by using conformal prediction to avoid high-uncertainty regions and provide formal guarantees, and by analysing how different uncertainty sources affect the robustness of counterfactual explanations (Bilkhoo et al., 2025; Christodoulou and Sun, 2026). In this thesis, we focus primarily on UQ for trustworthy decision-making, while treating XAI as a complementary perspective. Given the breadth of XAI methods and application domains, we refer the reader to recent surveys and taxonomies for a structured overview (Schwalbe and Finzel, 2024; Tjoa and Guan, 2020; Danilevsky et al., 2020).

UQ provides a principled means of formalizing trustworthiness in predictive systems. Rather than returning only a single prediction or probability, UQ seeks to estimate how confident the model is in its stated output. This distinction is essential: predictions accompanied by low uncertainty can typically support automated decision-making, whereas predictions with high uncertainty warrant caution, additional data, or human oversight. In this way, UQ offers an operational mechanism for incorporating trust into predictive workflows and for systematically managing the reliability of algorithmic decisions.

The practical relevance of UQ is particularly evident in business applications such as insurance. Consider the example of churn prediction, where predictive models identify customers likely to leave. Marketing teams then design and deploy retention actions for these customers. However, retention budgets are often limited,

making it impossible to intervene on every customer predicted to churn. In such cases, UQ plays a critical role: it enables decision-makers to prioritize customers for whom the model predicts a high likelihood of churn with low associated uncertainty. This ensures that limited resources are allocated efficiently, supports targeted interventions, and reduces ineffective or costly retention efforts. More broadly, by integrating uncertainty estimates with risk assessments and operational constraints, UQ helps with strategic and financially responsible business decisions. In large-scale operational settings such as insurance, banking, or healthcare, even small improvements in identifying high-uncertainty cases can translate into substantial financial savings or risk mitigation. UQ also supports compliance with emerging regulatory frameworks like the EU AI Act, which require transparency regarding model confidence and decision reliability.

Different research communities have approached UQ in different ways. In actuarial science and risk analytics, uncertainty has traditionally been modeled using frameworks such as belief functions and evidence theory, which explicitly distinguish between belief, plausibility, and residual ignorance (Walley, 1991; Shafer, 1976; Smets and Kennes, 1994; Hüllermeier and Waegeman, 2021). In machine learning and deep learning, uncertainty is commonly addressed through probabilistic modelling, Bayesian methods, calibration techniques, and ensemble-based approaches that produce predictive distributions and confidence scores (Shaker and Hüllermeier, 2020; Depeweg et al., 2018; Hüllermeier and Waegeman, 2021). More recently, Reinforcement Learning (RL) has incorporated UQ into exploration strategies and robustness mechanisms, relying on ensemble value functions and uncertainty-aware policies to improve stability in noisy or perturbed environments (Lockwood and Si, 2022; Singh et al., 2025d). Collectively, these developments underscore the broad relevance of UQ and motivate a systematic investigation across predictive modelling, ensemble learning, and RL.

Much of the prior work distinguishes between two types of uncertainty: aleatoric and epistemic. Aleatoric uncertainty captures the inherent randomness or noise present in the data generating process and cannot be reduced even with additional information (Hüllermeier and Waegeman, 2021). In contrast, epistemic uncertainty reflects a model's lack of knowledge, arising from limited data, insufficient model capacity, or ambiguity in the input and can often be reduced by acquiring more data or improving the model (Hüllermeier and Waegeman, 2021). This distinction is fundamental in prior work, as different methodologies target different uncertainty sources: belief-function models explicitly represent ignorance, probabilistic and Bayesian approaches quantify both forms through predictive distributions, and ensemble-based or exploratory RL methods predominantly address epistemic uncertainty by capturing model variability.

Although the aleatoric–epistemic distinction is now standard in machine learning (Hüllermeier and Waegeman, 2021), the term “uncertainty” is used differently

across disciplines. In statistics and operations research, it is often discussed as parameter uncertainty, model misspecification, and the reliability of estimates under changing conditions, with an emphasis on decision-making under limited information (Ben-Tal, Nemirovski, and El Ghaoui, 2009; Hansen and Sargent, 2001; Delage and Ye, 2010). In finance and macroeconometrics, uncertainty is commonly framed as unpredictability in the data-generating environment (e.g., time-varying volatility and regime changes) (Jurado, Ludvigson, and Ng, 2015) and as estimation error in inputs to optimization (e.g., expected returns and covariances) that can destabilize downstream decisions (Michaud, 1989; Black and Litterman, 1992). Recent work in asset pricing similarly argues that predictive uncertainty should be carried into portfolio construction rather than treated as a point forecast (Liao et al., 2025). This cross-disciplinary perspective clarifies a key theme of this thesis: uncertainty can describe (i) randomness in outcomes, (ii) lack of knowledge in a predictive model, and (iii) decision sensitivity to imperfect inputs (Hüllermeier and Waegeman, 2021; Ben-Tal, Nemirovski, and El Ghaoui, 2009).

Knightian uncertainty points to a limit of most uncertainty methods discussed above. Much of the finance and ML literature focuses on *quantifiable* uncertainty, meaning uncertainty that can be expressed with probabilities, variances, or estimated intervals. Knightian uncertainty instead refers to “unknown unknowns”, where the relevant distributions or even the correct model class are not known. Lehman et al. (2025) argues that standard ML, including common reinforcement learning formalisms, often assumes a closed world with fixed state and action spaces and approximately stationary data. By contrasting RL with biological evolution, they highlight robustness to qualitatively novel futures as a missing capability. This distinction matters for our setting: measuring environmental uncertainty (Jurado, Ludvigson, and Ng, 2015) or attaching confidence intervals to ML forecasts (Liao et al., 2025) addresses risk under modeled assumptions, but it does not fully cover regime shifts or structural breaks that can invalidate those assumptions. In this thesis, we therefore focus on uncertainty that can be estimated from data and communicated with calibrated scores or intervals, while acknowledging Knightian uncertainty as an important but separate research direction.

We therefore focus on uncertainty quantification as a practical mechanism to support trustworthy decision systems under limited oversight. Uncertainty quantification has been studied through multiple methodological perspectives. In work grounded in Dempster–Shafer theory and its extensions, such as the Belief Functions framework and the Transferable Belief Model, early research focuses on explicitly modelling belief, plausibility, conflict, and ignorance (Dempster, 2008; Shafer, 1976; Smets and Kennes, 1994). While these approaches yield interpretable representations of incomplete information, most formulations rely on handcrafted mass-assignment rules and do not easily scale to high-dimensional or modern machine learning pipelines. Conformal Prediction has evolved in parallel as a distribution-free framework offering finite-sample valid prediction regions (Fontana, Zeni, and

Vantini, 2023), with extensions addressing computational efficiency (Papadopoulos, Vovk, and Gammerman, 2007), non-exchangeable data (Barber et al., 2023), and complex modalities. However, CP primarily provides distribution-free intervals rather than structural uncertainty decompositions, and its applicability in decision-support tasks such as churn prediction remains limited. Moreover, CP offers only marginal coverage guarantees; under severe class imbalance, post-hoc calibration often favours the majority class, leading to inflated prediction sets for minority instances.

Machine learning research has introduced a wide range of Bayesian, ensemble-based, and calibration-aware methods that distinguish between aleatoric and epistemic uncertainty (Hüllermeier and Waegeman, 2021; Shaker and Höllermeier, 2020). Techniques such as MC-dropout (Gal and Ghahramani, 2016), deep ensembles (Lakshminarayanan, Pritzel, and Blundell, 2017), and Bayesian neural networks (Caldeira and Nord, 2020) offer practical uncertainty estimates, yet most existing studies focus on homogeneous ensembles, black-box architectures, or small benchmark datasets, and rarely examine how feature transformations or model heterogeneity affect uncertainty quality. In RL, uncertainty plays a central role in exploration and robustness, motivating ensemble based value functions, Bayesian critics, uncertainty weighted updates, and inverse-variance formulations (Liu et al., 2021; Lee et al., 2021; Mai, Mani, and Paull, 2022; Mavor-Parker et al., 2022). However, existing RL methods often emphasize exploration rather than decision stability, leaving open questions about how epistemic uncertainty can be directly incorporated into the Bellman update to improve robustness under noisy or perturbed environments.

Building on these complementary perspectives, this research investigates three major classes of uncertainty-quantification methods, namely Belief Functions (BF), conformal prediction (CP) (Chapter 2), and our heterogeneous ensemble framework Heterogeneous Ensemble for Uncertainty Quantification (Chapter 3), to quantify uncertainty in customer-churn prediction for our industrial partner. To ensure robustness and generalizability, we also evaluate these methods on publicly available churn datasets from the banking and telecom domains, alongside the proprietary industrial partner dataset (Section 2.5).

Beyond supervised learning, this thesis also extends the study of uncertainty to Reinforcement Learning (Chapter 4). In that chapter, epistemic uncertainty is incorporated into an uncertainty-aware variant of the Bellman update to improve learning stability under severe noise and real-world data perturbations. Taken together, the chapters provide a unified perspective on how uncertainty can be quantified, interpreted, and used across both predictive modelling and RL.

Despite substantial progress across these research streams, important gaps remain. We identify three main research gaps that motivate the research questions addressed in this thesis:

- **Belief Functions (BF) and Conformal Prediction (CP).** The first gap concerns the behaviour of BF and CP on noisy, imbalanced, and weakly separable real-world datasets. Our experiments on the proprietary churn datasets (Sections 2.5.1 and 2.5.2) show that both methods can struggle under severe class imbalance, noisy or missing labels, and limited feature separability. For BF, mass assignments often concentrated on composite sets such as {churn, non-churn}, indicating substantial residual ignorance and instability when conflicting evidence was combined. For CP, prediction sets became overly large under weak separability, reducing their usefulness for targeted intervention. More broadly, prior work does not yet clearly establish under which real-world data conditions these methods cease to provide actionable uncertainty estimates.
- **Supervised ML and Ensembles.** The second gap concerns uncertainty quantification in supervised Machine Learning (ML), particularly for heterogeneous ensembles under practical operational constraints. Most existing work focuses on homogeneous ensembles, even though heterogeneous ensembles often improve predictive performance in practice. In our empirical analysis, homogeneous ensembles often produced tightly clustered epistemic uncertainty estimates, suggesting limited diversity in the explored hypothesis space. In addition, the effect of feature transformations such as PCA and random Gaussian projections on both predictive performance and uncertainty estimates remains insufficiently studied.
- **Reinforcement Learning (RL).** The third gap concerns uncertainty-aware RL. Existing methods mainly use uncertainty for exploration and often rely on large ensembles, which makes them computationally expensive and sensitive to observation noise, non-stationarity, and correlated value estimates. Bootstrap-based methods can suffer from inter-head correlation, reducing the intended benefit of bootstrapping, while lightweight alternatives such as MC-dropout have been reported to underperform in RL settings (Charpentier et al., 2022; Durasov et al., 2021). This suggests a need for computationally efficient uncertainty-aware RL methods that improve stability under noisy observations and are relevant to practical business decision problems.

Taken together, these limitations define the main research gaps that this thesis addresses. We formalize these gaps into the following research questions:

1. **RQ1 (Belief Functions and Conformal Prediction).** For what kinds of data do belief-function methods and conformal prediction provide reliable and actionable uncertainty estimates, and how well do they scale to noisy, imbalanced business datasets such as operational churn prediction?
2. **RQ2 (Supervised ML & Ensembles).** How can heterogeneous ensembles combining different learning paradigms improve epistemic uncertainty estimation

compared to homogeneous ensembles, and what is the impact of feature transformations on the quality and stability of uncertainty estimates?

3. **RQ3 (Reinforcement Learning).** How can we design lightweight, uncertainty-aware RL architectures that are computationally efficient, robust to extreme observation noise, and applicable to real-world business decision problems such as churn management?

In response, we develop new uncertainty-aware architectures in both supervised ML and RL, and we systematically evaluate their predictive performance and uncertainty properties on real-world insurance data as well as standard benchmark environments.

To advance the state of the art in uncertainty-aware decision-making, this thesis proposes two novel architectures:

- **HEUQ (Heterogeneous Ensemble for Uncertainty Quantification).** A heterogeneous ensemble framework for supervised machine learning that provides per-customer uncertainty estimates and is tailored to churn and conversion prediction tasks.
- **MASURE (Masksembles for Stable and Uncertainty-Aware Reinforcement Learning Environments).** An uncertainty-aware RL framework that incorporates Masksembles into value-based RL to stabilize temporal-difference learning under extreme noisy observations and to exploit epistemic uncertainty for improved stability during training.

We focus on ML and RL because these paradigms underpin many state-of-the-art decision systems in industry, including insurance pricing, risk modelling, and customer management. Furthermore, our empirical results show that purely evidence-theoretic models (Belief Functions) and purely post-hoc calibration (Conformal Prediction) face important data-related limitations in our insurance case study. This motivates the move towards ensemble-based ML and RL as the main vehicles for scalable, uncertainty-aware decision support.

ML/HEUQ

1. We formalize total, epistemic, and aleatoric uncertainty for heterogeneous ensembles, extending the definitions of Depeweg et al. (2018) and Shaker and Hüllermeier (2020) to heterogeneous ensembles.
2. We propose HEUQ, a heterogeneous ensemble that delivers competitive or superior predictive performance and improved calibration compared with strong baselines, including deep ensembles, across churn and conversion tasks in multiple domains (banking, insurance, telecom).
3. We study how data transformations (Principal Component Analysis (PCA), Random Gaussian (RG) projections) affect both performance and uncertainty

estimates, showing that HEUQ remains robust under noise and transformations.

4. We perform an ablation study on homogeneous and heterogeneous ensemble composition, linking specific learners (e.g., shallow neural networks) to systematic changes in aleatoric and epistemic uncertainty and analyzing when epistemic uncertainty becomes less informative in homogeneous ensembles.

RL/Uncertainty aware RL (UARL)

1. We design an RL framework that integrates Masksembles into RL, enabling epistemic uncertainty estimates without large ensembles.
2. We propose an uncertainty-conscious extension to Bellman’s equation that improves training stability and robustness under extreme noisy environments.

Ancillary contributions. In addition to these main architectures, the thesis includes an ancillary contribution on UQ beyond machine learning approaches. In Chapter 2, we empirically investigate Belief Functions and Conformal Prediction on the real-life insurance churn data used by HEUQ and Masksembles for Stable and Uncertainty-Aware Reinforcement Learning Environments (MASURE). Belief functions allow us to model ignorance and conflicting evidence explicitly, while Conformal Prediction provides model-agnostic prediction sets with finite-sample coverage guarantees over probabilistic ML models. These experiments reveal practical limitations (e.g., instability under label noise, large prediction sets under weak separability) that motivate the ensemble-based architectures developed in later chapters.

To ensure that the methods developed in this thesis are robust, generalizable, and applicable to real-world decision-making, we evaluate them across a diverse collection of datasets, both proprietary datasets from our industrial partner and publicly available benchmark datasets. Details regarding the different versions of the proprietary churn datasets used throughout this thesis are provided in Chapter 2 (Section 2.5). Throughout this thesis, the terms “proprietary churn dataset” and “churn dataset” are used interchangeably to refer to the same underlying data source. A public repository for the code and public datasets used in this thesis is available at github.com/Rajawat23/DoctorateResearch.

Table 1.1 summarizes the links between datasets, chapters, and methodological frameworks.

This mapping highlights the empirical breadth of the thesis. Every methodological contribution is validated on both industrial datasets and open academic benchmarks when relevant. Chapters 2 and 3 focus on supervised learning and uncertainty quantification under real-world data (imbalance data, noisy data), whereas Chapter 4 extends these ideas to RL through a noisy-robust Bellman update. Together, these datasets provide a comprehensive foundation for assessing the reliability, stability, and applicability of the proposed uncertainty-aware methods.

TABLE 1.1: Overview of datasets and their connection to chapters and methodologies.

Dataset	Chapter	Methodology
Proprietary churn dataset (Version 1)	Chapter 2	Belief Functions
Proprietary churn dataset (Version 2)	Chapter 2	Conformal Prediction
Open credit card fraud dataset	Chapter 2	Conformal Prediction
Proprietary churn dataset (Version 2)	Chapter 3	Heterogeneous Ensemble for Uncertainty Quantification
Open Bank churn dataset	Chapter 3	Heterogeneous Ensemble for Uncertainty Quantification
Open Telecom churn dataset	Chapter 3	Heterogeneous Ensemble for Uncertainty Quantification
Proprietary Customer conversion dataset	Chapter 3	Heterogeneous Ensemble for Uncertainty Quantification
Proprietary churn dataset (Version 2)	Chapter 4	Masksembles for Stable and Uncertainty-Aware Reinforcement Learning Environments
OpenAI Gym environments	Chapter 4	Masksembles for Stable and Uncertainty-Aware Reinforcement Learning Environments

This thesis is organized as a collection of published and submitted scientific articles. Chapter 2 presents the work on belief functions and conformal prediction, Chapter 3 introduces the HEUQ framework for heterogeneous ensemble uncertainty quantification, and Chapter 4 addresses uncertainty in RL. In Chapter 5, we present the business use case developed in collaboration with our industrial research partners.

The main contributions of this thesis build on the following scientific publications. The poster and white paper *Beyond Yes or No: Making Reliable Decisions and Improving Personalised Customer Service* was presented as a single contribution that spans both Chapter 2 and Chapter 3:

- **Chapter 2 — Belief Functions and Conformal Prediction**

- Singh, A., Ittoo, A., Ars, P., & Vandomme, E. (2025). *Conformal Prediction: Calibrated Decision-Making*. Conference abstract accepted and presented at the Joint ORBEL–NGB Conference on Operations Research.
- Singh, A., Ittoo, A., Ars, P., & Vandomme, E. (2025). *Beyond Yes or No: Making Reliable Decisions and Improving Personalised Customer Service*. Poster presented at HEC Liège Research Day.
- Singh, A., Ittoo, A., & Ars, P. (2024). *Beyond Yes or No: Making Reliable Decisions and Improving Personalised Customer Service*. White Paper written with our industrial partners.

- **Chapter 3 — HEUQ: Heterogeneous Ensemble for Uncertainty Quantification**

- Singh, A., Ittoo, A., Ars, P., & Vandomme, E. (2025). *Multi-channel Multi-model Customer Conversion Prediction in the Insurance Domain*. Peer-reviewed chapter published and presented in 9th International Conference on Applied Economics and Business (Springer Nature Computer Science).
- Singh, A. (2024). *Heterogeneous Ensemble for Uncertainty Quantification (HEUQ) in Churn Management*. Poster presented at “Research in the Age of AI”, HEC Liège.

This poster presented part of the ongoing work that is being extended into a journal paper in preparation discussed in Chapter 3.

- Singh, A., Ittoo, A., Ars, P., & Vandomme, E. (2025). *Beyond Yes or No: Making Reliable Decisions and Improving Personalised Customer Service*. Poster presented at HEC Liège Research Day.
 - Singh, A., Ittoo, A., & Ars, P. (2024). *Beyond Yes or No: Making Reliable Decisions and Improving Personalised Customer Service*. White Paper written with our industrial partners.
 - Singh, A., Ittoo, A., Vandomme, E., & Ars, P. *HEUQ: A Model-Agnostic Heterogeneous Ensemble for Uncertainty Quantification*. Manuscript in preparation.
- **Chapter 4: Uncertainty-Aware Reinforcement Learning**
 - Singh, A., Ittoo, A., Vandomme, E., & Ars, P. (2025). *Uncertainty-Aware Reinforcement Learning Agents for Noisy Environments*. Peer-reviewed paper accepted for publication in ITT 2025 (IEEE). *To appear*.
- **Chapter 5: Conclusion**
 - Singh, A., Ittoo, A. (2023). *Uncertainty-Aware Learning for Customer Management and Decision Support*. Invited research talks presented at Young talent R&D seminar of our industrial research partners.

Chapter 2

Uncertainty Quantification with Belief functions and Conformal Predictions

This chapter investigates two uncertainty-quantification approaches, Belief Functions (BF) and Conformal Prediction (CP). Using proprietary and public datasets, we assess their suitability for decision-support and find that, despite strong theoretical foundations, their predictive performance is suboptimal in our applied setting. These limitations motivate machine learning based methods developed in later chapters. Parts of this work were presented at ORBEL (Singh et al., 2025b) and in poster and presentation sessions (Singh, Ittoo, and Ars, 2024; Singh et al., 2025a).

Contents

2.1 Introduction	12
2.2 Literature study	13
2.2.1 Belief Functions	13
2.2.2 Conformal Predictions	14
2.3 Background	16
2.3.1 Belief Functions	18
2.3.2 Conformal Prediction	22
2.4 Methodology	23
2.4.1 Belief Functions	23
2.4.2 Conformal Prediction	25
2.5 Data	26
2.5.1 Proprietary churn dataset (Version 1)	26
2.5.2 Proprietary churn dataset (Version 2)	28
2.5.3 Open credit card fraud dataset	29
2.6 Results	30
2.6.1 Belief Functions	30
2.6.2 Conformal Prediction	31
2.7 Discussion	33

2.1 Introduction

Before investigating modern machine learning-based UQ methods, our collaboration with our industrial partners motivated an initial assessment of whether established actuarial tools such as BF could support prediction tasks in insurance. This chapter therefore examines two classical approaches to uncertainty quantification: Belief Functions (BF) (Dempster, 1967) and Conformal Prediction (CP) (Papadopoulos, Vovk, and Gammerman, 2007). We subsequently evaluated CP as a complementary statistical framework to determine whether these classical methods offer practical value for real-world prediction problems.

Both BF and CP produce *set-valued predictions* rather than single labels, offering structured representations of uncertainty in the form of sets of plausible outcomes (Hüllermeier and Waegeman, 2021).

BF is an Uncertainty Quantification method rooted in Dempster–Shafer Theory (DST). The method represent uncertainty by distributing belief mass over all subsets of the label space, not only over single labels. This enables the model to express incomplete, conflicting, or ambiguous evidence in a principled way. For a binary classification task with labels $\{0, 1\}$, a BF model may assign belief to:

$$\{0\}, \quad \{1\}, \quad \{0, 1\}.$$

The set $\{0, 1\}$ corresponds to an explicit representation of uncertainty.

CP is a distribution-free statistical framework that constructs prediction sets by measuring how atypical a new observation is relative to a calibration dataset. Given a target confidence level (e.g., 90%), CP guarantees that the resulting prediction set contains the true label with at least that probability under minimal assumptions. Unlike BF, which relies on evidential reasoning, CP derives uncertainty from non-conformity scores produced by any underlying machine learning model.

The theoretical background for both approaches, including a brief illustrative example using the publicly available `iris` dataset, is presented in Section 2.3. Building on these foundations, this chapter investigates how BF and CP behave when applied to real-world noisy data, including churn prediction and credit-card fraud detection.

This chapter addresses **RQ1** by examining two UQ frameworks, namely Belief Functions and Conformal Prediction, on proprietary industrial datasets. More specifically, we study the following chapter-level research questions:

1. **BF:** Under which conditions are belief-function methods reliable, and how well do they perform on noisy, imbalanced real-life industrial datasets?
2. **CP:** Under which conditions Conformal Prediction deliver valid and useful prediction sets (e.g., 90% coverage), and how well do they handle class imbalance and noise?

The aim of this chapter is to assess the practical usefulness of these classical uncertainty quantification approaches in actuarial contexts, particularly their transparency, interpretability, and suitability for insurance applications such as churn prediction and credit-card fraud detection.

The remainder of this chapter is organized as follows. Section 2.2 presents the literature review covering both BF and CP. Section 2.3 provides the theoretical foundations of these two frameworks and includes an illustrative example using the iris dataset. Section 2.4 outlines the methodological framework adopted in this study. Section 2.5 describes the datasets used for the empirical evaluation. Finally, Section 2.7 discusses the results and their implications.

2.2 Literature study

2.2.1 Belief Functions

Dempster (2008) and Shafer (1976) propose a general framework for reasoning under uncertainty when information is incomplete, imprecise, or conflicting. Unlike Bayesian probability, which distributes belief over single hypotheses, BF assigns masses to subsets of the frame of discernment Θ , distinguishing *uncertainty* from *ignorance*. Smets and Kennes (1994) later generalized BF into the Transferable Belief Model (TBM) in their work, which separates credal and pignistic levels of reasoning (Smets and Kennes, 1994; Smets, 2002). The compendium by Yager and Liu (2008) consolidates foundational works and subsequent extensions connecting BF with fuzzy sets, possibility theory, and decision analysis. Early critiques such as Zadeh (1984) emphasized the limitations of Dempster’s normalization under high conflict, motivating alternative combination operators including Yager (1987), Dubois and Prade (1988), Inagaki (2002), and Smets (2002).

Subsequent studies expanded belief function theory toward data-driven uncertainty modelling and classification. Denoeux (1995), Zouhal and Denoeux (1998), and Denoeux (2000) demonstrated that classifiers such as evidential k -NN and neural networks can be expressed within the Dempster-Shafer framework, interpreting feature-based evidence as basic belief assignments (BBAs). Chen et al. (2014) proposed a Dempster–Shafer method for data classification, where each attribute acts as an independent source of evidence, combined through Dempster’s rule. Yong et al. (2004) introduced uncertainty modelling in classification using distance measures, while Wang, Liu, and Augusto (2010) presented systematic method to derive mass functions from multivariate data samples and an efficient approach to combine these derived mass functions. These formulations enable interpretable and conflict-aware decisions, often outperforming probabilistic baselines in the presence of ambiguity or overlapping classes.

Various studies proposed mechanisms for constructing BBAs directly from feature data or distributions. Denoeux (1995) proposed a k-nearest neighbor classification rule based on the Evidence Theory, which derives BBA through the Euclidean distance between test samples with k nearest neighbors. Which was further improved in subsequent studies of Zouhal and Denoeux (1998) and Liu, Pan, and Dezert (2013). Lefevre, Vannoorenberghe, and Colot (1999) proposed a classification framework within the Dempster–Shafer theory that initializes basic belief assignments from learned probability densities and then attenuates these beliefs using an information-theoretic coefficient based on the Hellinger distance between class distributions, effectively discounting unreliable sources before fusion. Wei and Yaowen (2013) proposed a supervised feature-based method that constructs basic belief assignments by optimizing reference vectors and dispersion parameters through learning, deriving beliefs from Euclidean distances without nearest-neighbor search and enabling robust uncertainty–ignorance differentiation. Other works such as Aregui and Denœux (2008) and Wang, Liu, and Augusto (2010) formalizes an empirical estimation strategy where the mass of each hypertuple (the frame of discernment) is defined to be proportional to the number of data items it covers. These approaches illustrate that uncertainty quantification can be achieved without strict probabilistic assumptions, leveraging BF’s ability to explicitly encode ignorance.

Complementary research explored computational scalability and extended applications. Reineking (2014) developed Monte Carlo and particle-filter approximations for evidential reasoning, enabling efficient belief updates in sequential problems such as SLAM. Zhou, Tian, and Deng (2023) further proposed a quantum-circuit formulation (BF-QC) for belief function operations as belief functions because of the computation complexity exploding exponentially with the increasing number of elements, the practical application scenarios of BF are limited.

Overall, belief functions have evolved from a mathematical formulation into a practical paradigm for interpretable uncertainty quantification. Their ability to separate confidence, ignorance, and conflict makes them well suited for data with low number of features or evidences, and reliability analysis.

2.2.2 Conformal Predictions

Fontana, Zeni, and Vantini (2023) present a comprehensive theoretical and conceptual survey of Conformal Prediction (CP), offering a unified review of its foundational principles and the wide range of methodological extensions that have emerged over time. The authors outline how CP generates statistically valid prediction sets under minimal distributional assumptions, establishing it as a general-purpose framework for uncertainty quantification.

Building upon these foundations, several studies have advanced CP to address key limitations in practical and theoretical settings. Papadopoulos, Vovk, and Gammerman (2007) introduce Inductive Conformal Prediction (ICP) for Neural Network

(NN), a computationally efficient adaptation of the original CP method. Their experiments on the Satellite, Shuttle, and Segment datasets demonstrate that ICP achieves accuracy comparable to traditional models while eliminating the high computational cost of standard CP. Similarly, Papadopoulos, Vovk, and Gammerman (2011) extend CP to the k-Nearest Neighbours Regression (k-NNR) algorithm, proposing six new nonconformity measures that yield significantly tighter predictive regions compared to traditional regression-based CPs.

Other extensions improve CP’s ability to handle non-exchangeable or structured data. Barber et al. (2023) generalize CP for non-exchangeable settings, such as distributional drift using weighted quantiles and a novel randomization scheme. Their approach maintains coverage stability and predictive validity even when exchangeability is violated, with empirical validation on electricity and election forecasting data. In a similar direction, Tibshirani et al. (2019) highlight the central role of the exchangeability assumption in guaranteeing finite-sample coverage and propose a weighted CP framework to relax it under covariate shift. Weighted CP has since been successfully applied in diverse fields including experimental design Fannjiang et al. (2022), survival analysis Candes, Lei, and Ren (2023), and causal inference Lei and Candès (2021). Extending this line, Kiyani, Pappas, and Hassani (2024) propose Partition Learning Conformal Prediction (PLCP), which learns uncertainty-guided features from calibration data to enhance conditional validity. Their theoretical and empirical analysis across real-world and synthetic datasets (e.g., US Census, MNIST) shows improved coverage and interval tightness relative to state-of-the-art methods.

Several studies propose methodological refinements tailored to data modality and structure. Lei et al. (2018) introduce Split CP, which pre-fits a model and evaluates it using a separate calibration set to obtain nonconformity scores, ensuring finite-sample marginal coverage. Vovk et al. (2018) extend this idea with Cross-CP, dividing the dataset into K disjoint subsets to generate cross-validation residuals and more stable predictions. Chernozhukov, Wüthrich, and Zhu (2021) further propose Distributional Conformal Prediction (DCP), which builds robust and conditionally valid prediction intervals using conditional distribution models such as quantile regression. Their approach achieves reliable k -step-ahead forecasts on financial datasets like the Fama/French 3 Factors data. Likewise, Xu and Xie (2023) introduce EnbPI (Ensemble-based Prediction Interval), a scalable CP variant for time-series data that avoids costly retraining while providing theoretical guarantees on conditional and marginal coverage.

Beyond methodological generalization, several studies extend CP for specialized learning contexts and complex data. Cauchois, Gupta, and Duchi (2021) apply CP to multi-class and multi-label image classification, addressing uneven coverage across easy and difficult samples through tree-structured classifiers. Ding et al. (2023) focus on class-conditional coverage, ensuring valid prediction intervals even in large-class settings where sample sizes per class are limited. Similarly, Straitouri et al.

(2023) propose an automated decision-support system leveraging CP to guide human experts by presenting a guaranteed set of possible labels, achieving nearly 80% reduction in expert misclassification on synthetic and CIFAR-10H datasets.

The Mondrian Conformal Prediction (MCP) framework Papadopoulos, Gamerman, and Vovk (2009) addresses label imbalance by treating each class separately, ensuring group-wise error control. Label-Conditional MCP (LCMCP) refines this idea further by conditioning on the true label. Pereira et al. (2020) employ MCP to estimate confidence intervals for predicting dementia progression using KNN and Naïve Bayes classifiers, demonstrating that the resulting conformal regions reliably contain the correct class within the desired confidence level, though further optimization is needed for clinical use.

Applications of CP across domains have demonstrated its broad utility. Vazquez and Facelli (2022) review fourteen studies applying CP to clinical medical sciences, including dementia progression, stroke risk, and cancer survivability prediction. Their analysis shows that CP-based methods such as ICP, MCP, and DCP, offer valuable per-instance uncertainty estimates not captured by global metrics, while highlighting the need for standardized clinical evaluation protocols. In a broader data-centric perspective, Zhou et al. (2025) propose a taxonomy that organizes CP developments by data type (static, dynamic, structured, or unstructured). They discuss challenges posed by dependencies in spatio-temporal and streaming data, and review specialized approaches such as Split CP and Weighted CP for handling covariate shifts and unbounded output spaces in large-scale models like LLMs.

Collectively, these studies demonstrate how Conformal Prediction has evolved from a theoretically elegant framework into a diverse family of methods capable of handling structured data, non-exchangeable distributions, and domain-specific challenges while retaining its core property of finite-sample validity.

2.3 Background

This section introduces the theoretical foundations of Belief Functions (BF) and Conformal Prediction (CP). We begin with a short description of the Iris dataset, which will be used throughout this section to provide illustrative examples.

Example Dataset: Iris

The Iris dataset¹ consists of 150 samples of iris flowers belonging to three species: *Setosa*, *Versicolor*, and *Virginica*. Each sample is described by four numerical attributes: sepal length, sepal width, petal length, and petal width, measured in centimeters. The dataset is widely used as a pedagogical example because the three species form relatively distinct clusters in feature space, while still exhibiting partial overlap in

¹UCI Machine Learning Repository: <https://archive.ics.uci.edu/dataset/53/iris>

specific attributes (e.g., petal length between Versicolor and Virginica). This mixture of separability and overlap makes it ideal for illustrating the principles of BF and CP.

Statistical Characteristics. The four numerical features in the Iris dataset exhibit distinct ranges and distribution patterns across the three species. Representative

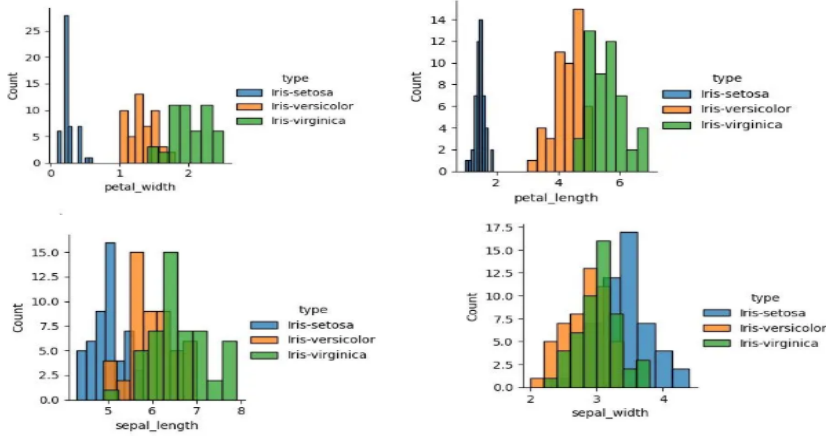


FIGURE 2.1: Data distribution of Iris features values by species.

intervals include:

Petal Length (cm): Setosa: [1.0, 1.9], Versicolor: [3.0, 5.1], Virginica: [4.5, 6.9],

Petal Width (cm): Setosa: [0.1, 0.6], Versicolor: [1.0, 1.8], Virginica: [1.4, 2.5],

Sepal Length (cm): Setosa: [4.3, 5.8], Versicolor: [4.9, 7.0], Virginica: [4.9, 7.9],

Sepal Width (cm): Setosa: [2.3, 4.4], Versicolor: [2.0, 3.4], Virginica: [2.2, 3.8].

These intervals highlight the strong separability of *Setosa* from the other species, particularly in petal-based features, while revealing substantial overlap between *Versicolor* and *Virginica*. As a result, features such as petal length and petal width provide both well-separated and overlapping regions, allowing clear demonstration of how set-based uncertainty representations—such as Belief Functions and Conformal Prediction—allocate mass or construct prediction sets in ambiguous zones.

Example 2.3.1. Feature Ranges in the Iris Dataset: The *Iris* dataset illustrates how class-overlapping regions naturally arise in real feature distributions. For instance, the feature *petal length* exhibits the following non-overlapping and overlapping intervals:

Setosa: [1.0, 1.9], Versicolor: [3.0, 5.1], Virginica: [4.5, 6.9].

Based on petal length feature, *Setosa* is clearly separated from the other two species, whereas *Versicolor* and *Virginica* exhibit substantial overlap, making the class prediction difficult. This overlap will later be used to illustrate how Belief Functions

allocate mass to composite hypotheses and how Conformal Prediction constructs multi-label prediction sets in such ambiguous regions.

2.3.1 Belief Functions

Belief Functions, based on the Dempster–Shafer Theory (DST), provide a flexible framework for representing and reasoning under uncertainty. Unlike classical probability theory, which assigns probability mass to singleton outcomes, BF permits belief mass to be allocated to *subsets* of outcomes, allowing explicit representation of uncertainty and ignorance.

Consider the classification task on the Iris dataset, where the set of possible classes (i.e., species) is:

$$\Theta = \{\text{Setosa}, \text{Versicolor}, \text{Virginica}\}.$$

Definition 1 (Frame of Discernment). The *frame of discernment* Θ is the finite set of all mutually exclusive and exhaustive hypotheses. In classification, it corresponds to the set of all possible class labels.

Definition 2 (Power Set). The *power set* 2^Θ is the set of all subsets of Θ , including singleton hypotheses, composite hypotheses, and Θ itself.

For $\Theta = \{\text{Setosa}, \text{Versicolor}, \text{Virginica}\}$, the power set contains eight subsets.

Definition 3 (Basic Belief Assignment). A *basic belief assignment* (BBA), or *mass function*, is a mapping

$$m : 2^\Theta \rightarrow [0, 1],$$

such that

$$\sum_{A \subseteq \Theta} m(A) = 1, \quad m(\emptyset) = 0.$$

The quantity $m(A)$ expresses the amount of belief committed exactly to the hypothesis A .

Definition 4 (Focal Elements). A subset $A \subseteq \Theta$ is called a *focal element* if $m(A) > 0$. Focal elements represent hypotheses to which evidence explicitly assigns support.

Definition 5 (Belief and Plausibility). For any subset $A \subseteq \Theta$, the *belief* and *plausibility* functions are defined as

$$Bel(A) = \sum_{B \subseteq A} m(B), \tag{2.1}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{2.2}$$

where B ranges over all subsets of Θ .

These represent, respectively, the lower and upper bounds on the probability of A .

Definition 6 (Dempster's Rule of Combination). Let m_1 and m_2 be two basic belief assignments defined on the same frame of discernment Θ , representing two independent sources of evidence. Their combination, denoted $m = m_1 \oplus m_2$, is defined for every non-empty $A \subseteq \Theta$ as

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}. \quad (2.3)$$

where $B, C \in 2^\Theta$.

The denominator

$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$$

quantifies the degree of conflict between the two sources. The rule normalizes by $1 - K$ to ensure that the resulting mass function satisfies $\sum_{A \subseteq \Theta} m(A) = 1$.

Example: Mass Assignment Using Histogram Overlap

In tabular datasets such as Iris(Example 2.3), each feature is interpreted as a source of evidence. For illustration, consider the feature *petal length*, whose empirical distributions for the three species partially overlap. Figure 2.2 shows the histogram distributions.

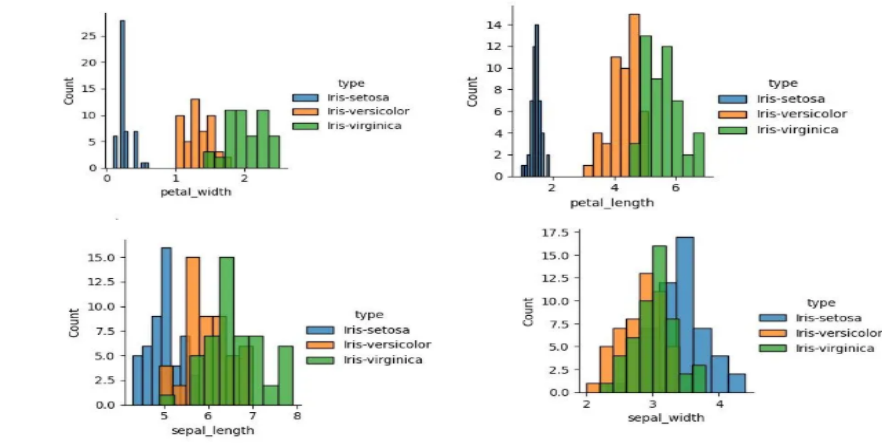


FIGURE 2.2: Histogram of petal length features by species. It is noticeable how Versicolor and Virginica overlap.

Let x denote the feature vector of a new sample, and let $x^{(j)}$ be the observed value of feature j (e.g., petal length). For each class $C \in \Theta$, the empirical density $p(x^{(j)} | C)$ is obtained from the histogram bin that contains $x^{(j)}$, that is, the value of the histogram-based density estimate at location $x^{(j)}$.

For each class $C \in \Theta$, the mass assigned to the singleton hypothesis $\{C\} \in 2^\Theta$ is defined as

$$m(\{C\}) = \frac{p(x^{(j)} | C)}{\sum_{c \in \Theta} p(x^{(j)} | c)}. \quad (2.4)$$

Note that $C \in \Theta$ is a class label (an element), while $\{C\} \subseteq \Theta$ is the corresponding singleton set; since $m : 2^\Theta \rightarrow [0, 1]$, masses are assigned to subsets such as $\{C\}$ rather than to labels directly.

When feature distributions overlap, belief is also assigned to composite subsets. For two overlapping classes C_k and C_l ,

$$m(\{C_k, C_l\}) = \frac{\min\{p(x^{(j)} | C_k), p(x^{(j)} | C_l)\}}{\sum_{c \in \Theta} p(x^{(j)} | c)}. \quad (2.5)$$

Any remaining uncertainty is attributed to total ignorance:

$$m(\Theta) = 1 - \sum_{c \in \Theta} m(\{C\}) - \sum_{k < l} m(\{C_k, C_l\}). \quad (2.6)$$

Equation 2.5 is not taken from a single canonical formula in the Belief Functions literature but follows a widely used principle in evidential classification: ambiguity between two classes should be quantified by the overlap of their evidential support. Dempster–Shafer theory does not prescribe how mass functions must be constructed (Chen et al., 2014). Thus the designed mass models that reflect domain-specific evidence characteristics. In regions of the feature space where two classes provide similar support for a given observation, the evidence cannot reliably discriminate between them. In such cases, BF require that a portion of the belief be assigned to their union $\{C_k, C_l\}$, representing explicit ambiguity rather than forcing a single-label decision. The minimum operator $\min\{p(x^{(j)} | C_k), p(x^{(j)} | C_l)\}$ quantifies the *shared evidence* that both classes support simultaneously, ensuring that the ambiguous mass is never larger than the evidence available for either class individually. This approach is consistent with overlap- and similarity-based heuristics commonly employed in evidential classification, where indistinguishability between hypotheses is captured by allocating mass to composite sets (Denoeux, 1995; Chen et al., 2014; Wang, Liu, and Augusto, 2010; Yong et al., 2004). The normalisation by $\sum_{c \in \Theta} p(x^{(j)} | c)$ ensures that all belief masses are non-negative and sum to one, as required for a valid BBA.

Thus, while not derived from a single canonical formula, this rule constitutes a valid and principled instantiation of a BBA construction tailored to histogram-based feature evidence, enabling DST to represent uncertainty arising from overlapping class distributions.

Example 2.3.2 (Belief Mass Assignment in an Overlapping Region). Building on Example 2.3.1, consider a flower whose petal length $x^{(j)}$ lies in the region of overlap between *Versicolor* and *Virginica* (see Figure 2.3). Assume that the belief masses become:

$$m(\{\text{Versicolor}\}) = 0.45, \quad m(\{\text{Virginica}\}) = 0.45, \quad m(\{\text{Versicolor}, \text{Virginica}\}) = 0.10.$$

From Definitions 5 and Equations (2.1)–(2.2), the belief and plausibility for the hy-

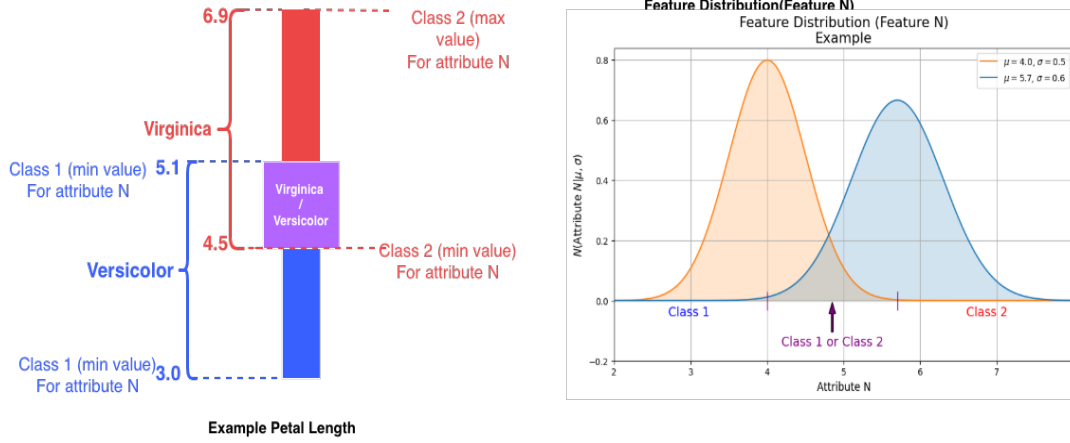


FIGURE 2.3: Illustration of petal-length overlap between Versicolor and Virginica.

pothesis “the flower is Versicolor” are

$$Bel(\{\text{Versicolor}\}) = m(\{\text{Versicolor}\}) = 0.45,$$

$$\begin{aligned} Pl(\{\text{Versicolor}\}) &= m(\{\text{Versicolor}\}) + m(\{\text{Versicolor}, \text{Virginica}\}) \\ &= 0.45 + 0.10 \\ &= 0.55. \end{aligned}$$

Here, $Bel(\{\text{Versicolor}\})$ captures the direct, unambiguous support for Versicolor, whereas $Pl(\{\text{Versicolor}\})$ also accounts for the ambiguous mass shared with Virginica. The interval $[0.45, 0.55]$ therefore quantifies the uncertainty regarding Versicolor: the true probability of Versicolor lies somewhere between these bounds, and the width 0.10 reflects how much evidence remains imprecisely allocated to the composite hypothesis.

Example 2.3.3 (Combining Evidence from Two Iris Features). Continuing with the same flower as in Example 2.3.2, suppose that petal length produces the masses

$$m_1(\text{Versicolor}) = 0.45, \quad m_1(\text{Virginica}) = 0.45, \quad m_1(\Theta) = 0.10,$$

let’s assume that petal width provides an additional, independent source of evidence:

$$m_2(\text{Versicolor}) = 0.60, \quad m_2(\text{Virginica}) = 0.30, \quad m_2(\Theta) = 0.10.$$

Using Dempster’s rule of combination in Equation (2.3), the combined masses become

$$m(\{\text{Versicolor}\}) = 0.62, \quad m(\{\text{Virginica}\}) = 0.32, \quad m(\Theta) = 0.06.$$

This result shows how evidence from multiple features can strengthen support for one class (Versicolor) while still retaining a small amount of residual ignorance through $m(\Theta)$.

2.3.2 Conformal Prediction

Conformal Prediction (CP) is a model-agnostic, post-hoc calibration method that constructs *prediction sets* with guaranteed finite-sample coverage. Instead of committing to a single predicted class, CP outputs a set of plausible labels whose size reflects the model's uncertainty.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote a labeled dataset. Let $\mathcal{Y}$ denote the label space (for Iris, $\mathcal{Y} = \{\text{Setosa}, \text{Versicolor}, \text{Virginica}\}$). Here x_i denotes the i -th sample (feature vector); when referring to a single feature we use $x_i^{(j)}$. Split Conformal Prediction partitions it into a training set $\mathcal{D}_{\text{train}}$ for fitting a probabilistic classifier f_θ , and a calibration set $\mathcal{D}_{\text{cal}}$ for estimating conformity thresholds.

Definition 7 (Nonconformity Score). Given a trained classifier producing probabilities $P_\theta(y|x)$, the nonconformity score for calibration instance (x_i, y_i) is:

$$s_i = \begin{cases} P_\theta(y_i = 1 | x_i), & y_i = 1, \\ 1 - P_\theta(y_i = 1 | x_i), & y_i = 0. \end{cases} \quad (2.7)$$

Higher values indicate better conformity with the model.

Definition 8 (Conformal Threshold). Let α denote the tolerated miscoverage level (e.g., $\alpha = 0.1$ for 90% confidence). Sorting the calibration scores $\{s_i\}$ in ascending order, the threshold is:

$$q_\alpha = \text{Quantile}_{\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil / n_{\text{cal}}} (\{s_i : (x_i, y_i) \in \mathcal{D}_{\text{cal}}\}). \quad (2.8)$$

Definition 9 (Prediction Set). For a new input x_{new} , the conformal prediction set is:

$$\hat{\Gamma}(x_{\text{new}}) = \{y \in \mathcal{Y} : P_\theta(y | x_{\text{new}}) \geq q_\alpha\}. \quad (2.9)$$

Depending on the predicted probabilities and threshold, the prediction set may contain one label, multiple labels, or (in the extreme) all labels in $\mathcal{Y}$:

$$\hat{\Gamma}(x_{\text{new}}) \subseteq \mathcal{Y}.$$

Singletons reflect confident predictions; the full set $\{0, 1\}$ indicates ambiguity.

Example 2.3.4 (Conformal Prediction on the Iris Dataset). Consider a classifier trained on the *Iris* dataset. For a new flower x_{new} , the model outputs the following estimated class probabilities:

$$P_\theta(\text{Setosa} | x_{\text{new}}) = 0.05, \quad P_\theta(\text{Versicolor} | x_{\text{new}}) = 0.60, \quad P_\theta(\text{Virginica} | x_{\text{new}}) = 0.35.$$

To illustrate the conformal procedure, suppose that the calibration set contains the following nonconformity scores (computed using Definition 7):

$$s = \{0.42, 0.47, 0.51, 0.58, 0.73\}.$$

For a target confidence level of 90% (i.e., $\alpha = 0.1$), the threshold $q_{0.1}$ is obtained using Equation (2.8). Sorting the scores in ascending order and applying the standard split-conformal quantile rule yields:

$$q_{0.1} = s_{(4)} = 0.58,$$

where $s_{(4)}$ denotes the fourth-smallest score.

Applying the inclusion rule of Equation (2.9), we obtain the prediction set

$$\begin{aligned} \hat{\Gamma}(x_{\text{new}}) &= \{ y \in \{\text{Setosa}, \text{Versicolor}, \text{Virginica}\} : P_{\theta}(y \mid x_{\text{new}}) \geq 0.58 \} \\ &= \{\text{Versicolor}, \text{Virginica}\}. \end{aligned}$$

The resulting set prediction reflects the inherent ambiguity between the two species, consistent with their overlapping petal-length distributions described in Example 2.3.1. Conformal Prediction therefore assigns a multi-label set rather than a single species, indicating that the classifier cannot distinguish between *Versicolor* and *Virginica* with high confidence at the chosen coverage level.

In this chapter, CP is applied as a wrapper around logistic regression and the HEUQ ensemble.

2.4 Methodology

This section outlines the methodological framework used in this study. The theoretical foundations of Belief Functions and Conformal Prediction are provided in Section 2.3. Although both methods rely on set-based representations of uncertainty, they operationalise and interpret uncertainty in different ways.

2.4.1 Belief Functions

The framework introduced in Section 2.3.1 was implemented as a feature-wise evidential classifier in which each attribute acts as an independent source of evidence about the true class (Chen et al., 2014). The methodology consists of three key components:

1. Construction of basic belief assignments (BBAs)
2. Combination of feature-wise evidence
3. Derivation of a final set-valued prediction

These three steps correspond directly to the procedures detailed below.

1. Mass Assignment via Histogram Overlap. To construct basic belief assignments (BBAs), each numerical feature was used to estimate empirical class-conditional distributions through histograms, as illustrated in Figure 2.2. For a new observation with feature value x_j , the strength of evidence supporting each class C was quantified by evaluating the class-conditional density $p(x_j | C)$ derived from these histograms. The belief mass assigned to a singleton class is obtained by normalising these densities according to Equation (2.4). In regions where two classes' empirical distributions overlap, a portion of the mass is allocated to the corresponding composite subset using the minimum-overlap expression in Equation (2.5). Any residual uncertainty is assigned to the universal set Θ using Equation (2.6), reflecting lack of decisiveness in the evidence.

This histogram-based formulation is entirely data-driven and empirical. The performance depends on the discriminative strength of the underlying features. When the feature distributions for the different classes overlap heavily or are noisy, the resulting evidence tends to be weak or ambiguous. The implications of such overlap particularly for the proprietary churn dataset are discussed in more details the Results (Section 2.6).

2. Combining Evidence Across Features. Once feature-wise BBAs are constructed, they are fused using Dempster's Rule of Combination (Definition 6). This rule aggregates consistent evidence (subsets with non-empty intersections) while discounting conflicting evidence (subsets with empty intersections). The normalisation term ensures that the combined mass function remains valid by excluding conflict, as formalised in Equation (2.3). In practice, this iterative combination produces a global BBA per data record.

3. Decision Rule. The final prediction of the BF classifier is derived directly from the combined mass function. If the largest mass is assigned to a singleton class, that class is returned as the prediction. If the dominant mass is assigned to a composite subset (e.g., $\{0, 1\}$ in a binary problem), the classifier outputs this subset as an *uncertain prediction*. This decision rule preserves the set-based semantics of the framework, enabling explicit representation of ambiguity rather than forcing a point prediction when the evidence is weak or contradictory.

As noted in Section 2.3.1, the approach is most effective when features exhibit well-separated, low-noise empirical distributions. Its behaviour under weakly discriminative or high-dimensional feature sets is discussed later in the Results (Section 2.6).

2.4.2 Conformal Prediction

Conformal Prediction was implemented as a model-agnostic, post-hoc calibration layer applied to probabilistic classifiers. Unlike BF, which generate set-valued predictions directly from evidential reasoning, CP transforms raw model probabilities into prediction sets that satisfy formal finite-sample coverage guarantees.

The methodology consists of four key components:

1. Data splitting for conformal calibration
2. Construction of probability-based nonconformity scores
3. Computation of the conformal threshold
4. Construction of prediction sets

These steps correspond directly to the procedures detailed below.

1. Split Conformal Classification. The methodology follows the split conformal classification procedure introduced in Section 2.3.2. The available dataset is partitioned into two disjoint subsets: a training set for fitting the underlying classifier and a calibration set for computing nonconformity scores. No model-specific retraining or modification is required, reinforcing the model-agnostic nature of CP.

2. Probability-Based Nonconformity Scores. For each calibration example (x_i, y_i) , a nonconformity score is computed based on the model’s class probability, following Equation 2.7. This score quantifies how well the model’s prediction aligns with the observed label: higher scores indicate conformity, while lower scores indicate atypicality. Only this probability-based nonconformity function is considered in this work.

3. Threshold Computation. After computing the calibration scores, the significance level α is used to derive a threshold via the empirical quantile rule in Equation (2.8). Under the exchangeability assumption, this threshold ensures that prediction sets constructed for future data achieve the desired coverage level $(1 - \alpha)$ in finite samples.

4. Prediction Set Construction. For a new observation x_{new} , the classifier’s probability outputs $P_\theta(y \mid x_{\text{new}})$ are compared to the threshold obtained above. The final prediction set is given by the inclusion rule in Equation (2.9). In binary classification tasks, the resulting prediction can be a confident singleton (e.g., $\{0\}$ or $\{1\}$) or an uncertain set (e.g., $\{0, 1\}$) when the model is not confident relative to the calibration threshold.

2.5 Data

Proprietary Data Versions During the course of this research, two versions of the proprietary datasets were made available by our industrial partners, each reflecting a different stage of the company’s internal data infrastructure.

Version 1 was the dataset initially provided by the Marketing team at the start of the collaboration. This version was used for early experiments, including the evaluation of BF, since it was the only data accessible at that time.

Approximately eighteen months later, as our industrial partners established a dedicated AI team, better curated **Version 2** of the dataset was provided. This second version is a cleaner version of the first one contained improved feature engineering, updated customer records, and more reliable target annotations.

Because BF was evaluated early in the project and showed limited performance, all subsequent methodological developments, such as the HEUQ model introduced in Chapter 3, the application of CP, and the reinforcement-learning experiments, were conducted using Version 2. This chronological progression explains why different methodological components in this thesis rely on different dataset versions: each experiment was performed on the highest-quality data available at the time it was undertaken.

2.5.1 Proprietary churn dataset (Version 1)

TABLE 2.1: Key feature distributions in the proprietary churn dataset version 1

Feature	Description	Most frequent values / distribution
Target (Churn)	Binary churn indicator	0 > 95%, 1 < 5%
Policyholder age	Age of the subscriber	Range 18–70; highest frequencies at ages 48–60 (~2.5–3.0% each).
Product type	Main policy category	AUTO, Non-AUTO
Vehicle category	Vehicle type code	110 (95%), 120 (4.7%), others <1%
Sex	Customer gender	Gender 1 = 47%, Gender 2 = 53%
Language	Language	Language 1 = 61%, Language 2 = 39%
Life stage	Customer segment	SEN = 26%, SIN = 19%, MED = 14%, ENS = 13%, others = 28%
Has children	Child indicator	Yes = 38%, No = 10%, Missing = 52%
Profession	Professional group	Cat. 4 = 25%, 7 = 16%, 3 = 12%, others <10%
Payment fractioning	Installment vs. full payment	0 = 57%, 1 = 43%
Domiciliation	Automatic debit usage	0 = 99.9%, 1 < 0.1%

The Version 1 dataset used in Section 2.6.1 originates from the internal marketing portfolio of our industrial partners and was developed for churn prediction. Each record corresponds to a customer active in March 2022, labeled as either *churned* (1)

or *retained* (0) in the subsequent month. Features were constructed over a 24-month observation window (January 2020–January 2022), followed by a one-month waiting period (February 2022) to prevent data leakage i.e. customers who submitted request to change provider in or by the end of January. The prediction target was defined for March 2022. The Table 2.1 summarizes some of the features of the proprietary data, due to confidentiality agreements, we can only share a limited variables and their descriptions and challenges with this specific dataset.

The raw dataset contained more than 250,000 records with more than 300 variables spanning demographic, behavioural, and product attributes. Because slightly more than 2,500 churners were initially observed, churn cases from two consecutive months (March–April 2022) were combined, yielding 7,523 labeled instances. After preprocessing and feature selection, 65 predictors were retained. Categorical attributes (e.g., sex, language, life stage, policy type) were one-hot encoded, continuous variables standardized, and missing values imputed using mean or mode (see Table 2.1).

Several data-related challenges were observed in Version 1 of the dataset. First, user identifiers were not consistent across consecutive months, which caused partial overlap and missing connections between customer records. Second, some customers marked as churned in March 2022 were still active in April, indicating label noise and possible misclassification. Third, the dataset showed extreme **class imbalance**, with a churn ratio of less than 5%, making model training and evaluation difficult. Finally, many variables were incomplete or weakly informative, resulting in **fragmentary features**, meaning features that were incomplete, sparsely populated, or inconsistently recorded, reducing predictive reliability.

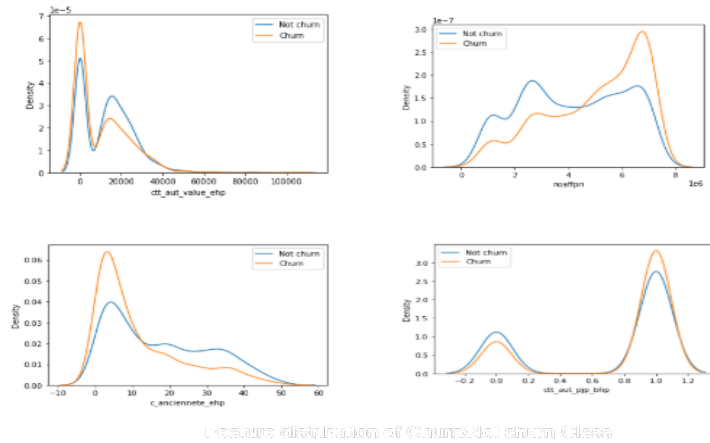


FIGURE 2.4: Feature Distribution for Classes Churn/Not Churn. The complete overlap across all features indicates a severe lack of separability between the classes. This absence of clear, distinctive evidence prevents the belief function (mass function) from assigning mass with certainty, leading to high ambiguity and an inability to accurately classify data points.

Belief Functions limitations: The proprietary churn data presented several limitations that affected the reliability of downstream modelling via BF. The dataset exhibited strong class imbalance, missing or inconsistent churn labels, and irregularities across monthly updates.

Because the features offered limited separation between churners and non-churners, the corresponding mass functions tended to assign most belief to broad or non-specific sets rather than clearly supporting one class (Figure 2.4). Because of these issues, only a fraction of the records could be classified with confidence, and the available evidence for each customer was often weak, noisy, or contradictory (Section 2.6).

2.5.2 Proprietary churn dataset (Version 2)

For the empirical results presented in the Section 2.6.2, and Chapter 3, 4, we used an anonymized policy-level dataset provided by our industrial partners. The dataset contains more than 250,000 observations, each representing a specific insurance policy held by a customer (e.g., motor, fire, assistance, travel). This policy-level granularity is a deliberate design choice by our industrial partner, as it enables identifying not only customers likely to churn a given contract but also those who may subsequently cancel other policies in their portfolio. The raw data was pre-processed internally by our industrial partners using domain-specific procedures handling missing values, correcting invalid entries, and mitigating outliers before being made available for modelling.

The final modelling frame consists of more than 50 cleaned and engineered predictors capturing demographic characteristics (e.g., age, household size, occupation, region), vehicle attributes (e.g., category, age, power, ownership type), insurance conditions (bonus–malus, discounts, coverage options), behavioural signals (tenure, interactions, replacement patterns), the breadth of each customer’s non-life portfolio, and recent claims experience. Categorical attributes were label-encoded for modelling, while continuous variables such as age, vehicle age, and claim amounts were retained in raw or standardized form depending on the experimental setting. The single dependent variable indicates whether the policy churned within the prediction window. Table 2.2 provides a snapshot of representative features and their empirical distributions.

All experiments on the proprietary churn dataset (Version 2) use stratified splits of 70% for training, 15% for validation, and 15% for testing, to preserve the original class proportions. For the conformal prediction experiments (Section 2.6.2), we additionally require a calibration set; we construct it by splitting the validation set into validation and calibration subsets.

TABLE 2.2: Key feature distributions in the proprietary dataset version 2

Feature	Description	Most frequent values / distribution
Policyholder age	Age of the subscriber	Range 18–70; highest frequencies at ages 48–60 (~ 2.5 – 3.0% each).
Household size	Number of persons in the household	1 = 75.6%, 2 = 22.0%, ≥ 3 = 2.4%
Customer tenure	Years since first subscription	0–5 = 35.8%, 6–10 = 21.2%, 11–20 = 28.4%, ≥ 21 = 14.6%
Sex	Policyholder gender	Female = 45.5%, Male = 54.5%
Region	Residential region (encoded)	Region 0 = 10%, Region 1 = 40%, Region 2 = 50%
Occupation	Professional category (encoded)	Cat. 2 = 25%, Cat. 4 = 24%, Cat. 11 = 9.6%, others $< 10\%$ each
Vehicle category	Type of insured vehicle	Cat. 1 = 92.5%, Cat. 2 = 7.1%, others $< 1\%$
Second-hand vehicle	Vehicle new vs. used	Used = 49.9%, New = 7.1%, Unknown = 43.1%
Bonus–Malus	Bonus–Malus coefficient	0 = 67.6%, 1–4 = 11.6%, ≥ 5 = 20.8%
Auto portfolio size	Number of auto policies per policyholder	1 = 87.5%, 2 = 11.4%, ≥ 3 = 1.1%
Total portfolio size	Number of non-life contracts	1 = 33%, 2 = 23%, 3 = 18%, 4 = 13%, ≥ 5 = 13%
Claims last year	Number of claims last year	0 = 88.9%, 1 = 10.9%, ≥ 2 $< 0.3\%$

2.5.3 Open credit card fraud dataset

A second dataset was included to strengthen the analysis, assess, and illustrate the CP method in a cleaner and more controlled setting suitable for publication. This dataset, known as the Credit Card Fraud Detection dataset, was released through a collaboration between Worldline and the Machine Learning Group of the Université Libre de Bruxelles (Dal Pozzolo et al., 2015). It consists of credit card transactions carried out by European cardholders over a two-day period in September 2013 and is widely used as a benchmark for fraud detection.

The dataset contains **284,807** transactions, among which only **492** are fraudulent, resulting in an extreme class imbalance of **0.172%**. All input variables are numerical: 28 PCA-transformed components (V1–V28) and two untransformed variables, Time and Amount. The Class variable serves as the target label, taking value 1 for fraud and 0 otherwise. Because of confidentiality constraints, the original raw features are not available, the PCA representation therefore offers a compact and denoised feature space suitable for statistical calibration and uncertainty quantification.

This dataset was employed specifically for the CP experiments presented in this chapter. For the experiments, the dataset was split into **train** (182,276 records), **calibration** (45,569 records), and **test** (56,962 records) sets. A total of 30 independent

variables were available, but to maintain dimensional consistency with our uncertainty framework, we restricted the analysis to the top 10 PCA components, which capture most of the variance in the data and reduce the influence of noise.

The corresponding CP results were presented at a scientific conference, illustrating the applicability in highly imbalanced fraud detection scenarios.

2.6 Results

Before presenting the results, it is important to note that the exact number of churn and non-churn customers in the dataset cannot be disclosed due to confidentiality agreement with our industrial partners. To comply with this agreement while still enabling meaningful interpretation, all result tables report class proportions or rounded values rather than precise customer counts. Whenever absolute numbers are shown, they are ceiled or floored to prevent the disclosure of sensitive operational volumes. This approach preserves the structure of the analysis while ensuring full compliance with confidentiality requirements.

2.6.1 Belief Functions

The BF classifier exhibited substantial performance limitations on the churn dataset version 1 (Section 2.5.1). The dataset was split in 70-30% stratified split, where 70% was employed for training and 30% for testing purposes. These limitations are directly tied to the data quality issues identified in Section 2.5.1, which significantly weakened the evidential support required for reliable mass assignment in BF.

TABLE 2.3: Prediction vs Actual Percentages (Belief Function Classifier) Proprietary dataset version 1

	Class 0	Class 1	Class 0 or 1
Predicted correctly	38.69%	12.04%	–
Misclassified	6.07%	6.31%	–
Uncertain (0 or 1)	–	–	36.85%
Actual class population	51.63%	48.37%	–

Table 2.3 summarises the predictive outcomes. Less than 3,000 non-churners and 900 churners were correctly classified, representing 50.73% of all customers. Conversely, 12.38% samples were misclassified, and a notably large subset i.e. 36.85% of customers were found uncertain {Class 0, Class 1}. This high abstention rate indicates that, for a substantial portion of the population, the available features did not provide sufficient evidential strength to favour one class over another.

A closer examination of the misclassified samples further highlights the instability of the approach. Approximately half of the errors stem from *Class 1 samples being misclassified as Class 0* (6.07% cases), while the remainder correspond to *Class 0*

samples misclassified as Class 1 (6.31% cases). This symmetry reflects the fact that the evidence in the feature space was often weak or contradictory.

These empirical results confirm the dataset challenges discussed earlier. A major limitation stemmed from the fact that features were showed little separability between churners and non-churners. Because the feature distributions largely overlapped across classes, the belief functions rarely received strong evidence in favour of a specific label. Consequently, the model often assigned mass to uncertain sets ($\{0, 1\}$), resulting in a high proportion of uncertain predictions. The lack of discriminative features thus prevented BF from forming confident decisions and significantly reduced its predictive effectiveness on this dataset.

Overall, these findings demonstrate that BF struggled fundamentally on this dataset, not because of the theory itself but because the underlying data did not satisfy the evidential requirements needed for BF modelling. In settings where features are sparse, noisy, or weakly discriminative, the BF framework tends to predict uncertain sets. This substantially limits its usability for churn prediction in real-world, operational contexts such as those encountered at our industrial partners.

2.6.2 Conformal Prediction

Conformal Prediction was applied to Version 2 and Credit card fraud detection datasets introduced in Section 2.5.2 and Section 2.5.3.

Proprietary churn dataset (Version 2)

As a model-agnostic framework, CP was applied on top of two underlying classifiers: an ensemble, namely HEUQ model introduced in Chapter 3 and a logistic regression (LR) baseline.

Two significance levels were considered, $\alpha = 0.05$ and $\alpha = 0.02$, corresponding to target coverage rates of 95% and 98%, respectively. For calibration, approximately 41,000 samples were used. The test sets contained about 25,000 customers for the $\alpha = 0.05$ configuration and about 25,000 customers for the $\alpha = 0.02$ configuration, with the remaining records used for training. Across both splits, churn represented less than 10% of the customer population, indicating a strong and persistent class imbalance.

TABLE 2.4: Conformal Prediction results for LR and HEUQ at $\alpha = 0.05$ and $\alpha = 0.02$ (percentages of actual class)

Model	Significance level(α)	Class	{Not churn}	{Churn}	{Not churn, Churn}	Observed (%)	Expected (%)
LR	0.05	Not churn	81.1%	2.1%	16.8%	97.8	95.0
		Churn	24.6%	0.0%	75.4%	75.4	95.0
LR	0.02	Not churn	50.1%	1.5%	48.4%	98.4	98.0
		Churn	23.5%	0.0%	76.5%	76.5	98.0
HEUQ	0.05	Not churn	78.3%	2.4%	19.2%	97.6	95.0
		Churn	33.6%	0.0%	66.4%	66.4	95.0
HEUQ	0.02	Not churn	50.1%	1.5%	48.4%	98.5	98.0
		Churn	23.5%	0.0%	76.5%	76.5	98.0

Table 2.4 summarises the Conformal Prediction outcomes for both LR and HEUQ under $\alpha = 0.05$ and $\alpha = 0.02$. Across all configurations, the models achieve coverage close to or above the expected rates for the *non-churn* class, with observed values ranging between 97.6% and 98.5%. This is largely driven by the strong class imbalance, where most customers belong to the majority (non-churn) class and therefore receive either singleton predictions or broad set predictions containing the true label. However, for the *churn* class, coverage falls substantially below the nominal levels, with observed values between 66% and 76%, far from the required 95-98%.

In practice, the vast majority of churn cases are assigned to the uncertainty set $\{0, 1\}$, indicating that the models rarely produce confident singleton predictions for this minority class. These violations reflect the difficulty of constructing reliable prediction sets for the minority class. CP over HEUQ exhibits similar behaviour to LR at $\alpha = 0.02$, whereas at $\alpha = 0.05$ it produces fewer set-valued predictions for non-churn cases but still fails to achieve valid coverage for churners. Overall, while CP provides calibrated guarantees for the dominant class, it struggles to meet its theoretical coverage bounds for churn prediction, highlighting the intrinsic challenges posed by the data distribution and class imbalance.

Open credit card fraud dataset

TABLE 2.5: Conformal Prediction results for Logistic Regression on the Credit Card Fraud dataset

Significance level(α)	Class	{Not fraud}	{Fraud}	{Not fraud, Fraud}	Observed (%)	Expected (%)
0.01	Not Fraud	28218	521	28108	99.1	99.0
	Fraud	2	104	9	98.3	99.0
0.02	Not Fraud	38148	1081	17618	98.1	98.0
	Fraud	4	106	5	96.5	98.0
0.05	Not Fraud	47010	2780	7057	95.1	95.0
	Fraud	5	108	2	95.7	95.0

Table 2.5 reports the Conformal Prediction (CP) performance of the logistic regression classifier on the Credit Card Fraud dataset. Across all significance levels, CP achieves coverage close to the nominal targets for both the non-fraud and fraud classes. For $\alpha = 0.01$ and $\alpha = 0.02$, the observed coverage for non-fraud transactions is 99.1% and 98.1%, respectively, matching the expected theoretical guarantees. The fraud class also exhibits high empirical coverage (98.3% and 96.5%), indicating that CP reliably includes the true label within the prediction set even for the minority class. At the more permissive level $\alpha = 0.05$, observed coverage remains aligned with expectations, with 95.1% for non-fraud and 95.7% for fraud. These results demonstrate that, on this dataset, CP provides valid and well-calibrated uncertainty quantification for fraud detection, achieving its theoretical coverage bounds across both majority and minority classes.

2.7 Discussion

Although BF and CP adopt fundamentally different approaches to uncertainty quantification, their behaviour on the churn dataset reveals a strikingly similar pattern. BF struggled because the underlying features were weakly discriminative, causing most basic belief assignments to remain diffuse and leading to frequent allocation of mass to composite sets such as $\{0, 1\}$. CP, in contrast, does not rely on evidential structures; instead, it calibrates prediction sets from model residuals or nonconformity scores. Yet, despite these methodological differences, both frameworks exhibit the same under performance for the minority churn class. Neither is able to provide confident, class-specific predictions. BF predicts a broad belief sets, while CP produces prediction sets that include both labels. In contrast, when CP was applied to the Credit Card Fraud dataset, it achieved coverage almost perfectly aligned with its theoretical guarantees, highlighting that CP performs reliably when the underlying features provide stronger class separability.

This behaviour is driven by both statistical and algorithmic factors. Statistically, churn is an extreme minority class with substantial feature overlap with non-churn customers. As a result, the classifier used inside CP assigns nonconformity scores to churn examples that are often indistinguishable from those of non-churn cases. Algorithmically, CP includes all labels whose nonconformity scores fall below the $(1 - \alpha)$ quantile of the calibration distribution. For churn cases whose scores frequently lie near or above this threshold, CP must include both labels in the prediction set, yielding $\{0, 1\}$. Consequently, the vast majority of churn customers end up in the uncertainty set, mirroring the behaviour observed under BF.

Taken together, these results indicate that the primary limitation is not the uncertainty framework itself, but rather the weak signal and severe class imbalance of the dataset. When features offer little discriminative evidence, both BF and CP despite their theoretical differences converge toward highly uncertain predictions for churn, exposing structural constraints of the data that neither method can overcome.

Operational implications on the churn decision workflow. In the churn prediction workflow of our industrial partner, uncertainty estimates are most valuable when they support *actionable* choices, such as selecting customers for retention campaigns under limited budgets and handling borderline cases more conservatively. In practice, the set-valued outputs of BF and CP can be useful as a safety layer: when the methods return a single label, they provide a clear recommendation, while ambiguous cases can be explicitly flagged for cautious handling. This is attractive in an operational churn workflow, because it supports conservative policies such as deferring action, requesting additional evidence, or routing selected customers to human review.

At the same time, large-scale prioritization typically requires a calibrated continuum of risk scores that supports fine-grained ranking and stable operating thresholds over time. Because BF and CP are not designed to directly provide such calibrated risk scores and because ambiguity can remain common under class imbalance and limited separability; their outputs may be less convenient when decisions must be made at scale. This motivates complementing set-based uncertainty with *model-based* uncertainty quantification, where uncertainty is represented through predictive distributions that can support both ranking and triage.

From set-based UQ to model-based UQ. The results on the proprietary churn setting provided by our industrial partner are therefore primarily *diagnostic*: they suggest that practical performance depends not only on the uncertainty formalism, but also on data characteristics such as imbalance and weak separability. Accordingly, Chapter 3 investigates heterogeneous ensemble learning on the proprietary dataset version 2 (Section 2.5.2) as a way to obtain sharper and better calibrated uncertainty estimates while maintaining competitive predictive performance.

Chapter 3

Heterogeneous Ensemble framework for Uncertainty Quantification (HEUQ)

This chapter introduces HEUQ, a heterogeneous ensemble framework designed to quantify uncertainty by leveraging diversity across multiple learning paradigms. Using proprietary and public churn datasets, we demonstrate that HEUQ provides more reliable uncertainty estimates and improved predictive performance compared to homogeneous ensemble methods like Deep ensemble. A detailed paper on this work is currently in preparation. Portions of the results were previously presented at ICAEB (Singh et al., 2025c), summarized in a white paper (available on request), and shared in a poster session (Singh et al., 2025a).

Contents

3.1 Introduction	36
3.2 Literature review	39
3.2.1 Uncertainty Estimation in Machine Learning	39
3.2.2 Uncertainty in Machine Learning for Operations Research	41
3.2.3 Use case: Churn prediction	41
3.2.4 Use case: Customer conversion	42
3.2.5 Literature review findings	43
3.3 Methodology	43
3.3.1 Uncertainty in machine learning	43
3.3.2 Uncertainty in HEUQ: a <i>Basic Composition Rule (BCR)</i> approach	45
3.3.3 Heterogeneous Ensemble for Uncertainty Prediction: HEUQ	49
3.4 Data	50
3.4.1 Proprietary churn dataset (Version 2)	52
3.4.2 Proprietary Customer conversion dataset description	54
3.5 Experimental Setup	55
3.5.1 Performance measure	56

3.5.2 Training and implementation details	57
3.6 Results	59
3.6.1 Experiments: Churn Prediction	59
3.6.2 Experiments: Uncertainty quantification churn prediction	61
3.6.3 Ablation study	62
3.6.4 Customer conversion (Case study)	65
3.7 Discussion and Conclusion	66

3.1 Introduction

In Chapter 2, we first examined uncertainty quantification methods commonly employed in actuarial contexts, namely Belief Functions (BF) and Conformal Prediction (CP). While these approaches offer set-based representations of uncertainty, our experiments showed that their predictive performance on real-world insurance data was limited. This motivates a shift toward machine learning-based methods for Uncertainty Estimation (UE). In particular, we ask whether the limitations observed on our industrial partner’s churn setting stem from the UQ (Chapter 2) or from the predictive models and their inductive biases. To answer this, Chapter 3 evaluates heterogeneous ensembles on the dataset (Version 2.5.2), targeting improved calibration and more informative epistemic uncertainty without sacrificing predictive accuracy.

Recent advances in machine learning have produced highly capable predictive models that increasingly support decisions in business, public administration, healthcare, and other high-stakes domains. As these systems gain influence, concerns regarding the reliability and the trustworthiness of their predictions have grown. Most predictive models provide only point estimates without indicating the degree of uncertainty associated with them. When data are noisy, imbalanced, limited, or subject to distributional shifts, relying solely point estimates alone provide an incomplete picture of model behaviour.

As introduced in Chapter 1, Uncertainty Quantification offers a principled way to address this challenge by complementing model predictions with measures of confidence. Prior research has explored probabilistic and ensemble-based techniques for representing uncertainty (Lakshminarayanan, Pritzel, and Blundell, 2017; Shaker and Hüllermeier, 2020). Homogeneous ensembles, composed of models of the same type, have been widely used to estimate total, aleatoric, and epistemic uncertainty (Lakshminarayanan, Pritzel, and Blundell, 2017; Depeweg et al., 2018; Shaker and Hüllermeier, 2020). However, such approaches fail to exploit the complementary strengths of different model families.

Despite their popularity, research on uncertainty estimation has focused predominantly on these homogeneous ensembles, even though heterogeneous ensembles, composed of diverse learning paradigms, are known to improve predictive performance in many settings, including churn prediction (Bogaert and Delaere, 2023). A

gap remains in the literature regarding how model heterogeneity affects the estimation and decomposition of uncertainty. Specifically, it is unclear whether diversity among ensemble members reduces epistemic uncertainty by strengthening consensus, or instead increases it by introducing multiple, potentially conflicting hypotheses about the data-generating process.

Our working premise is that heterogeneity can lead to more informative uncertainty estimates. By covering a broader hypothesis space, heterogeneous ensembles capture multiple plausible mappings from features to outcomes, thereby offering richer insight into where the model is confident and where uncertainty is warranted. Moreover, little is known about how common data transformations (e.g., dimensionality reduction) or modifications in ensemble configuration (e.g., adding or removing specific models) affect the uncertainty estimates produced by such ensembles.

To address these gaps, this chapter introduces the *Heterogeneous Ensemble for Uncertainty Quantification* (HEUQ), a general and model-agnostic framework for estimating predictive uncertainty. Central to our approach is the idea that a heterogeneous ensemble, drawing from a diverse set of learning paradigms, can capture a wider range of inherent variability and complex relationships within the data itself, thus potentially improving uncertainty estimates. Individual models within a heterogeneous ensemble, unlike their homogeneous counterparts, will each learn diverse hypotheses and make independent errors, resulting in reduced overall error. Subsequently, heterogeneous ensembles are likely to achieve improved performance (Petrakova, Affenzeller, and Merkurjeva, 2015).

Heterogeneous Ensemble for Uncertainty Quantification (HEUQ) framework, which extends the ensemble paradigm by combining diverse model types. HEUQ aims to improve predictive performance while producing well-calibrated and informative uncertainty estimates suitable for real-world applications.

Building on the decomposition introduced by Depeweg et al. (2018), we quantify total uncertainty via the Shannon entropy of the aggregated predictive distribution, epistemic uncertainty as the divergence between individual model predictions and the ensemble prediction, and aleatoric uncertainty as the residual component inherent to the data. This formulation provides fine-grained insight into both data-driven variability and model-driven disagreement.

Heterogeneous ensembles are particularly well suited to the context of churn prediction and customer acquisition in the insurance industry, where decisions have substantial financial implications and where uncertainty plays a critical operational role. Low epistemic uncertainty signals high consensus among models and supports confident automated decisions, while high epistemic uncertainty highlights cases where human expertise or additional information may be required.

Investigating uncertainty in this context is particularly promising. Although numerous machine learning models have been proposed for churn prediction and

customer conversion, the question of how confident these models are in their predictions has received far less attention. This gap is scientifically important, as uncertainty estimates provide insight into whether model outputs should be trusted, especially for high-stakes decisions. As highlighted in Chapter 1, understanding the certainty of predictions is crucial for operational decision-making in insurance settings, where actions such as customer retention or targeted acquisition depend not only on the predicted outcome but also on the model’s confidence in that outcome. Motivated by these considerations, this chapter examines the following research questions:

- **RQ1:** Do heterogeneous ensembles provide more reliable and better-calibrated UE and predictions than homogeneous ensembles traditionally used for UE?
- **RQ2:** How do common data transformation techniques, such as PCA and Random Gaussian projections, affect UE in ensemble predictions?
- **RQ3:** In research with ensemble methods, ablation studies are often performed to investigate the contribution of different models and model combinations on the performance. However, the question as to whether adding or removing models in an ensemble impacts UE has yet to be addressed.

To address these questions, we ground our investigation in the context of churn prediction and customer conversion within the insurance domain, leading to the following contributions:

- **A general framework for uncertainty estimation in heterogeneous ensembles.** We introduce HEUQ, a model-agnostic ensemble framework that extends existing uncertainty decomposition methods (Shaker and Hüllermeier, 2020; Depeweg et al., 2018) to heterogeneous settings. HEUQ quantifies total uncertainty using Shannon entropy and epistemic uncertainty through model divergence, allowing aleatoric uncertainty to be obtained as the residual component.
- **Empirical evidence that heterogeneous ensembles improve reliability and calibration.** Across three churn datasets (one proprietary insurance dataset and two public benchmarks) and a proprietary conversion dataset, HEUQ achieves competitive or superior balanced accuracy, Negative Log-Likelihood (NLL), and Brier Score (BS) compared to strong baselines including Logistic Regression, Random Forests, CatBoost, Neural Networks, and Deep Ensembles.
- **Robustness under data transformations.** We show that HEUQ maintains stable predictive performance and uncertainty behaviour under common transformations such as Principal Component Analysis (PCA) and Random Gaussian (RG) projections, an essential property in real-world modelling pipelines.

- **Insights from ablation studies on model diversity.** Systematic removal and substitution of learners reveal that individual model families contribute differently to uncertainty components. Neural networks tend to reduce aleatoric uncertainty, while tree-based learners reduce epistemic uncertainty through high consensus.
- **Relationship between uncertainty and performance.** We demonstrate a negative correlation between predictive uncertainty and balanced accuracy: higher uncertainty reliably indicates lower accuracy. This relationship, although intuitive, has not been explicitly documented in prior research.
- **Contributions to customer conversion modelling.** Using a second proprietary dataset, we benchmark predictive performance across physical and digital sales channels, an underexplored setting in the insurance domain. We identify CatBoost and Neural Networks as the strongest learners and demonstrate how uncertainty estimates help assess the reliability of channel-specific models.

The remainder of this chapter is organized as follows: Section 3.2 reviews existing literature on ensemble learning and uncertainty quantification. Section 3.3 presents the theoretical background, uncertainty formulation, and the proposed HE-UQ framework, along with the datasets and experimental setup. Section 3.6 reports the empirical results, including robustness analyses, ablation studies, and managerial implications. Section 3.7 concludes with key findings and directions for future research.

3.2 Literature review

This section discusses related work on uncertainty estimation in machine learning, in operations research. Finally, we highlight the gaps in the literature that our study aims to address.

3.2.1 Uncertainty Estimation in Machine Learning

As discussed previously, ML literature distinguishes between two main types of uncertainty: *aleatoric* and *epistemic* (Hüllermeier and Waegeman, 2021; Shaker and Höllermeier, 2020). Aleatoric uncertainty is inherent noise in the data-generating process and is typically irreducible, even with more data. (Hüllermeier and Waegeman, 2021) Epistemic uncertainty reflects limited knowledge about the model and its parameters, for example due to an imperfectly learned hypothesis or an inadequate model class. (Abdar et al., 2021; Höllermeier and Waegeman, 2021; Shaker and Höllermeier, 2020; Hoffmann, Fortmeier, and Elster, 2021; Kläs and Vollmer, 2018)

Recently, the estimation of uncertainty has started to play an increasingly prominent role in the evaluation of models' predictions (Klås and Vollmer, 2018; Hüllermeier and Waegeman, 2021; Wang, Lawrence, and Niepert, 2020; Nicora et al., 2022). Uncertainty estimation (UE) provides valuable information on the confidence of a model in its predictions, enabling a better model calibration (Hüllermeier and Waegeman, 2021; Abdar et al., 2021). Calibration in machine learning indicates if a trained model's predicted probabilities agree with the performance, indicating if the model reflects the underlying distribution of data. In addition, as shown by Nicora et al. (2022) models that are trained to explicitly reduce uncertainty (measured by metrics such as negative log-likelihood) are less prone to overfitting. The use of uncertainty measures, such as negative log-likelihood during the training process, results in a trained model that is more robust to data variations and less prone to overfitting (Nicora et al., 2022). Furthermore, as mentioned in the Introduction, information on the degree of uncertainty present in a model's predictions is essential for effective decision-making (Nicora et al., 2022; Hüllermeier and Waegeman, 2021; Abdar et al., 2021). For example, in churn management, if a model predicts customer churn with low uncertainty, marketers can confidently devise higher-value retention strategies. Conversely, high epistemic uncertainty indicates disagreement among learners, prompting marketers to be more cautious.

Several approaches have been developed to represent and formalize uncertainty. Techniques such as imprecise probability Walley (1991), belief functions, and evidence theory Shafer (1976) and Smets and Kennes (1994) build on and expand probability theory to offer more flexible frameworks for handling uncertain or incomplete information.

In our study, we focus on Machine Learning based approaches to UQ. Wang et al. (2019) study aleatoric and epistemic uncertainties for Convolutional Neural Network (CNN)-based 2D and 3D medical image segmentation tasks at both pixel level and structure level. Depeweg et al. (2018) propose a Bayesian deep learning framework the decomposition of aleatoric and epistemic components, and their incorporation in decision-making. The work of Gal and Ghahramani (2016) show how uncertainty in deep learning can be efficiently approximated via dropout mechanisms. In particular, they show that a NN with dropout before every weight layer is mathematically equivalent to an approximation of probabilistic deep Gaussian process. Lakshminarayanan, Pritzel, and Blundell (2017) introduced a simpler alternative to the Bayesian method with deep (learning) ensembles. Lakshminarayanan, Pritzel, and Blundell (2017) proposed to retrain the same network many times with different initializations (e.g. different weights).

Caldeira and Nord (2020) study uncertainty estimation using deep ensembles, Bayesian neural networks and concrete dropout. They estimate epistemic uncertainty as the spread of the prediction between different models. Aleatoric uncertainty is estimated by fitting the mean and standard deviation of a normal distribution to maximize the log-likelihood of the data (Caldeira and Nord, 2020). Hoffmann, Fortmeier, and Elster (2021) extend UQ to deep learning ensembles for optical form measurements. The authors define epistemic ensemble uncertainty as standard deviation over prediction. Gustafsson, Danelljan, and Schon (2020) compare Bayesian neural network ensembles and Monte-Carlo dropout for multiple computer vision tasks including image segmentation. The authors' comparison demonstrates that ensembling consistently provide more reliable and practically useful uncertainty estimates Gustafsson, Danelljan, and Schon (2020). The study of Shaker and Hüllermeier (2020) extend Depeweg et al. (2018)'s regression-based analysis to quantify aleatoric and epistemic uncertainty to a classification problem statement. They employ a homogeneous ensemble of random forests, and study aleatoric and epistemic uncertainty. We focus on heterogeneous ensembles and uncertainty estimation. Our study can therefore be considered as an extension of the work of Shaker and Hüllermeier (2020) who considered homogeneous ones. For a full review, we refer the reader to Abdar et al. (2021) and Hüllermeier and Waegeman (2021).

3.2.2 Uncertainty in Machine Learning for Operations Research

Machine learning for uncertainty estimation in operations research is an emerging field and can find applications across various domains like routing, credit scoring, demand prediction, churn prediction, and outlier detection. However, much of the existing research has prioritized deep learning approaches given their black box nature (Thuy and Benoit, 2023; Ning and You, 2019; Garcin and Stéphan, 2021; Watson-Daniels et al., 2024; Feng et al., 2022). In the available work, however, uncertainty estimates are very specific to the proposed model or are merely monitored as additional metric (Thuy and Benoit, 2023). Distinct from the previously proposed studies we propose a model-agnostic framework that takes into consideration human experts based on estimated epistemic uncertainty.

Moreover, there is a lack of literature on the impact of the data transformation on uncertainty estimates with applications in Operations Research. To add to that the uncertainty estimates in the previous studies are always evaluated on small benchmark datasets e.g. MNIST (60,000 observations), and edX platform MOOC (2000 observations).

3.2.3 Use case: Churn prediction

A complete literature survey on churn prediction is beyond the scope of our work. We provide a concise overview of churn literature, as it serves as an application of

our proposed framework for quantifying uncertainty employing heterogeneous ensembles. Popular methods in churn include Logistic Regression (LR) (Jain, Yadav, and Manoov, 2020; Buckinx and Poel, 2005; De Bock and Poel, 2012), Artificial Neural Networks (NN) (Khan et al., 2019; Tsai and Lu, 2009), Decision Tree (DT) (Ahmad, Jafar, and Aljoumaa, 2019; Mauritsius et al., 2020), Random Forest (RF) (Ullah et al., 2019; Jain, Yadav, and Manoov, 2020; De Bock and Poel, 2012), Naive Bayes (NB) (Kassem et al., 2020; Mauritsius et al., 2020), Support Vector Machines (SVM) (Jain, Yadav, and Manoov, 2020; Coussement and Poel, 2008), Deep Learning (Zhang et al., 2017). For complete surveys, we refer the interested reader to De and Prabu (2022), Geiler, Affeldt, and Nadif (2022), Ahn et al. (2020). For conciseness and following Bharathi S, Pramod, and Raman (2022), we synthesize the main findings of our literature reviews in Table A.3 in Appendix A.3. Also, following Bogaert and Delaere (2023), we indicate whether the churn prediction methods are ensemble, and if so, whether they are heterogeneous ensembles. We added a new column indicating whether the question of uncertainty in predictions was addressed.

3.2.4 Use case: Customer conversion

Yeo et al. (2020) focus on e-commerce conversion prediction, introducing a joint modelling approach integrating customer and product-level conversion patterns based on the buying decision process. Their study employs XGBoost on a simulated dataset based on real-world data. Wendi, Xiaoling, and Feida (2017) propose a time-aware recommendation model for customer conversion, integrating recommendation relevance and conversion time using survival analysis. Recommendation algorithms often fail to account for the mismatch between a product's actual conversion period and the target conversion period set by applications (Wendi, Xiaoling, and Feida, 2017). Tokuç and Dağ (2024) leverage LightGBM for purchase intent prediction using e-commerce clickstream data, addressing high-dimensionality, class imbalance challenges, non-purchase sessions, and the temporal variability of user behaviour. Martínez et al. (2020) propose a predictive analytics framework for customer behaviour in non-contractual settings, using gradient tree boosting to forecast future purchases. Their model, tested on a large dataset (10,000 customers and 200,000 transactions), achieves 89% accuracy and 0.95 Area Under the Curve (AUC)¹ for predicting next-month purchases. Semih, Emin, and Ahmet (2023) introduce two models designed for real-time customer conversion prediction, aiding competitive cost-per-click bidding strategies in digital marketing campaigns, particularly for Google Ads. Soni (2022) evaluate the effectiveness of temporal, demographic, and behavioural factors in enhancing customer conversion. The authors employed the Decision Tree, Random Forest, XGBoost, and Gradient Boosting machine to evaluate factors influencing e-commerce customer conversion. Their study found that

¹AUC indicates a model's ability to discriminate between different classes. It represents the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance by the model (Bradley, 1997).

behavioural-based personalization most effectively enhances conversion rates. Bag et al. (2022) examine how AI-driven digital engagement increases online sales and how satisfying online shopping experiences influence repurchase intentions. Using data from Indian consumers, their research indicates that AI positively affects user engagement and conversion, leading to satisfying user experiences.

Despite these advances, most studies focus on digital sales, with limited research on conversion prediction in physical sales settings. Additionally, comparative analyses between digital and physical sales channels remain scarce, highlighting the need for further research in this area.

3.2.5 Literature review findings

With regards, to uncertainty estimation, our review shows that it has gained traction with machine learning. This can be attributed to the heightened importance of trustworthiness due to the “black-box” nature of deep learning models. However, most of the studies focus on model-specific uncertainty estimation in operations research. Furthermore, with regard to ensemble methods, most studies to date have focused on homogeneous ensembles (Bogaert and Delaere, 2023).

The success of machine learning hinges on data. Data transformation is an essential step in the impact model performance. However, our extensive literature review reveals a gap in research – the effect of feature transformation on model uncertainty remains unexplored. Thus, with this chapter, we aim to fill these gaps in the literature.

3.3 Methodology

In this section, we elaborate on our methodology to address the gaps in the literature. We first lay out our theoretical formulation of uncertainty (Section 3.3.1, 3.3.2, also in Appendix A.1) for our framework. We present the decomposition of the total uncertainty into its epistemic and aleatoric components. The theoretical grounding is then used to motivate our proposed heterogeneous ensemble, HEUQ (Section 3.3.3). Subsequently, Section 3.4 presents the three different datasets employed in the research, and the data transformation methodologies investigated. Finally, our experimental setup, performance measures, and implementation are described (Section 3.5).

3.3.1 Uncertainty in machine learning

Definition 10 (Aleatoric uncertainty). Following Hüllermeier and Waegeman (2021), **aleatoric uncertainty** is uncertainty inherent to the data-generating process, arising from variability or randomness in the observations. It is typically difficult to reduce, even with additional data.

Definition 11 (Epistemic uncertainty). Following Hüllermeier and Waegeman (2021), **epistemic uncertainty** (model uncertainty) stems from limitations of the learned

model and its parameters. It is, in principle, reducible, for example by collecting more informative data, improving hyperparameters, or using a better model.

Definition 12 (Total uncertainty). **Total uncertainty** is the combination of aleatoric and epistemic uncertainty. In this chapter, we compute total uncertainty using Shannon entropy (Eq. (3.3)) and estimate epistemic uncertainty by quantifying predictive disagreement across ensemble members (Eq. (3.4)).

In binary classification, high epistemic uncertainty signifies strong learner disagreement (e.g., half predict class 0, half predict class 1). This level of disagreement indicates that the strong learners are unsure about the true classification.

Hypothesis space and uncertainty

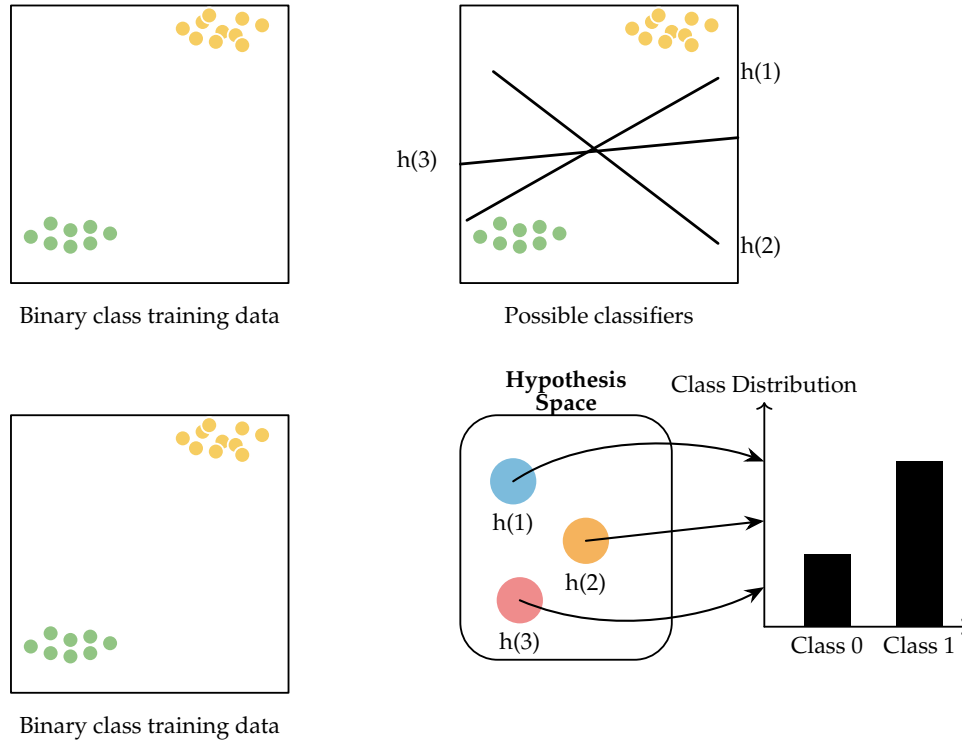


FIGURE 3.1: An example of the binary classification task. Three different optimal hypotheses are represented as three separating lines that divide the data into two classifying regions. Although these hypotheses are learned differently and correspond to distinct points in hypothesis space, they can yield similar class distributions and comparable classification performance.

In machine learning, the hypothesis space is all the possible hypotheses that can be considered. The goal in classification is to learn the best hypothesis, mapping $\mathcal{X} \rightarrow \mathcal{Y}$, from the entire hypothesis space $\mathcal{H}$. For a learner, the best hypothesis, $h^* \in \mathcal{H}$, minimizes the difference between the predicted and actual output labels. That is, it minimizes the loss or uncertainty in its predictions. An illustration is given in Figure 3.1 (Hüllermeier and Waegeman, 2021).

As shown in Figure 3.1 there can be multiple hypotheses learned by different learners. A heterogeneous ensemble combines models from different learning algorithms and different optimization functions. By using different learning algorithms, a heterogeneous ensemble effectively expands the hypothesis space it explores. This diversity is advantageous because different algorithms make different assumptions about the data.

As previously discussed, aleatoric uncertainty is an inherent property of the data and cannot be reduced. However, in real-world scenarios, the actual noise level in the data is often unknown, and the underlying data-generating function is inaccessible. By combining models with different biases, a heterogeneous ensemble can potentially capture a wider range of relationships in the data, leading to improved estimates of aleatoric uncertainty. Our ablation study (Table 3.6) demonstrates this behaviour, where the inclusion of a neural network (NN) in a heterogeneous ensemble results in higher epistemic uncertainty estimate but lower aleatoric uncertainty estimate compared to an ensemble consisting solely of tree-based models. This suggests that the NN captures data patterns that tree-based ensembles may overlook. Combining the strengths of different strong learners through the ensemble allows for a more comprehensive search within the hypothesis space, thus leading to a better model in terms of predictive capability, calibration and lower total uncertainty estimate. Our results attest to lower total uncertainty estimate (Table 3.5). In a homogeneous ensemble of strong learners, we can create diversity with different hyperparameters however the optimization algorithm remains the same. Thus the learned hypotheses lie close to each other (Table 3.5 and Table 3.6).

This is good for potentially lower epistemic uncertainty estimates but not necessarily for total uncertainty estimates.

3.3.2 Uncertainty in HEUQ: a *Basic Composition Rule* (BCR) approach

This section elaborates on our formulations of aleatoric and epistemic uncertainties in machine learning. We interchangeably use *model* and *hypothesis* for clarity. As mentioned earlier in Section 3.3.1, uncertainty can be inherent in the data (due to its finite size in practice) or can result from the choice of the model and its configurations (in terms of hyperparameters and parameters). Hence, all predictions are associated with a certain degree of uncertainty (Hüllermeier and Waegeman, 2021).

As the focal point in our development, we will consider a specific instance (denoted as x). We adopt such a perspective for two reasons: i) it enables us to make our development more concrete, and ii) in various applications, marketing decisions have to be made on a specific set of data points (e.g. portfolio optimization) by involving humans in the loop (Chapter 5). To present our definition of uncertainty, we adopt a heuristic approach based on the following notations/assumptions.

1. As usual, the data set

$$\mathcal{D} := \{(x_1, y_1), \dots, (x_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$$

is split into a training set and a test set:

2. For a hypothesis h , $p(\cdot|h, x)$ denotes the predictive probability distribution on $\mathcal{Y}$ provided by the running of hypothesis h on the training set $\mathcal{D}_{Train}$. So :

$$p(\cdot|h, x) : \mathcal{Y} \longrightarrow [0, 1]$$

and satisfies the classical axioms of probability.

3. We consider a binary classification setting. $\mathcal{Y} = \{y_0, y_1\} = \{0, 1\}$.

Total uncertainty for a given model/hypothesis

Let us consider a specific hypothesis (model) h_0 . A classical way to evaluate the predictive quality h_0 is to compute the “likelihood function” L_0 on the testing set :

$$L_0 = \prod_{i=N_1+1}^{i=N} p(y_0|h_0, x_i)^{1-Y_i} p(y_1|h_0, x_i)^{Y_i}.$$

Taking $l_0 = -\log_2(L_0)$ gives us the standard negative log-likelihood loss function computed on the testing set.

As mentioned previously, reformulating standard negative log-likelihood loss function for a specific profile x . In that equation, $p(\cdot|h_0, x)$ can be computed for any x . However, there are no guarantees that the testing set contains labels y for that specific value of x . A solution to circumvent this problem is to consider the conditional expectation of y given x and hypothesis h_0 , which is : $(p(y_1|h_0, x))$. We then get, for a specific x , an estimation of the loss function at x , based on the training set can be rewritten as :

$$l_0(x) = - \sum_{k=0}^1 p(y_k|h_0, x) \log_2(p(y_k|h_0, x)) \quad (3.1)$$

We see that $l_0(x)$ is the entropy to the predictive probability $p(\cdot|h_0, x)$. We can then define $l_0(x)$ as the *total uncertainty at x according to hypothesis h_0* .

Total uncertainty for an ensemble model

For the heterogeneous ensemble, there is a set $\mathcal{H} = \{h_1, \dots, h_M\}$ of hypotheses. For a given set $\mathcal{H} = \{h_1, \dots, h_M\}$ the *Basic Combination Rule (BCR)* to combine models in the ensemble is given by

$$p(y_k|BCR, x) = \frac{\sum_{j=1}^M p(y_k|h_j, x)}{M} \quad (3.2)$$

By analogy with Equation (3.1), we can then define :

$$l_{BCR}(\mathbf{x}) = - \sum_{k=0}^1 p(y_k|BCR, \mathbf{x}) \log_2(p(y_k|BCR, \mathbf{x}))$$

which is in fact the entropy-related formulation of Hüllermeier and Waegeman (2021) (more precisely their Equation (23)). A more detailed generalization from the BCR to the Convex combination rule (CCR) can be found in the Appendix A.

$$u_t(\mathbf{x}) = l_{BCR}(\mathbf{x}) = - \sum_{k=0}^1 p(y_k|BCR, \mathbf{x}) \log_2(p(y_k|BCR, \mathbf{x})) \quad (3.3)$$

where, $u_t(\mathbf{x})$ is the *total uncertainty at $\mathbf{x}$ according to hypotheses ensemble $\mathcal{H}$* .

Decomposition of total uncertainty for an ensemble of hypotheses

Focusing on a particular profile $\mathbf{x}$, we now formalise aleatoric and epistemic uncertainty in line with Hüllermeier and Waegeman (2021).

Let $u_a(\mathbf{x})$ (resp. $u_e(\mathbf{x})$) denote the aleatoric uncertainty (resp. epistemic uncertainty) for a given profile $\mathbf{x}$.

We now state the properties that these uncertainty measures should satisfy.

1. $u_t(\mathbf{x}) = u_a(\mathbf{x}) + u_e(\mathbf{x})$
2. The aleatoric uncertainty $u_a(\mathbf{x})$ needs to give us an estimation of the irreducible uncertainty, relative to the fact that the ultimate (and unknown) best model h will always provide us an estimation of event/class probability (e.g. $p(y_k|h, \mathbf{x})$), with a minimal “uncertainty²”. Since the best hypothesis is generally unknown, $u_a(\mathbf{x})$ can be a good surrogate to the estimation of the total uncertainty. Indeed, if $u_e(\mathbf{x})$ (i.e. epistemic uncertainty) is close to zero, then $u_a(\mathbf{x}) = u_t(\mathbf{x})$, the total uncertainty for $\mathbf{x}$.
3. The epistemic uncertainty $u_e(\mathbf{x})$ should allow us to evaluate the *level of consensus* between the different hypotheses, and to estimate the margin still available for an improvement of the approach when adding new hypotheses and/or increasing the size of the datasets.

Epistemic uncertainty, relative to ensemble set $\mathcal{H}$. In HEUQ, each participant produces a predictive distribution, so epistemic uncertainty can be operationalized as *distributional disagreement* across heterogeneous hypotheses. A principled way to quantify such disagreement is through information-theoretic divergences (Shaker and Hüllermeier, 2020; Hüllermeier and Waegeman, 2021). In particular, epistemic uncertainty is commonly linked to the mutual information between hypotheses and outcomes, which can be expressed as a Kullback–Leibler (KL) divergence between

²There will always exist a part of the uncertainty in the definitive decision choice of classification, a binary random choice (e.g. between to churn or not)

the joint distribution of hypotheses and outcomes and the product of their marginals. This captures how strongly predictions depend on the uncertain hypothesis (Hüllermeier and Waegeman, 2021). Motivated by this connection, we define $u_e(\mathbf{x})$ as the average divergence between each participant's predictive distribution and the ensemble predictive distribution produced by BCR:

$$u_e(\mathbf{x}) = \frac{1}{M} \sum_{j=1}^M KL[p(\cdot | h_j, \mathbf{x}), p(\cdot | BCR, \mathbf{x})]. \quad (3.4)$$

Here, $KL[q_1, q_2]$ denotes the Kullback–Leibler divergence between discrete³ probability distributions q_1 and q_2 . When $p(\cdot | BCR, \mathbf{x})$ is the uniform mixture of $\{p(\cdot | h_j, \mathbf{x})\}_{j=1}^M$, Eq. (3.4) corresponds to the Jensen–Shannon divergence Fuglede and Topsøe (2004). Essentially, Eq. (3.4) measures epistemic uncertainty as the average discrepancy between each hypothesis prediction and the aggregated (BCR) prediction.

It is then clear that $u_e(\mathbf{x})$ is always positive and can be made as low as possible by an appropriate management (adding/removing) of hypotheses. In other words, minimizing $u_e(\mathbf{x})$ can be one of the useful objective functions for guiding the selection of classifier models for the ensembles.

Aleatoric uncertainty, relative to ensemble set $\mathcal{H}$. Let us define the aleatoric uncertainty as the difference between total uncertainty and epistemic uncertainty:

$$u_a(\mathbf{x}) = u_t(\mathbf{x}) - u_e(\mathbf{x}). \quad (3.5)$$

This can be expressed as:

$$u_a(\mathbf{x}) = -\frac{1}{M} \sum_{k=0}^1 \sum_{j=1}^M p(y_k | h_j, \mathbf{x}) \log_2(p(y_k | h_j, \mathbf{x})). \quad (3.6)$$

From Equation (3.6), it can be seen that $u_a(\mathbf{x})$ is always positive. This definition thus corroborates with that of Hüllermeier and Waegeman (2021). Given the adopted definitions and the fact that they are conditioned on $\mathcal{D}$ we see that for a single hypothesis ($\mathcal{H} = \{h_0\}$), the epistemic uncertainty $u_e(\mathbf{x})$ is 0.

While Equation (3.2) presents a simple formulation for the Basic Combination Rule (BCR) using a standard average, it might be appealing to assume that more complex convex approaches could yield better results. Interestingly, our additional experiments discussed in Appendix A suggest otherwise. Even with a more sophisticated *convex combination rule*⁴. However, our experiments show that it is difficult to achieve significant improvements over those obtained with BCR. This warrants further investigation and could yield significant insights.

³Binary.

⁴defined as a linear combination with non-negative coefficients summing to 1 Rockafellar (1970) pp. 11-12, significant improvements over BCR were difficult to achieve.

3.3.3 Heterogeneous Ensemble for Uncertainty Prediction: HEUQ

Our heterogeneous ensemble for uncertainty prediction, HEUQ, is presented in Figure 3.2. HEUQ extends the previous approach presented in Shaker and Hüllermeier (2020) on uncertainty prediction from homogeneous ensembles to a heterogeneous one. We also draw inspiration from the results of Lakshminarayanan, Pritzel, and Blundell (2017) who explored homogeneous strong learner ensembles (Deep ensemble), and from Bogaert and Delaere (2023), who showed the outstanding performance of heterogeneous ensembles. Finally, our work builds upon the prior research on uncertainty quantification (UQ) in Machine Learning of Hüllermeier, Destercke, and Shaker (2022) and Hüllermeier and Waegeman (2021), and Depeweg et al. (2018).

We adopted such an approach since it lent itself to our definition of epistemic uncertainty, defined in Equation (3.4), as the average divergence between the predictions of the individual models in the ensemble and the ensemble's overall prediction. We favored a heterogeneous ensemble since such ensembles have been shown to outperform individual classifiers and homogeneous ensembles in past research (Mohammed and Kora, 2023; Ekbal and Saha, 2011; Prusa, Khoshgoftaar, and Dittman, 2015; Shahzad and Lavesson, 2013; Bogaert and Delaere, 2023; Sesmero, Ledezma, and Sanchis, 2015). As discussed previously, the question of whether heterogeneous ensembles increase or reduce uncertainty estimates has not yet been addressed in extant literature. The use of several varied models, each learning their hypotheses (Figure 3.1), leads to the exploration of the larger hypothesis space, and potentially, to lower uncertainty in predictions (Woźniak, Grana, and Corchado, 2014).

We followed the following steps to create our ensemble.

1. **Base classifier selection:** Drawing insights from previous research in the use case of customer churn prediction (De Caigny, Coussement, and De Bock, 2018; Spiteri and Azzopardi, 2018; Zhang et al., 2017; Verbraken, Verbeke, and Baensens, 2012; Scriney, Nie, and Roantree, 2020; Bolancé, Guillen, and Padilla-Barreto, 2016; Ahn et al., 2020; Mukhopadhyay et al., 2021; Lalwani et al., 2022; Bogaert and Delaere, 2023), we select the following base learners for HEUQ:
 - (a) Deep Neural network(DNN)
 - (b) Random Forest (RF)
 - (c) Logistic Regression (LR)
 - (d) Bagging of decision trees (BG)
 - (e) XGBoost (XGB)
 - (f) CatBoost (CB)

We employ DNN because of their excellence in modelling and capturing complex nonlinear relationships compared to simpler models like LR (Vafeiadis et al., 2015). We define DNN as networks that have more than two hidden layers between input and output layers.

RF and LR, besides being robust to noise (Zhang et al., 2017; Breiman, 2001) and are frequently used in the use case of churn prediction, with LR often considered a benchmark (Bogaert and Delaere, 2023).

BG and XGB are well known to improve predictive performance while avoiding overfitting and improving overall generalization (Fan et al., 2006), (Chen and Guestrin, 2016).

CB is particularly well suited for categorical data, a common characteristic (Prokhorenkova et al., 2018). Furthermore, it has received less attention compared to other methods in OR and churn prediction (Bogaert and Delaere, 2023). More information on hyperparameters and parameters are discussed in Section 3.5.2.

2. **Classifier combination:** While various classifier selection and combination techniques exist, we employ the *standard average* method. The choice is motivated by: i) the objective of our research was not to study classifier combinations in ensembles, and ii) it is a simple, yet efficient method, enabling us to focus (and isolate) on the issue of uncertainty. Despite its simplicity, it outperforms a more complex combination approach, as described in Appendix A.
3. **Uncertainty estimation:** The probability distributions of all base classifiers are utilized to estimate the aleatoric and epistemic uncertainty for individual data points, as discussed in Section 3.3.2 and Section 3.3.2 with Equations (3.3), (3.4), (3.5), (3.6).

Although the proposed HEUQ comprises of six base classifiers, it can be readily extended to encompass additional models, catering to diverse business and data requirements, as illustrated in Figure 3.2.

3.4 Data

This chapter brings together two complementary industrial case studies based on the HEUQ framework, each originally evaluated on different datasets. To avoid ambiguity, we clearly distinguish the datasets used for (i) churn prediction and (ii) customer conversion prediction.

Churn Prediction Datasets. Three datasets are used for the churn experiments and are summarised in Table 3.1:

- **Proprietary Insurance Churn Dataset (Version 2):** The internal dataset from our industrial partners, previously introduced in Section 2.5.2.
- **Bank Churn Dataset (Public):** an open-source benchmark widely used in academic studies.

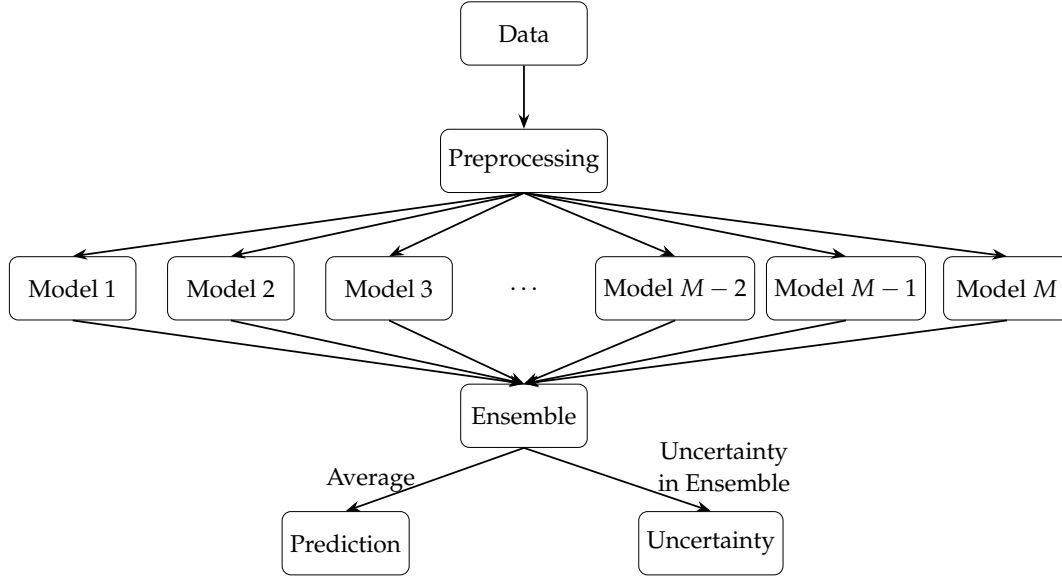


FIGURE 3.2: HEUQ architecture, heterogeneous ensemble approach offers the advantage of combining various independent models/hypotheses. This also allows the flexibility in the rule of combination, such as convex or basic combination rule.

- **Telecom Churn Dataset (Public):** Another standard churn benchmark.

Table 3.1 summarizes the data source, sector, number of instances, distribution of numerical and categorical features, churn rate, and example variables for the churn datasets. For all public datasets, we provide URLs so that readers can access the original data and detailed feature descriptions.

Customer Conversion Dataset. For the conversion prediction task, we use a second proprietary dataset from our industrial partners, described in Section 3.4.2.

Most publicly available datasets tend to be of limited size and may not fully reflect real-world data, our proprietary dataset helps in addressing it.

Id	Source	Sector	Instances	Numeric features	Discrete features	Churn rate	Example variables
1	Private	Insurance	250,000+	24	25	>10	- age, gender, product type, tenure, postcode, family, contract type premium discount
2	Kaggle ^a	Bank	10,000	6	4	20	- credit score, age, tenure, gender, area, active, has credit card
3	Kaggle ^b	Telecom	7043	3	16	27	- tenure, monthly/total charge, gender, streaming TV

TABLE 3.1: Overview and characteristics of Churn Data set

^a<https://www.kaggle.com/shubh0799/churn-modelling>

^b<https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

Due to confidentiality agreements, we can only share a limited sample of the variables and their descriptions and challenges with this specific dataset. By leveraging a combination of private and publicly available datasets, our study encompasses

data from multiple industries and captures diverse characteristics. This approach enhances the generalizability of our findings to real-world scenarios.

We also elaborate on the data transformation schemes (Section 3.4.1) applied to investigate how any transformation and distortion in the data impact uncertainty in predictions. The data was randomly split with the test and validation set comprising of 30% of the original data and the train set comprising of remaining 70%. We ensured that the ratio between majority and minority class was maintained during the train test creation.

3.4.1 Proprietary churn dataset (Version 2)

For the main churn experiments, we received an anonymized dataset with over 250,000 records from our industrial partner, a major insurance provider. A detailed description of the original data can be found in Section 2.5.2. Here, we briefly summarize the data again. The data had been pre-processed and cleaned (dealing with missing values, removing null values, dealing with outliers) by our partner given their domain knowledge. An illustrative snapshot of the data can be found in Table 3.2. Each data point represents a customer’s specific policy (e.g. car, fire, theft...), allowing us to predict churn (canceling) at the policy level. Our prediction for churn (or not churn) is therefore made at the policy level per customer profile. Such a policy-level data structure is a design choice by our industrial partner. This structure helps identify customers at risk of canceling other policies with the provider. The features (independent variables) provided a comprehensive description of each policy and the associated customer profile, encompassing customer-related and product-related information. The dataset consisted of a single dependent variable representing whether a policy had churned or not.

POLICY	AGE	CUSTOMER TIME	POST CODE	FAMILY	...	CONTRACT TYPE	CATEGORY	PREMIUM	CHURN
XYZ1	53	2	2020	1	...	1	1	2	0
XYZ2	35	2	3001	0	...	2	1	2	0
XYZ3	24	1	5021	1	...	11	1	2	0
XYZ4	24	1	4000	2	...	23	1	2	0
XYZ5	51	2	1000	1	...	6	1	2	0
...	...	...	...	...	...	...	...	...	...
XYZ99	34	1	1020	2	...	12	2	2	1
XYZ92	61	1	1210	1	...	98	2	2	0
XYZ96	55	1	5003	0	...	12	1	2	0
XYZ101	48	2	4053	1	...	11	1	2	1
XYZ119	27	2	9041	0	...	2	1	2	0

TABLE 3.2: Data snapshot

Data challenges and mitigation strategy

Class imbalance: All churn datasets, and especially the proprietary dataset, exhibit strong class imbalance, with churn representing less than 10% of all observations. Data imbalance biases the training towards the majority class, leading to overfitting.

Undersampling and oversampling are common approaches to address imbalance. Although these techniques can be effective in improving model performance,

they also carry certain drawbacks. Undersampling reduces the data of the majority class, aiming for balance. However, this discards potentially valuable information and relationships in data while reducing the overall data size. Oversampling replicates or generates new data points for the minority class. Oversampling, potentially introduces the risk of overfitting, as the model may learn to rely heavily on the replicated data points, leading to poor performance on unseen data (Batista, Prati, and Monard, 2004). Furthermore, oversampling techniques often involve generating new data points, which may not accurately represent the actual distribution or attribute values of the data. We experimented with SMOTE (Synthetic Minority Oversampling Technique), which generates synthetic instances based on the similarity (distance) to existing ones (Chawla et al., 2002). However, our results were unconvincing since SMOTE generated data points that were unrealistic, e.g. generating postal codes that did not exist in reality.

Consequently, for our study, we relied on *weighted training* also known as cost-sensitive learning to avoid losing samples with under-sampling or generating new data employing oversampling (Thölke et al., 2023). Weighted training involves placing more importance on some samples or a class (Hashemi and Karimi, 2018).

Homogeneity: The majority of features had very similar values across both classes (i.e. churners vs. non-churners). This compounded the difficulties of learning how to accurately separate both classes.

Data transformation for churn experiments

We conducted further experiments to study the robustness of HEUQ (Section 3.3). In particular, we will investigate whether transformations and distortions applied to the data influence HEUQ’s predictive ability and associated uncertainty.

We, therefore, considered the following three data transformation schemes

- No transformation (Vanilla)
- Principal Component Analysis (PCA)
- Gaussian Random Projections (RG)

PCA: PCA is a popular dimensionality reduction technique based upon identifying the principal components characterizing the data (Jolliffe and Cadima, 2016)

RG is a dimensionality reduction technique employing the Johnson-Lindenstrauss lemma, stating that points in a vector space of sufficiently high dimensions can be projected to new dimensional space while approximately preserving pairwise distances between points (Lindenstrauss, 1984). RG employs Gaussian distribution to generate random projections. It allows for controlled distortion (similar to noise) to be introduced in the data.

3.4.2 Proprietary Customer conversion dataset description

The customer conversion experiments are based on a second proprietary dataset containing approximately 190,000 records from our industrial partners in the vehicle insurance segment. The target variable indicates whether a marketing offer (for vehicle insurance) has been converted into a successful sale. The dataset includes a rich set of features describing customer demographics, portfolio composition, vehicle specifications, and behavioural attributes.⁵

Customer demographics encompass data related to the characteristics of our customer base. 21.2% of customers are 25 years old or younger, 46.8% fall between 26 and 45 years old, 26.1% are aged between 46 and 65, and 5.9% are 66 years old or above. The gender distribution reveals a predominance of male customers (62.4%) compared to females (37.6%). Customers are also categorized into lifecycle segments such as young professionals and retirees, providing further granularity in demographic profiling.

Portfolio composition reflects the insurance products that customers are currently subscribed to, offering insights into their coverage preferences and product bundling behaviour.

Vehicle specification includes specifications such as vehicle age and engine power. The average vehicle age is 7.5 years, with a standard deviation of 5 years; vehicle ages range from 0 to 60 years. Engine power distribution shows that 78% of registered vehicles have engines rated at 100 kW or less, 20.7% fall within the 101–200 kW range, and only 1.3% exceed 200 kW.

Additionally, the dataset incorporates behavioural variables such as whether customers have opted in to receive marketing communications, the number of insurance quotes requested in the past 30 days, and any history of insurance claims. These behavioural indicators provide important signals of customer intent and engagement.

We divide our data into two primary groups: Physical (Phy) and Digital (Web) channels. Physical selling refers to any form of customer engagement and offers that are made by a representative, such as face-to-face, or through a phone interactions. This can include insurance offers made directly in the insurance provider's office or over the phone with a representative assisting the customer. On the other hand, digital selling involves interactions that occur online, through digital platforms such as a company's website, mobile application, or through email marketing campaigns. These digital channels enable offers to be made without direct human interaction. In our data, conversion rates are substantially higher in physical channels than in digital channels, although precise figures cannot be disclosed for confidentiality reasons. This multi-channel structure motivates part of our analysis in the customer conversion study. The dataset provided to us by our partners for this study was of

⁵Due to the confidential nature of the data, we disclose only information that will not cause any prejudice to the insurer.

exceptionally high quality, with an extremely low percentage of missing values, approximately only 0.0001%. This minimal amount of missing data suggested that the dataset was largely complete and reliable for analysis. For the few instances where missing values were present in certain categorical columns, we employed mode imputation as the method for filling these gaps. This approach was chosen with the assumption that substituting missing data with the mode would not cause any substantial distortion in the overall distribution of the data. Additionally, we applied label encoding to transform the categorical variables into numerical representations. The customer conversion prediction dataset presented challenges similar to those detailed in Section 3.4.1, which we addressed using weighted training. However, we did not evaluate the effects of data transformation on uncertainty for this specific dataset. Portions of this research were previously presented at the ICAEB conference.

3.5 Experimental Setup

After presenting our proposed method, HEUQ, the dataset and data transformation schemes, we now describe our experiment setups to address the gap in the literature as discussed in Section 3.1. We define three sets of experiments.

1. **Churn Prediction Performance:** We are concerned here with the predictive performance, i.e. whether HEUQ outperforms previously proposed methods, including Deep ensembles and other homogeneous ensembles of strong learners. Performance is evaluated using balanced accuracy, a widely used metric for churn prediction (Ahn et al., 2020; Vafeiadis et al., 2015; Xie et al., 2009; Spiteri and Azzopardi, 2018; Bolancé, Guillen, and Padilla-Barreto, 2016; Mukhopadhyay et al., 2021) Results are presented in Table 3.3, and are organized into three settings corresponding to the aforementioned data transformation schemes (Section 3.4.1):

- (a) Vanilla
- (b) PCA
- (c) RG

These settings allow us to analyze the performance of HEUQ under varying data representations and sensitivity to data distortions.

2. **Uncertainty Estimation:** We evaluate and compare the uncertainty estimation (UE) capabilities of HEUQ against homogeneous ensembles of LR, BG, XGB, RF, Cat and Deep ensemble. Thus, we follow the study on UQ of Shaker and Hüllermeier (2020). We employ stratified Brier score and Negative log-likelihood as metrics (Wallace and Dahabreh, 2012) to compare the performance of the learners.

3. **Ablation Experiments:** These experiments aim to address the impact of different participating strong learners HEUQ’s composition on predictive accuracy and uncertainty estimation.

For the **customer conversion** case study, we adopt the same set of base learners but focus on predictive performance rather than uncertainty decomposition. In particular, we assess model quality using balanced accuracy and F1 score, and analyze the impact of model choice across physical vs. digital channels. We do not apply PCA or RG transformations to this dataset, in line with the more application-oriented nature of the conversion study.

3.5.1 Performance measure

We employ Balanced Accuracy, Negative Log-Likelihood, and Brier score as performance metrics for churn prediction task (Section 3.4.1). Balanced accuracy assesses the predictive (classification) performance, while Negative Log-Likelihood and Brier score are commonly used for uncertainty estimation / quantification metrics (Lakshminarayanan, Pritzel, and Blundell, 2017; Guo et al., 2017). For the task of Customer conversion (Section 3.4.2), we relied only on Balanced accuracy and F1 score.

Balanced Accuracy (BA): BA is a useful metric for evaluating the performance of classification tasks where the training data is imbalanced. BA evaluates imbalanced classification by averaging TPR and TNR, as described below.

$$\text{Balanced Accuracy} = (TPR + TNR) / 2.$$

F1 Score: The F1 Score represents the harmonic mean of Precision and Recall, and is especially useful under uneven class distributions, and is calculated as follows:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Brier score (BS): Brier score is a popular metric to evaluate the calibration and sharpness of probabilistic predictions. In the context of uncertainty, the score measures whether the model’s predictions reflect the underlying distribution of labels. A low Brier score suggests the model predictions are well-calibrated and are consistent with observed data. Brier score is defined as the mean squared difference between predicted probability and the observed outcome. BS is also indicative if a model is overconfident or underconfident with the prediction.

We employ the *stratified* Brier score as proposed by Wallace and Dahabreh (2012) as:

$$\frac{\sum_{i=0}^N (y_i - \hat{P}\{y_i | x_i\})^2}{N}.$$

The stratified version takes into account the imbalance in the data. It overcomes the issue of poor calibration and imbalance of minority classes of the original Brier score (Wallace and Dahabreh, 2012). The BS further was interpreted as $BS = \text{uncertainty} - \text{resolution} + \text{reliability}$, by Ovadia et al. (2019). Resolution measured the deviation of point predictions against the marginal, and reliability measured calibration as the average violation of long-term true label (Ovadia et al., 2019).

Negative Log-Likelihood (NLL): NLL in the context of UE is indicative of how well the model's predictive distribution accounts for the uncertainty in the data (Ovadia et al., 2019). A low NLL score indicates that the model's predictive distribution is a good fit and it effectively captures the variability and uncertainty in the data.

$$- \left(\sum_{i=1}^n y_i \log p(x_i) + (1 - y_i) \log (1 - p(x_i)) \right)$$

where i is an observation, n is the number of observations and y_i is the outcome for i , while $p(x_i)$ is the predicted probability of the model.

A combination of low BS and NLL scores is indicative that the model is well-calibrated and sharp, and its predictions have low uncertainty (Ovadia et al., 2019).

Friedman test Rank: To evaluate and compare the performance of classifiers across all datasets and their transformations (Vanilla, PCA, RG), our study employs the non-parametric Friedman test (Bogaert and Delaere, 2023). This test assesses whether statistically significant differences exist in the average ranks of the classifiers across the different datasets and transformations.

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right]$$

The test is non-parametric and compares the average rank R of the classifiers k across multiple data sets N , with the null hypothesis that states that all ranks are equal, resulting in the test statistic (Bogaert and Delaere, 2023). Here, R_j gives the average rank of the classifier j over the N data sets

3.5.2 Training and implementation details

Scope and provenance. This chapter consolidates two empirical studies that used the same HEUQ methodology but different business problems and data-generation processes (customer churn vs. customer conversion). Because the studies were conducted as separate projects, their experimental protocols (data splitting and model selection) are not identical. In the thesis, we describe both studies using a common structure and notation, while keeping the original protocols to preserve comparability with their respective baselines.

Data splitting and model selection. Churn prediction: As discussed in Section 3.4 for all datasets, we split the data into a training and testing set, with 70%

of the stratified sampled for training, and the remaining 30% for testing and validation. Validation data is used for model selection and early stopping. Final results are reported on the held-out test split.

Customer conversion: Due to limited data availability, we use a stratified 80:20 train–test split. Model selection is performed within the training split using an exhaustive grid search (GridSearchCV) with stratified 5-fold cross-validation. Final results are reported on the held-out test split.

Homogeneous ensemble baselines. For each dataset, we construct homogeneous ensembles that use the same outer train–test split as HEUQ and are evaluated on the same held-out test set. Following Shaker and Hüllermeier (2020), for each base learner type (LR, BG, XGB, RF, CB, NN), we train an ensemble of ten models. Each ensemble member is trained on a different stratified 90% subsample of the training split, while the remaining 10% is left out for that member. Ensemble predictions are obtained by averaging member probabilities, as in HEUQ. This design yields bootstrapped ensembles of size ten trained on stratified subsets of the data and enables fair performance and uncertainty comparisons under matched evaluation protocols.

Hyperparameters. We conduct extensive hyperparameter tuning for all baseline algorithms and report the best-performing configurations (with respect to BA) used in the final experiments. For baseline models (including homogeneous ensembles), we select hyperparameters via exhaustive grid search (GridSearchCV) over predefined parameter grids. For HEUQ, hyperparameters are tuned *jointly* across heterogeneous participants; since this yields a large combined search space, we rely on Bayesian hyperparameter optimization. At each iteration, a candidate joint configuration is proposed, the participating learners are trained, and the resulting HEUQ ensemble (combined by standard averaging) is evaluated using an ensemble objective (high BA, low BS, and low NLL). In both studies, the selected configurations were very similar across learners; a plausible explanation is that both datasets were collected by the same industrial partner and therefore share similar data choices and underlying data-collection processes. The selected configurations are listed below.

- **Baseline LR:** L2 penalty, 1000 maximum iterations, 1bfgs solver.
- **Baseline BG with decision trees:** maximum depth 6, 700 estimators.
- **Baseline RF:** maximum depth 4, 700 estimators, maximum features 6.
- **Baseline XGB:** maximum depth 3, 100 estimators, learning rate 0.1; `scale_pos_weight` for weighted training.
- **HEUQ:** Participants trained on the same dataset as the baselines; combined by standard averaging. Selected participant hyperparameters:
 - LR: 1000 maximum iterations, L2 penalty, 1bfgs solver.
 - RF: 300 estimators, maximum depth 9.

- BG with decision trees: maximum depth 6, maximum samples 30%, 700 estimators.
 - CB: depth 5, learning rate 0.1, class weight 12 for churn samples.
 - XGB: 300 estimators, learning rate 1.
 - NN: two hidden layers (2048, 256), dropout 50%, 30 epochs, learning rate 0.00025; weighted cross-entropy loss; Adam optimizer; weight decay 0.0001.
- **Data Transformation:** We explore the impact of data transformation, with PCA and RG, on the performance of our proposed model. For PCA, 95% of the data variance was explainable with an embedding size of 40 for private, 15 for Telecom, and 10 for Bank data. We evaluated different embedding sizes for RG using various distortion rates (*Epsilon*: 0.1, 0.3, 0.5, 0.9) on private data (10, 20, 30, 40), Telecom data (5, 10, 15, 17), and Bank data (5, 7, 10). *Epsilon* parameter controls the level of noise or distortion injected in the data during RG projection. A higher value indicates a lower distortion. Note that embedding size refers to the size of the resulting vectors after their projections on a new, reduced-dimension space by PCA and RG.

All experiments were performed on a Linux machine with 16 cores, 16 Gb RAM, and an Intel i7 chipboard. We employed Python as the programming language and utilized widely-used open-source libraries such as Sklearn and Pytorch for modelling, and Numpy, Pandas, and Scipy for data transformation, handling, and the Friedman test.

3.6 Results

We now present and discuss the empirical results obtained for churn prediction, customer conversion, uncertainty quantification, and the ablation study, as described in Section 3.5.

3.6.1 Experiments: Churn Prediction

We evaluate the performance of HEUQ against all baseline methods presented in Section 3.5 for the task of churn prediction. BA is used as the performance metric. Results are presented in Table 3.3. Along the rows are all the baseline methods (Section 3.5). Three variants of HEUQ, each with a specific data transformation scheme, viz. vanilla, PCA, and RG, are also presented. As already mentioned, these variants aim at investigating the impact of data transformation (PCA) and controlled distortion (using RG) on HEUQ's performance. For conciseness, we report only the embedding size and distortion values that yield the best-BA.

We refer the reader to Appendix A.2 for HEUQ's performance with the full set of embedding sizes and epsilon values.

Model	Data Transformation	Private data		Bank data				Telecom data		
		Embedding Size	Distortion	Balanced Accuracy	Embedding Size	Distortion	Balanced Accuracy	Embedding Size	Distortion	Balanced Accuracy
Baseline LR	Vanilla ^a	-	-	64.01	-	-	66.56	-	-	76.4
Baseline BG	Vanilla	-	-	63.6	-	-	76.06	-	-	77.51
Baseline RF	Vanilla	-	-	62.7	-	-	76.48	-	-	76.22
Baseline XGB	Vanilla	-	-	63.4	-	-	77.2	-	-	76.8
Baseline Cat	Vanilla	-	-	64.97	-	-	77.73	-	-	76.8
Baseline NN	Vanilla	-	-	64.02	-	-	75.7	-	-	77.1
Baseline LR	PCA	40	-	63.74	10	-	68.3	15	-	76.3
Baseline BG	PCA	40	-	63.4	10	-	76	15	-	76.13
Baseline RF	PCA	40	-	62.63	10	-	75.47	15	-	76.2
Baseline XGB	PCA	40	-	60.52	10	-	76.67	15	-	75.04
Baseline Cat	PCA	40	-	63.85	10	-	76.95	15	-	75.84
Baseline NN	PCA	40	-	64	10	-	75.8	15	-	76.2
Baseline LR	RG	40	0.3	64.03	10	0.9	68.31	5	0.1	74.15
Baseline BG	RG	40	0.5	62.31	7	0.9	71	10	0.1	73.74
Baseline RF	RG	40	0.9	61.48	7	0.1	70.8	15	0.1	73.54
Baseline XGB	RG	30	0.5	62.15	7	0.5	70.7	15	0.9	73.86
Baseline Cat	RG	30	0.5	62.87	10	0.5	69.8	5	0.3	73.83
Baseline NN	RG	30	0.1	62.3	10	0.9	70.1	15	0.5	73.68
HEUQ	Vanilla	-	-	64.95	-	-	77.49	-	-	77.82
HEUQ	PCA	40	-	64.45	10	-	77.50	15	-	77.8
HEUQ	RG	40	0.5	64.46	10	0.3	72.16	17	0.3	74.73

TABLE 3.3: Churn prediction performance across datasets, models, and data transformations (Balanced Accuracy)

^aNo data transformation

Model	Friedman test Rank ↓
LR	73.5
BG	73.31
RF	72.77
XGB	72.43
CB	72.15
NN	71.83
HEUQ	68.49

TABLE 3.4: Friedman test ranking comparing all baseline learners for all datasets and different data transformation (p -value:0.0051)

From Table 3.3, it can be seen HEUQ has marginally better performance compared other methods in terms of Balanced Accuracy (BA) for most of datasets and transformations. The results are in line with the findings of Bogaert and Delaere (2023) on the better performance of heterogeneous ensembles.

Overall the results in Table 3.3 confirms our findings with regard to our research question that a heterogeneous ensemble like HEUQ are comparatively more robust to transformation and distortions introduced in data than individual learners. Interestingly, all methods including HEUQ exhibit some sensitivity to data transformations. Specifically, all of them experience performance loss. However, the magnitude of the loss varied across models. (Baseline) LR was the least impacted by all data transformations, experiencing a maximum performance decrease of 2.25% on the BA dataset when RG was applied (Telecom data). However, LR consistently underperformed compared to HEUQ, particularly on the Bank dataset, where its BA score was below 11%. We observed the performance jump by 1.74 and 1.75% in Bank data when applied PCA and RG. This enhancement may be attributed to nature of more relevant linear features or the removal of extraneous noise through dimension reduction. XGB was generally found to be the most sensitive to transformations, compared to other methods for all data transformations, and experienced a

maximum performance drop of 6.5% in BA with RG in Bank data. Cat also demonstrates higher sensitivity to data distortion introduced by RG when compared to other methods for all datasets. It is evident that Cat’s baseline performance (64.97% and 77.73% BA score) was nearly identical to that of HEUQ (64.95% and 77.49%) for private data without any data transformations. Nonetheless, upon applying PCA and RG, CB’s performance drastically decreased by 0.78 to 7.93% across various datasets. In contrast, HEUQ exhibited a minor improvement of 0.1 to a comparatively smaller decrease of 5.33%.

Table 3.3 also indicates that HEUQ’s performance remains relatively stable, regardless of whether the dataset is left unperturbed in its initial state (vanilla), transformed by PCA, or distorted via RG. These results suggest that, on average, a heterogeneous ensemble, such as HEUQ, is more resilient to performance degradation due to the presence of noise or other distortions in the data. We thus address our research question first research question, on the impact of data transformation on our ensemble’s performance.

To assess whether the better performance of HEUQ is statistically significant, we conducted a Friedman test to compare the balanced accuracy, which produces a ranked-list of the classifiers evaluated at at given p -value. We construct a Null hypothesis that there is no significant difference in performance between HEUQ and baseline methods. This test considers all the classifier variants, i.e. all data transformations (PCA, RG, vanilla and all datatesets). A lower-ranked classifier is considered as having a better performance. Results of the Friedmann test are given in Table 3.4, with a p -value of p -value of 0.0051. Since this value is lower than our significance level of 0.05, we can conclude that HEUQ (lowest-ranked according to the Friedmann test) has the best performance across all the classifiers for all datasets and different data transformations. considered.

3.6.2 Experiments: Uncertainty quantification churn prediction

We compare the performance of these homogeneous ensembles against HEUQ employing the negative log-likelihood (NLL) and Brier score (BS), our two metrics for uncertainty quantification (Section 3.5.1). We also report on the BA to assess whether a gain (loss) in predictive performance (BA) engenders a loss (gain) in the associated uncertainty. Ideally, we would like methods that achieve high predictive performance with the lowest associated uncertainty. Results are presented in Table 3.5. The arrow direction for each metric indicates whether a higher ($\uparrow$) or a lower ($\downarrow$) value is desirable. For e.g., lower values of NLL, BS, and uncertainty are desirable as opposed to BA. Note that HEUQ’s BA score is different from that reported in our earlier experiments in Table 3.3. This is explained because we employ 90% bootstrap subset of original training data but emoloy same test data as previously discussed in Section 3.5.2. Table 3.5 shows that HEUQ consistently achieved lower or, in most

cases, the lowest uncertainty scores (i.e. NLL and BS) compared to the other ensembles. In addition, it also achieved the highest BA score among all the homogeneous baseline ensembles. While HEUQ’s performance is comparable, it surpasses state-of-the-art deep ensembles by 1.8% and 4.34% on the bank and telecom datasets, respectively. Thus, HEUQ outperforms the other ensembles in terms of predictive accuracy. Moreover HEUQ demonstrates better or comparable scores with lower BS and NLL values. When considering all three metrics collectively, HEUQ emerges as the better-performing method, in terms of calibration and predictive accuracy. These results enable us to conclude that the predictions from heterogeneous ensembles, such as HEUQ, are better calibrated. They tend to be associated with lower uncertainty, thereby addressing our second research question. To validate our results, we conducted a Friedman test with a similar Null hypothesis as discussed in Section 3.6.1 and found that HEUQ marginally rank better compared to the other baseline methods with a p-value of 0.007.

Model	Private data			Bank data			Telecom data			Friedman test Rank ↓
	NLL ↓	BS ↓	Balanced Accuracy ↑	NLL ↓	BS ↓	Balanced Accuracy ↑	NLL ↓	BS ↓	Balanced Accuracy ↑	
Homogeneous LR	0.63	0.221	64.19	0.45	0.15	75.32	0.48	0.16	76.6	24.394
Homogeneous BG	0.633	0.22	63.53	0.53	0.17	75.55	0.46	0.15	77.1	24.174
Homogeneous RF	0.623	0.212	62.31	0.48	0.15	75.96	0.49	0.16	76.23	24.151
Homogeneous XGB	0.65	0.23	64.11	0.46	0.15	76.07	0.46	0.15	77.2	24.074
Homogeneous Cat	0.57	0.206	62.58	0.48	0.16	76.17	0.48	0.16	76	23.923
Deep ensemble	0.71	0.24	64.20	0.72	0.16	73.05	0.75	0.21	76	23.767
HEUQ ^a	0.6	0.204	64.36	0.42	0.18	77.39	0.46	0.14	77.8	23.484

TABLE 3.5: Uncertainty quantification baseline. *p-value:0.007* for Friedman test ranking

^aBase learners of HEUQ ensemble trained on 90% of training data

^bAll models tested on 30% original test data

3.6.3 Ablation study

To gain a deeper understanding of our results, we performed an ablation study by replacing and removing learners from our ensemble for private churn data. The ablation study helps us answer the effects of modifying an ensemble configuration on the performance and uncertainty, corresponding to our third research question.

Heterogeneous strong learners

We created several variants of HEUQ and tested on our challenging private churn data to gain more insight into the performance of the ensemble by discarding one learner from the ensemble. That is, each HEUQ variant consisted of five learners. They are presented in Table 3.6. Each learner that was discarded from the original HEUQ is indicated as ~~learner~~. In addition, we created a heterogeneous ensemble consisting only of tree-based learners (last row in the table). Along the columns, the performance metrics are reported, including Negative Log Likelihood (NLL), Brier Score (BS), and Balanced Accuracy (BA). We also report on the estimation of the

uncertainty by measuring the aleatoric and epistemic uncertainty. The arrow listed with each metric indicates whether a higher ($\uparrow$) or a lower ($\downarrow$) value is desirable. For e.g., lower values of NLL, BS ($\downarrow$), and uncertainty are desirable as opposed to BA($\uparrow$).

Model type	NLL $\downarrow$	BS $\downarrow$	Balanced Accuracy $\uparrow$	$u_a\downarrow$	$u_e\downarrow$	$u_t\downarrow$
HEUQ(LR,RF,BG,CB,XGB,NN)	0.624	0.219	64.92	0.8960	0.0181	0.9081
HEUQ(LR,RF,BG,CB,XGB,NN)	0.621	0.215	64.87	0.8645	0.0311	0.8956
HEUQ(LR,RF,BG,CB,XGB,NN)	0.628	0.216	64.85	0.8694	0.0304	0.8998 ^b
HEUQ(LR,RF,BG,XGB,CB,NN)	0.627	0.215	65.05	0.8547	0.0290	0.8837
HEUQ(LR,RF,BG,XGB,CB,NN)	0.629	0.214	65.05	0.8580	0.0307	0.8887
HEUQ(LR,RF,BG,XGB,CB,NN)	0.626	0.216	64.9	0.8632	0.0309	0.8941
HEUQ(LR,RF,BG,CB,XGB,NN) ^a	0.625	0.228	64.64	0.9175	0.0081	0.9256

TABLE 3.6: HEUQ ablation with heterogeneous strong learners

^aOnly tree learners.

^bMean Aleatoric and epistemic uncertainty reported on the test set.

Several interesting observations can be made from the Table 3.6. Specifically, Aleatoric uncertainty accounts for a substantial fraction of the total uncertainty. This is to be expected given the definitions of uncertainties and the characteristics of our private dataset. As noted earlier, we had a highly imbalanced distribution between churners vs. non-churners, and several features had the same values for churners and non-churners (homogeneity) (Section 3.4.1).

The above results also show that some of the variants achieve a higher BA than the original HEUQ (see Table 3.5). Most notable variants HEUQ(LR,RF,BG, XGB, CB, NN) and HEUQ(LR,RF, BG, XGB, CB, NN) outperform HEUQ by 0.1 and HEUQ(LR, RF, BG, XGB, CB, NN) underperforms by 0.31 in BA scores. However, the original HEUQ achieved better results when considering uncertainty-specific metrics, such as NLL and BS. We can also observe that ensembles with neural networks had lower levels of aleatoric uncertainty in their data. Neural networks are indeed well-known for capturing complex nonlinear relationships in data. Also, ensembles restricted to tree-based methods achieved the lowest levels of epistemic uncertainty estimate. These observations seem to suggest that neural networks could effectively deal with the challenges present in our dataset and that tree-based methods achieve high consensus in their predictions. This further suggests that the hypotheses learned by tree-based methods are very close to each other in the hypothesis space. NLL and BS scores suggest that HEUQ is better calibrated and provides a higher level of certainty (lower uncertainty) in its predictions than the ablated variants (with fewer learners).

We can therefore conclude that the predictions of a heterogeneous ensemble, consisting of several diverse methods, tend to be associated with lower levels of uncertainty estimate. We attribute the higher calibration and certainty to the heterogeneity and larger hypothesis space explored by HEUQ. Our results further reinforce the notion that multiple strong learners with distinct loss functions, capable of capturing diverse hypotheses, exhibit enhanced learning and performance. While we could not test all possible combinations of strong learners proposed in HEUQ, from

Table 3.6, the results suggest that only tree methods do not perform well. The empirical evidence also indicates that increasing the number of methods employed can lead to better-calibrated results. The combination of tree-based learners with strong learners like LR and NN leads to slightly more calibrated ensembles. The result confirms the notion that different learners explore the hypothesis space in different ways, suggesting that tree-based learners may have some overlap in their hypothesis space. We attribute this overlap to their building block.

To further understand the relationship of uncertainty with BA, we employed the testing dataset, since all performance and uncertainty scores have been computed on it. As columns, we consider the epistemic uncertainty, aleatoric uncertainty, actual class labels, and predicted probabilities. The predicted probabilities serve to determine the predicted class labels. We set a threshold of 0.5 such that any predicted probability lower than this threshold will be classified as class 0 (not churn). These labels are then used to compute a BA score. The scores for epistemic and aleatoric uncertainties are used to compute the total uncertainty.

To analyse the relationship between predictive uncertainty and model performance, we partitioned the test set into four uncertainty-based quartiles, from the lowest (0–0.25) to the highest (0.75–1.00) uncertainty levels. Balanced accuracy decreases sharply across these groups, dropping from 82.53% in the lowest-uncertainty quartile to 50.70% in the highest (Figure 3.3). This strong negative correlation indicates that higher predictive uncertainty is reliably associated with lower classification performance, reflecting the expected behaviour of a well-calibrated uncertainty measure. These results confirm that HEUQ assigns high confidence to instances it predicts accurately while appropriately flagging cases where errors are more likely.

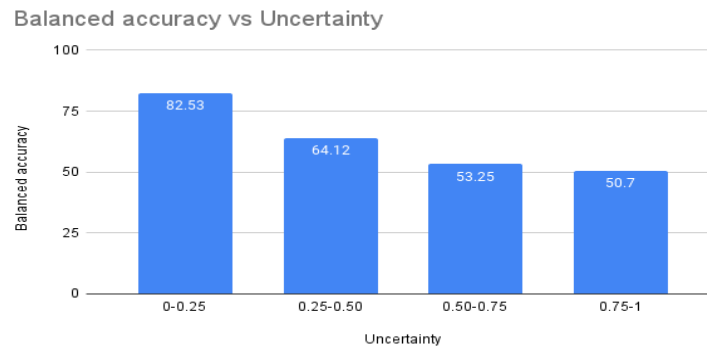


FIGURE 3.3: Balanced accuracy (BA) across uncertainty quartiles on the test set. BA decreases from 82.53% in the lowest-uncertainty quartile (Q1) to 50.70% in the highest-uncertainty quartile (Q4), confirming the negative relation between uncertainty and performance.

This verified that *higher uncertainty is indicative of lower performance* which is naturally expected.

3.6.4 Customer conversion (Case study)

We now turn to the customer conversion case study described in Section 3.4.2. Here, we evaluate the same set of base learners and HEUQ on two subsets of the dataset, corresponding to digital (Web) and physical (Phy) sales channels. Performance is assessed using Balanced Accuracy (BA) and F1 score (Table 3.7).

TABLE 3.7: Performance metrics of machine learning models on Web and Physical Datasets

Model	Dataset	BA	F1
LR	Web	80.64	87.5
RF	Web	81.22	87.4
BG	Web	81.65	85.9
XGB	Web	82.01	87.1
CB	Web	82.20	87.7
NN	Web	82.10	87.9
HEUQ	Web	82.05	87.3
LR	Physical	71.66	71.9
RF	Physical	72.61	72.7
BG	Physical	72.34	72.1
XGB	Physical	72.51	72.6
CB	Physical	73.66	73.6
NN	Physical	72.70	72.5
HEUQ	Physical	73.37	73.4

Web Dataset Performance

Across all models BA ranged from 80.64% to 82.2%, while the F1 score varied from 85.9% to 87.9%. Among the models, CB achieved the highest BA of 82.2%, slightly outperforming other models. This was followed by NN with a BA of 82.1% and then XGB with BA of 82.01%. Although CB achieved the best performance in BA, NN demonstrated a slight advantage in terms of F1 score (87.9% vs. 87.7%), which indicates that NN offers a somewhat better precision-recall trade-off. This may imply that while CB maximized overall classification accuracy, NN provided a more balanced performance by minimizing false positives and false negatives. LR, despite its simplicity, performed adequately well (80.64% BA, 87.5% F1), which suggests that the Web dataset is close to being linearly separable. HEUQ achieves BA and F1 scores that are essentially on par with the strongest individual models, showing that combining heterogeneous learners does not degrade performance in this setting.

Physical Dataset Performance

The Physical dataset presented a comparatively challenging classification task, with all models showing lower performance compared to the Web dataset. BA ranged from 71.66% to 73.66%, while F1 scores were between 71.9% and 73.6%. CB once again emerged as the top performer, achieving the highest BA (73.66%) and F1 score (73.6%). The performance gap between the best and worst models was notably smaller in the Physical dataset (approximately 2% in BA) compared to the Web dataset (approximately 1.5%). This suggests that the Physical dataset may contain

more complex or noisy patterns that are equally challenging for all model architectures. Once again, HEUQ demonstrates performance comparable to the best single model, confirming that heterogeneity does not come at the expense of predictive accuracy, and that HEUQ can be used consistently across different business tasks (churn vs. conversion) and channels (digital vs. physical).

Taken together, these results show that HEUQ delivers BA and F1 scores that are essentially on par with the best single model (CatBoost) in both the Web and Physical channels, while remaining consistently competitive across architectures. Although HEUQ is not always the top performer in terms of BA or F1 on the conversion task, it has an important additional advantage: by construction, it provides uncertainty estimates and an epistemic–aleatoric decomposition for every profile and channel. In the churn experiments (Section 3.6.1 and Table 3.5), HEUQ was shown to be better calibrated than homogeneous ensembles (lower NLL and Brier scores). This suggests that, in multi-channel customer conversion as well, HEUQ is a natural choice when decision-making requires not only high predictive accuracy but also reliable uncertainty information for prioritising leads or designing human-in-the-loop workflows.

3.7 Discussion and Conclusion

We presented Heterogeneous Ensemble for Uncertainty Quantification, a heterogeneous ensemble for uncertainty quantification, a model-agnostic uncertainty estimation framework. The study aimed to bridge the gap in extant literature in the field of uncertainty estimation with heterogeneous ensemble methods. Following Hüllermeier and Waegeman (2021), the total uncertainty is expressed as the sum of aleatoric and epistemic uncertainty. The latter is expressed as the average divergence between an ensemble’s overall prediction and the individual model’s prediction. Our premise is that the combination of a diverse set of methods will allow the learning of a wide variety of hypotheses, allowing a better understanding of the different sources of uncertainty and leading to actionable insights for management decisions/actions. Across the churn prediction experiments, HEUQ demonstrated competitive or superior balanced accuracy relative to strong single classifiers and bootstrapped homogeneous ensembles. More importantly, it consistently achieved lower negative log-likelihood and Brier scores, indicating improved calibration and more reliable confidence estimates. The method remained robust across different data transformations, including PCA and random projections, suggesting that heterogeneity in the ensemble helps mitigate the impact of distortions or noise in data representations. The ablation study further confirmed the importance of diversity: removing learners generally increased uncertainty and degraded calibration, highlighting the role of varied hypotheses in producing well-calibrated predictions.

The customer conversion analysis provides an additional empirical perspective on the applicability of HEUQ. While CatBoost slightly outperformed all methods

in balanced accuracy and F1 score on both Web and Physical channels, HEUQ remained highly competitive across both datasets. The experiments revealed a substantial performance gap between digital and physical channels, with physical interactions yielding 10% higher BA and 15% higher F1 scores. These findings point to fundamental behavioural differences between online and offline customers, physical interactions appear to foster greater trust and engagement, whereas digital interactions may require more personalised or targeted interventions to drive conversions. Although uncertainty metrics were not evaluated for this dataset, the strong calibration results observed for churn prediction (Table 3.5) suggest that HEUQ is well suited for applications where reliability, confidence scoring, or human-in-the-loop decision-making are important. In this sense, the conversion study illustrates HEUQ's practical relevance, even when performance differences between models are small.

Future research offers several promising directions building on the findings presented in this chapter. First, alternative uncertainty quantification frameworks could be examined to further strengthen the robustness and interpretability of heterogeneous ensembles. Exploring evidential, distribution-free, or hybrid UQ approaches may provide additional insight into how epistemic uncertainty can be more effectively captured in high-dimensional and imbalanced settings.

Second, the heterogeneous ensemble architecture developed here can be extended beyond churn prediction to a broader spectrum of operations research and decision-analytics problems. While the underlying methodology is expected to remain largely applicable, adapting it to domains characterised by structural constraints, sequential decisions, or multi-objective trade-offs represents an important avenue for future study.

Third, although this work primarily examined the BCR and the convex combination rule for aggregating classifier outputs, a wider family of optimization-based strategies could be incorporated. Techniques such as genetic algorithms, particle swarm optimization, differential evolution, or other evolutionary search methods may offer more flexible mechanisms for identifying optimal ensemble configurations under uncertainty.

Overall, this chapter contributes to a broader research agenda on uncertainty-aware heterogeneous ensembles. In particular, we argue that explicitly modelling epistemic uncertainty provides an additional criterion for selecting, weighting, and extending ensemble members, thereby supporting more informed and transparent decision-making processes.

Operational implications for trustworthy risk scoring. For our industrial partner, the practical value of uncertainty estimation is not limited to higher balanced accuracy, but extends to *reliability* in decision-making: risk scores must remain interpretable across customer segments and time, and must support stable operating points (e.g., intervention thresholds) under changing data conditions. Across the

churn experiments on the proprietary dataset (Version 2), HEUQ consistently improved calibration (e.g., lower NLL and Brier scores) while remaining competitive in predictive performance, and similar trends were observed on most additional benchmark datasets. Operationally, this enables policies that combine *risk* and *confidence*, for example prioritising high-risk customers with low epistemic uncertainty for automated retention actions, while routing uncertain cases to human review or to targeted data collection. In this way, epistemic uncertainty becomes actionable: it supports safer automation, better triage, and more transparent justification of when a model should be trusted.

From calibrated prediction to uncertainty-aware learning. The analysis above treats uncertainty primarily as an *assessment* tool applied after training in a supervised learning setting. However, many real decision processes are sequential and operate under noise, perturbations, and distributional shift, where errors can compound through feedback loops. This motivates the next step: embedding uncertainty directly into the learning dynamics so that it can modulate updates and exploration during training. Accordingly, Chapter 4 extends uncertainty quantification to reinforcement learning by integrating ensemble-style epistemic signals (via Masksembles) into Bellman-based learning, and evaluates robustness both on the proprietary dataset (Version 2) and on controlled OpenAI Gym environments.

Chapter 4

Uncertainty Aware Reinforcement Learning (UARL)

This chapter presents MASURE, an uncertainty-aware reinforcement learning framework designed to stabilize value learning under severe noisy observations. Building on Masksembles for efficient epistemic-uncertainty estimation, MASURE incorporates uncertainty directly into the Bellman update, reducing volatility in noisy environments and improving robustness across both online and offline settings. We evaluate the approach on classical open AI gym and real-world datasets, demonstrating substantial gains in stability and performance. This work have been published in the ITT conference (Singh et al., 2025d).

Contents

4.1 Introduction	69
4.2 Literature study	72
4.3 Methodology	73
4.3.1 Masksembles	73
4.3.2 Q-Learning and Contextual Bandits	74
4.3.3 MASURE: Masksembles for Stable and Uncertainty-aware RL Environments	75
4.3.4 Experimental Setup	76
4.4 Results	79
4.5 Conclusion	82

4.1 Introduction

In this chapter, we continue our study of Uncertainty Quantification but extend it to a different learning paradigm, i.e. Reinforcement Learning (RL). Most prior work on UQ has been developed within the context of classical machine learning methods, whereas here we depart from that literature by examining how UQ can be embedded directly within the RL learning process. As will be explained later, UQ in RL can play a crucial role in stabilizing the learning process. In Chapter 3, uncertainty was used primarily to assess prediction reliability in supervised learning. Here,

we go one step further and use uncertainty as a *learning signal* in RL, integrating it into Bellman-based updates via Masksembles to improve robustness while requiring fewer trained parameters than classical ensembles. In particular, reinforcement learning is especially sensitive in environments with extreme noise, perturbations, or distributional shift.

Reinforcement Learning (RL) is a key AI paradigm in which an agent learns optimal behaviour through trial and error, guided by rewards. Unlike supervised learning, RL does not rely on labeled data and is therefore well-suited for sequential decision-making problems. RL has gained substantial prominence in various notable fields like robotics, autonomous driving, recommender systems, finance, and resource management (Shakya, Pillai, and Chakrabarty, 2023). Deep RL underpins several major AI breakthroughs, including AlphaGo (Silver et al., 2016) and more recently RL with Human Feedback (RLHF) in systems such as ChatGPT (Achiam et al., 2023).

Despite its success, RL remains highly sensitive to noisy or unstable training environments, where perturbations in observations or transitions can disrupt learning dynamics. Such noise often causes instability and catastrophic forgetting, with agents overwriting previously learned policies (Gu, Jia, and Choset, 2019). These vulnerabilities limit the applicability of RL to real-world systems, where noise is the norm rather than the exception.

To address these challenges, we propose a novel uncertainty-aware RL framework that leverages epistemic uncertainty to stabilise the learning process and enable efficient training even under severe noise. The core of our approach lies in estimating uncertainty of the states. In our formulation, uncertainty in a given state reflects the noise or perturbations in the observed state, e.g. inherent noise in real-world environments.

In our study, these uncertainty estimates are used to adjust the RL agent’s value updates. Specifically, the higher the estimated uncertainty, the smaller are the corresponding updates. The underlying hypothesis is that noisy states lead an RL agent to misestimate the value of the state that has been perturbed with noise. This will lead to incorrect updates during learning (value or policy functions) such that the agent gets destabilized, forgetting all valuable states seen before (catastrophic forgetting).

Prior research has primarily exploited uncertainty to improve exploration (Lockwood and Si, 2022). In contrast, far less attention has been given to the role of uncertainty in *stabilizing* the learning process. Here, uncertainty is not used to drive exploration, but rather to help the agent detect when its value estimates are unreliable, preventing abrupt updates that may destabilize training.

Learning instability is a well-known challenge in deep RL: noisy or perturbed observations can lead to highly fluctuating value estimates, causing overshooting in temporal-difference updates, divergence of Q-values, and in severe cases catastrophic forgetting of previously learned behaviour. Ensuring stability is therefore

essential, as unstable learning dynamics often result in poor sample efficiency, failure to converge, or highly inconsistent policies.

Efforts to address stability have often relied on deep ensembles (Lee et al., 2021; Charpentier et al., 2022). However, deep ensemble approaches suffer from significant drawbacks. They require multiple forward passes and the training, evaluation, and inference of numerous neural networks. This becomes computationally very expensive as the ensemble size grows. Other RL methods with uncertainty, such as Bootstrapped DQN, also exist. However, they often suffer from correlation among ensemble heads, which reduces diversity and limits the reliability of their uncertainty estimates (Charpentier et al., 2022). More efficient approaches like Monte Carlo (MC)-Dropout have been proposed, but they under perform when compared to deep ensembles (Durasov et al., 2021; Charpentier et al., 2022).

Our approach, described in Section 4.3, addresses these challenges by stabilizing the learning of deep RL(DRL) agents in noisy environments. Specifically, we adopt an architecture known as Masksembles (Durasov et al., 2021). Masksembles is an efficient uncertainty estimation (UE) neural network architecture that trains a single neural network using multiple parameterized binary masks (Durasov et al., 2021). By controlling the overlap between these masks, the method effectively creates an ensemble of diverse, uncorrelated sub-models within a single network, combining the benefits of deep ensembles and MC-Dropout. This method serves as a drop-in replacement, achieving performance on par with computationally expensive deep ensembles while maintaining the efficiency of MC-Dropout (Durasov et al., 2021). Building on this foundation, we introduce **MASURE** (*Masksembles for Stable and Uncertainty-aware Reinforcement Learning Environments*), a novel architecture designed for uncertainty-conscious value update to stabilize learning under extreme noisy conditions. MASURE extends the Masksembles framework to RL, reducing computational overhead while improving robustness in both online and offline environments.

To evaluate the performance of our proposed method, we conducted two sets of experiments, ensuring generalizability:

- **Online setting:** We employed standard RL benchmarks, including *CartPole-v1*, *MountainCar-v0*, and *LunarLander-v2*. These environments are widely used in RL research because their simplicity and controlled dynamics make them well-suited for studying the impact of noise and the stability of learning. The choice of environments is inspired from recent state of the art work on uncertainty in RL (Mai, Mani, and Paull, 2022) and builds on prior studies (Charpentier et al., 2022; Gadgil, Xin, and Xu, 2020; Gal, McAllister, and Rasmussen, 2016).
- **Offline setting:** We evaluated on proprietary real-world churn datasets from the insurance domain, where noise is inherently present in customer data.

We compared our approach against (i) popular baseline algorithms without explicit UE (DQN, BootstrapDQN (Osband et al., 2016)) and (ii) UE-based baselines

(SunriseDQN, IVDQN (Lee et al., 2021; Mai, Mani, and Paull, 2022)). Performance was assessed using average episodic return in the online tasks and balanced accuracy (BA) in the churn prediction task. Across both settings, MASURE consistently outperformed the baselines under noisy conditions, achieving higher and more stable returns in online benchmarks and the best BA in offline churn prediction (64.3% vs. 63.5% for DQN, 63.8% for BootstrapDQN, 61.9% for SunriseDQN, and 62.0% for IVDQN).

4.2 Literature study

Effective decision-making in real-world applications is often challenged by uncertainty and noise. The literature distinguishes between two main types of uncertainty in RL: aleatoric (inherent noise in the environment) and epistemic (uncertainty due to limited data/knowledge). The authors in Charpentier et al. (2022) introduce a framework for disentangling aleatoric and epistemic uncertainty in RL, proposing four desiderata to capture the desired behaviour for uncertainty estimates (UE) in RL setup. The desiderata cover aleatoric and epistemic uncertainty at both training and testing time for UE in RL models (e.g., MC-dropout, ensemble, evidential networks) inspired by supervised learning to instantiate these uncertainties. This distinction is crucial for improving exploration and stabilizing learning.

Several works focus on improving exploration efficiency by leveraging uncertainty estimates. For epistemic uncertainty and exploration, Liu et al. (2021) propose an approach to improve exploration efficiency for distributional RL inspired by Bayesian Deep Learning, using estimated epistemic uncertainty derived from ensembles of the quantile function network. Similarly, the study by Lee et al. (2021), is closely related to our work, proposes reweighting sample transitions based on uncertainty estimates from a Q-ensemble. They use an Upper-Confidence Bound (UCB) based on the mean and variance of Q-functions to select actions for efficient exploration. The work by An et al. (2021) also points out gaps in ensemble methods and proposes an ensemble-diversified actor-critic algorithm for offline RL that leverages the generalization ability of deep neural networks and clipped Q-learning to penalize Out-of-Distribution (OOD) data points with high prediction uncertainties. The work by authors of Mai, Mani, and Paull (2022) introduces Inverse-Variance RL (IV-RL), a Bayesian framework combining probabilistic ensembles and Batch Inverse Variance weighting, which is currently a state-of-the-art method that we compare against.

Approaches for handling aleatoric uncertainty and noise are also prevalent. The authors of Mavor-Parker et al. (2022) addresses the issue where curiosity-driven exploration fails due to action-dependent noise sources ('noisy TVs') by proposing Aleatoric Mapping Agents (AMAs) that generate separate forward predictions for the mean and aleatoric uncertainty of future states to reduce intrinsic rewards for unpredictable transitions. The G-learning algorithm proposed by Fox, Pakman, and

Tishby (2016) tackles the slow learning problem in early stage Q-learning in noisy environments by regularizing the value estimates by penalizing deterministic policies.

Other approaches include Fortunato et al. (2018), which proposes Noisy Nets as an extension for various Deep RL (DRL) algorithms (like DQN, A3C) to yield better performance through efficient exploration. Furthermore, the authors in Flet-Berliac and Preux (2020) proposes filtering out noisy samples during training by employing the variance of the samples. More complex techniques include Li et al. (2024) use of Reward Machines for Deep RL in noisy and uncertain environments to leverage task structure and prior knowledge. The work by Sachdeva et al. (2024) introduces a novel DRL algorithm for uncertainty and noise-aware decision-making using Bayesian Neural Networks and skew-geometric Jensen-Shannon divergence. Investigating the statistical properties, the authors in Zhu, Dong, and Lam (2024) study the central limit theorem behaviours of estimated Q-values and value functions to construct asymptotically valid confidence regions. Finally, the authors in Luis et al. (2023) proposes recursively learning the uncertainty of the Bellman equation to solve for an over-approximation that may lead to inefficient exploration.

4.3 Methodology

Our approach builds upon recent advances in UE-based RL. In particular, we integrate uncertainty-based sample reweighting, exploration strategies, and reward shaping, drawing inspiration from prior work such as Lee et al. (2021), Mai, Mani, and Paull (2022), and Lin, Chen, and Qi (2020).

4.3.1 Masksembles

For UE, a wide range of methods have been proposed in machine learning and RL like Monte Carlo (MC) dropout (Gal and Ghahramani, 2016; Charpentier et al., 2022), deep ensembles (Lakshminarayanan, Pritzel, and Blundell, 2017; Charpentier et al., 2022), deep kernel learning (Charpentier et al., 2022; Maddox et al., 2019), and evidential networks (Malinin and Gales, 2018; Charpentier et al., 2022).

Deep ensembles estimate predictions by averaging outputs from independently trained models, where aleatoric uncertainty is captured by the average of predicted variances σ_i^2 and epistemic uncertainty from the variance of mean predictions μ_i . Although highly effective, this approach is computationally expensive as training, evaluation, and inference scale with the number of ensemble members (Durasov et al., 2021; Charpentier et al., 2022). In contrast, MC-Dropout provides a cheaper approximation by running a single trained network K times with different random dropout masks, combining the outputs to compute the mean prediction, aleatoric uncertainty from predicted variances σ_k^2 , and epistemic uncertainty from the variance of mean predictions μ_k . Despite their computational efficiency, MC-Dropout

often underperforms relative to deep ensembles, since simply adding randomness does not always yield diverse predictions (Durasov et al., 2021).

To overcome the above mentioned gaps, Masksembles (Durasov et al., 2021) were proposed as a new architecture combining the strengths of both approaches. Instead of random dropout, Masksembles employ a fixed, predefined set of binary masks with controlled overlap. This reduces correlation across sub-models, ensures diversity, and yields uncertainty estimates that match deep ensembles at significantly lower computational cost (Table 4.1)(Durasov et al., 2021). We leverage these properties in MASURE, where the variance across Masksemble heads quantifies epistemic uncertainty to stabilize learning under noise. MASURE is instantiated on a Q-learning backbone (Section 4.3.2) and formalized through an uncertainty-conscious update rule (Section 4.3.3).

TABLE 4.1: Compute efficiency comparison of MASURE and ensemble-based baselines on LunarLander-v2.

Model	Params	Inference FLOPs	Training FLOPs
DQN	5.00K	9.86K	24.64K
SunriseDQN (Deep Ensemble)	24.98K	49.28K	123.20K
IV-DQN (Deep Ensemble)	26.28K	51.84K	129.60K
MASURE (Ours)	6.02K	36.69K	91.72K

4.3.2 Q-Learning and Contextual Bandits

We denote the state by s (possibly vector-valued), the action by a (discrete), and write $Q(s, a)$ for their state-action value.

The standard Q-learning update is defined as:

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha \cdot \delta_t \quad (4.1)$$

where δ_t denoting the temporal difference(TD), defined as:

$$\delta_t = r_t + \gamma \cdot \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) \quad (4.2)$$

and where α is the learning rate, r_t is the reward received at time t , $\gamma \in [0, 1)$ is the discount factor, and s_{t+1} denotes the next state. In a one-step contextual bandit setting, there are no transitions ($\gamma = 0$). The update therefore simplifies to:

$$Q_{t+1}(s_t, a_t) \leftarrow Q_t(s_t, a_t) + \alpha (r_t - Q_t(s_t, a_t)). \quad (4.3)$$

We frame the customer churn prediction problem as a contextual bandit. At each time step, the agent observes a context vector x_t (customer profile and behaviour) and chooses an action a_t from a discrete set (e.g., “Churn,” “Not Churn”). The agent then receives an immediate reward r_t , representing customer retention or churn.

The goal is to learn a policy $\pi(a|x)$ that maximizes total expected reward, thereby classifying churn.

While these Q-learning provide the learning backbone, they do not account for the destabilizing effect of noise. We therefore propose MASURE, which integrates Masksemble-based uncertainty estimation directly into the Q-learning update. The key idea is to modulate the update magnitude according to epistemic uncertainty, thereby dampening unstable updates in noisy states.

4.3.3 MASURE: Masksembles for Stable and Uncertainty-aware RL Environments

We formalize MASURE with an uncertainty-conscious update designed to stabilize Q-learning under consecutive noisy states by scaling the temporal difference step inversely with epistemic uncertainty. Real-world data is inherently noisy, and such noise induces epistemic uncertainty that can destabilize learning. MASURE employs DQN (Mnih et al., 2015) as its backbone, estimating epistemic uncertainty from variance to adaptively scale Q-value updates. We use an ensemble of n Masksemble subnetworks, where each mask $i \in \{1, \dots, n\}$ defines a sub-model that produces a Q-value estimate $Q_t^{(i)}(s_t, a_t)$. The ensemble mean estimate is:

$$\hat{Q}_t(s_t, a_t) = \frac{1}{n} \sum_{i=1}^n Q_t^{(i)}(s_t, a_t) \quad (4.4)$$

and the epistemic uncertainty is given by the empirical variance across subnetworks:

$$\sigma_t^2(s_t, a_t) = \frac{1}{n} \sum_{i=1}^n (Q_t^{(i)}(s_t, a_t) - \hat{Q}_t(s_t, a_t))^2. \quad (4.5)$$

The Bellman target is defined as:

$$y_t = r_t + \gamma \cdot \max_{a'} \hat{Q}_t(s_{t+1}, a') \quad (4.6)$$

with temporal difference error:

$$\delta_t = y_t - \hat{Q}_t(s_t, a_t). \quad (4.7)$$

Finally, our proposed uncertainty-conscious Q-value update is:

$$\hat{Q}_{t+1}(s_t, a_t) \leftarrow \hat{Q}_t(s_t, a_t) + \frac{1}{1 + \sigma_t^2(s_t, a_t)} (\alpha \cdot \delta_t) \quad (4.8)$$

where α is the learning rate.

The update is made uncertainty-conscious by scaling it inversely with the epistemic variance $\sigma^2(s, a)$, so that higher uncertainty results in smaller effective updates to $\hat{Q}(s, a)$. This mechanism prevents large, unstable changes during consecutive noisy states and improves stability in both contextual bandits and full RL

environments. The $1/(1 + \sigma^2(s, a))$ is chosen for two reasons: (i) it decreases the update magnitude smoothly with increasing uncertainty, thereby dampening instability, and (ii) the denominator guarantees bounded, well-behaved updates without introducing additional hyperparameters. We adopt this simple inverse relation for its ease of implementation, and empirical robustness across tasks without extra tuning.

4.3.4 Experimental Setup

TABLE 4.2: Experimental setup, noise schedule, and hyperparameters.

Category	Parameters
Training	400 episodes; replay buffer size = 100k; ϵ -greedy exploration ($\epsilon_{\text{start}} = 1.0$, $\epsilon_{\text{end}} = 0.01$).
Noise schedule	No noise for episodes 0–99; burst probability = 0.01 per step; burst length = 500 steps; cool-down period = 20k steps; Gaussian noise $\mathcal{N}(0, 1)$.
Hyperparameters	DQN: batch = 128, learning rate = 0.0005, $\gamma = 0.99$, $\tau = 0.005$, ϵ-decay = 0.99. BootstrapDQN: batch = 64, lr = 0.0005, mask prob. = 0.9, prior = 10, Bernoulli masks. SunriseDQN: batch = 64, lr = 0.0005, mask prob. = 0.9, prior = 10, ensemble size = 5. IV-DQN: batch = 64, lr = 0.0005, $\gamma = 0.99$, $\tau = 0.005$, dynamic ϵ, min batch = 48, mask prob. = 0.5, ensemble size = 5. Masksemble (Ours): masks = 4, scale = 2, hidden layers = [64, 64], mask combine = average, lr = 0.0005, $\gamma = 0.99$, $\tau = 0.005$, ϵ-decay = 0.99, batch = 128.

We evaluate our method in both online and offline RL settings, testing robustness under controlled noise in simulations and validating real-world applicability on a large-scale industrial dataset.

Online RL: To destabilize learning, we inject structured observation noise using a StepBurstNoiseObservation wrapper. This wrapper is an extension of proposed RandomNormalNoisyObservation from Khraishi and Okhrati (2023) that adds Gaussian noise to the observation for multiple steps with a probability p . Unlike prior work Chen et al. (2022), Fox, Pakman, and Tishby (2016), Sun et al. (2022), and Zhang et al. (2020), which primarily examines mild or short-lived perturbations, our design introduces sustained and severe noise to deliberately destabilize the agent, thereby testing its ability to maintain stable learning. In the burst-style noise schedule, training begins with 100 warm-up episodes without noise. Afterwards, Gaussian noise $\mathcal{N}(0, 1)$ is injected with probability $p = 0.01$ per step. When triggered, a burst lasts 500 steps (potentially crossing episode boundaries, with episodes capped at 1000 steps) and is followed by a 20,000-step cool-down before another burst can occur. The noise amplitude was fixed at 1.0 across all environments, producing rare but

severe perturbations. Even under noise, states are clipped to the valid ranges defined by the Gym environments.

Baselines Reproducibility. This setup creates rare but severe perturbations, in contrast to prior work that typically assumes mild noise, thereby testing whether agents can recover and maintain stability. To validate our framework, we ran experiments on three standard OpenAI Gym environments: *CartPole-v1*, *MountainCar-v0*, and *LunarLander-v2*. The environments were chosen for their simplicity, interpretability, suitability based on previous research (Mai, Mani, and Paull, 2022; Charpentier et al., 2022) for studying noise effects and learning stability.

CartPole-v1 is a classic control task where an agent must balance an inverted pendulum mounted on a moving cart. The environment is fully observable, low-

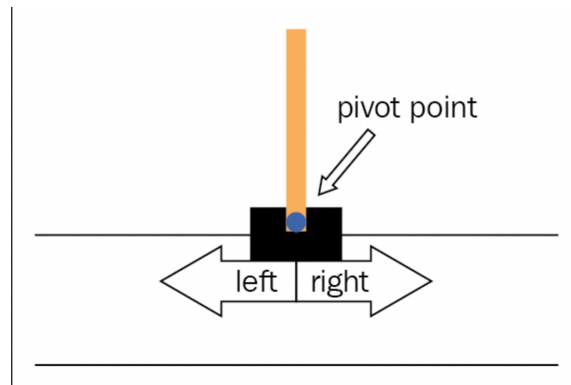


FIGURE 4.1: The CartPole environment used in our experiments. The agent must balance a pole atop a moving cart by applying left or right forces.

dimensional, and highly sensitive to perturbations, making it suitable for studying stability and recovery under sudden observation noise.

MountainCar-v0 is a sparse-reward environment in which a low-powered car must build momentum by oscillating between two hills to reach a goal state. Its

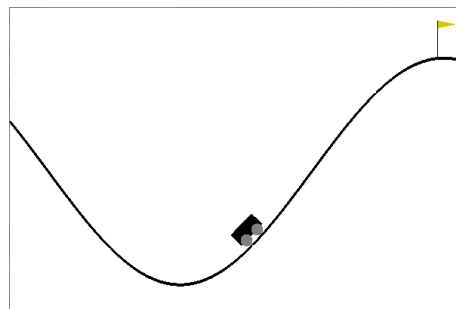


FIGURE 4.2: The MountainCar environment used in our experiments. An agent must learn to build momentum in an underpowered car by driving back and forth to reach the hilltop goal.

inherent difficulty, long horizons, and sensitivity to noisy transitions make it an informative testbed for robustness analysis.

LunarLander-v2 features a higher-dimensional state space and a more complex, multi-modal reward structure. The agent controls a spacecraft’s descent using dis-

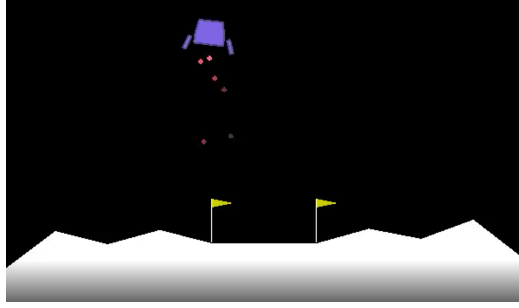


FIGURE 4.3: The LunarLander-v2 environment, where the agent learns to control a lander using thrusters to touch down safely on a designated pad.

crete thrusters, requiring stable control, precise state estimation, and resilience to abrupt perturbations.

Together, these environments cover a gradient of complexity; from simple dynamics and dense rewards to harder tasks with sparse, delayed feedback, allowing us to systematically assess the robustness of UARL methods under burst-style noise.

Offline RL: We also evaluate our method in an offline RL setting using proprietary churn data described version 2 in Section 2.5.2. The data had been preprocessed, cleaned and transformed (dealing with missing values, removing null values, dealing with outliers) by our partner given their domain knowledge yielding a 40-dimensional context representation. At each step, the agent observes a context vector x_t , selects an action $a_t \in \{\text{Churn}, \text{Not Churn}\}$, and receives an immediate reward r_t corresponding to retention or churn. This contextual bandit formulation naturally incorporates the inherent noise present in real-world data. Model performance is evaluated using BA, a popular metric for imbalanced datasets.

Inspired from Lin, Chen, and Qi (2020), the dataset was split into 70% training and 15% testing and 15% validation using a random stratified split to preserve class proportions. Each row corresponds to a single policy, and no policy record leaks across splits. Due to Non disclosure agreement restrictions, we cannot disclose the exact churn ratio. However, as discussed previously the proportion of churners is substantially lower than 10%.

In imbalanced datasets, minority samples are harder to classify, so the reward function emphasizes their importance by assigning higher costs. The agent receives $+1/-1$ for correct/incorrect predictions on minority samples and $\lambda/-\lambda$ on majority samples, with $\lambda \in [0, 1]$:

$$R(s_t, a_t, l_t) = \begin{cases} +1, & a_t = l_t \text{ and } s_t \in D_P, \\ -1, & a_t \neq l_t \text{ and } s_t \in D_P, \\ \lambda, & a_t = l_t \text{ and } s_t \in D_N, \\ -\lambda, & a_t \neq l_t \text{ and } s_t \in D_N, \end{cases}$$

TABLE 4.3: Balanced accuracy (BA) (%) of RL agents for churn prediction (offline learning).

Agent	Balanced Accuracy
DQN	63.52 ± 1.2
BootstrapDQN	63.78 ± 0.9
SunriseDQN	61.88 ± 1.5
IVDQN	62.00 ± 1.3
MASURE (Ours)	64.30 ± 1.8

where D_P and D_N denote the minority and majority sets, s_t is the state at time t , a_t the agent’s action (predicted label), and l_t the true class label. For $\lambda < 1$, minority errors carry higher cost. The best performance is observed when $\lambda = \rho = \frac{|D_P|}{|D_N|}$. When data is balanced, we see that $\lambda = 1$ and costs are equal.

For both online and offline settings, we compare against standard baselines like DQN (Mnih et al., 2015), BootstrapDQN (Osband et al., 2016) and state-of-the-art UE-based methods SunriseDQN (Lee et al., 2021), IVDQN (Mai, Mani, and Paull, 2022).

Table 4.2 presents the main hyperparameters of the various models. Intuitively, increasing the number of masks in a relatively small network increases the overlap between active neurons across subnetworks. This can reduce the effective diversity of the ensemble and may reintroduce the correlation in BootstrapDQN, where multiple heads learn similar representations despite independence.

For online RL benchmarks, performance is reported as the average cumulative reward over the last 50 evaluation episodes, averaged across 8 random seeds. For the offline task, performance is measured using BA on the test set, also averaged across 8 random seeds. All experiments were run on an RTX 3080Ti GPU with an Intel i7 CPU and 16 GB RAM. For more details, please refer to the article and the publicly available code repository of (Mai, Mani, and Paull, 2022).

4.4 Results

We evaluate the impact of noisy observations in benchmark environments. As shown in Figure 4.4, which compares IVDQN and MASURE under noise-free and noisy settings, the SOTA UE-based IVDQN’s performance degrades under burst-style noise injection, whereas MASURE maintains stable learning dynamics, demonstrating robustness to perturbations. In Figure 4.5 we present the performance comparison in all 3 environments. MASURE outperforms all baselines in LunarLander and Cart-Pole, providing clear evidence of improved stability under extreme noise.

In the MountainCar NoisyEnv, MASURE exhibits a similar limitation to IVDQN, struggling to exploit exploration effectively. This outcome is expected, as MountainCar provides sparse rewards with limited feedback in most states. The proposed uncertainty-conscious update scales the TD step by $\frac{1}{1+\sigma^2}$, ensuring stability under

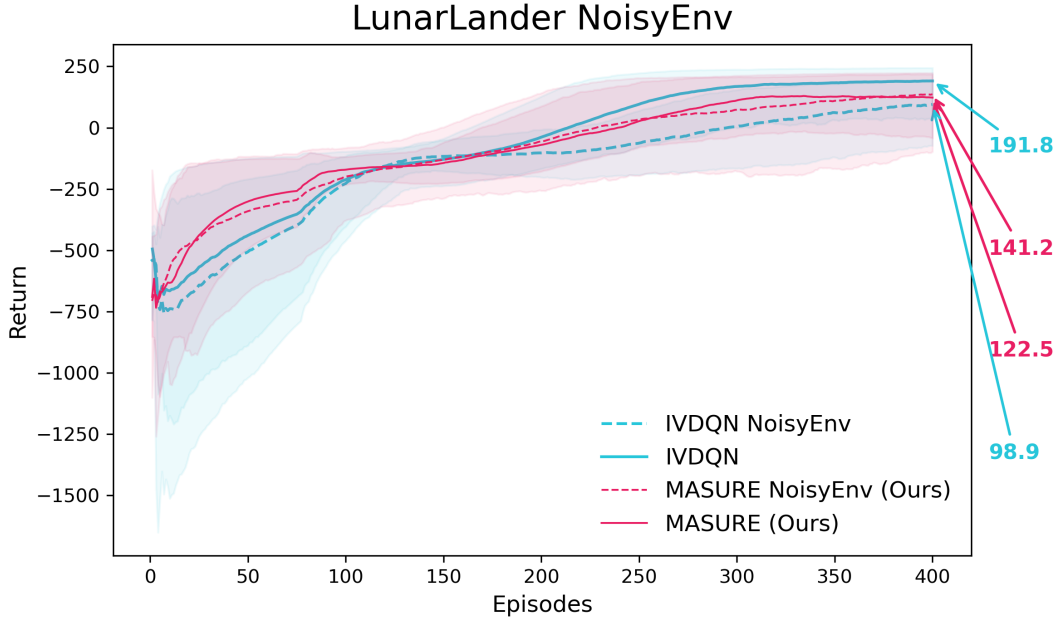


FIGURE 4.4: Comparison of SOTA IVDQN and MASURE in noisy environments, showing improved stability of MASURE.

noisy observations. However, in sparse-reward environments such as Mountain-Car, states along unexplored trajectories tend to exhibit high epistemic uncertainty but low immediate reward. The resulting reduction in effective learning rate slows value propagation and biases the agent toward familiar states, thereby discouraging exploration. Under these conditions, BootstrapDQN performs best, while IVDQN’s weighted strategy does not consistently improve results. Nevertheless, MASURE still achieves better performance than IVDQN, confirming its robustness advantage even in exploration-heavy settings. Overall, while the proposed weighting scheme enhances robustness in noisy, dense-reward tasks, it may underperform in environments where exploration and long-horizon reward propagation are essential for success.

This limitation could be mitigated by coupling the uncertainty-weighted update with an exploration bonus, similar to the UCB-based mechanism in SunriseDQN (Lee et al., 2021). Alternatively, intrinsic motivated approach to stimulate Policy Exploration (Bagot, Mets, and Latré, 2020) could be integrated to encourage exploration as well.

For the offline setting, we evaluate on a real-world churn prediction dataset, where we assume inherent noise is present in all features and therefore do not inject additional perturbations. As summarized in Table 4.3, MASURE achieves the highest BA (64.30%), outperforming DQN (63.52%), BootstrapDQN (63.78%), SunriseDQN (61.88%), and IVDQN (62.00%). MASURE significantly enhances stability in NoisyEnv and generalizes effectively to the inherently noisy real-world offline data. Thus, MASURE provides a stable and accurate approach for UE offline learning on real-world datasets.

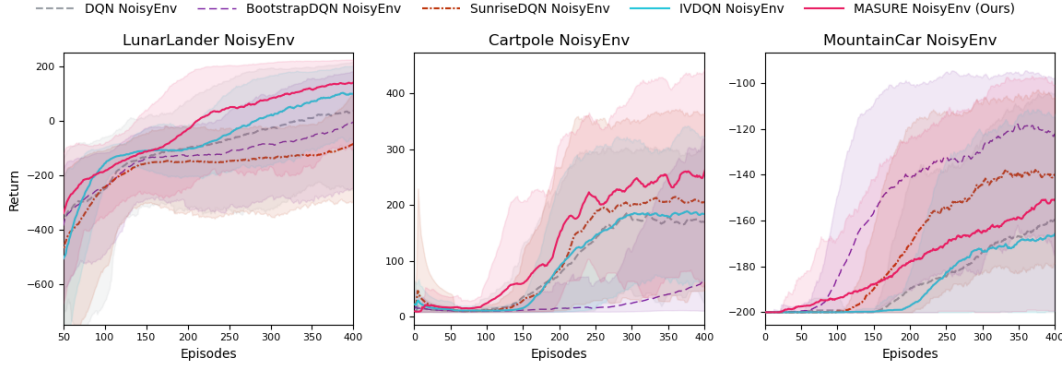


FIGURE 4.5: Comparison of performance of baseline agents and MASURE in noisy environments, showing improved stability of MASURE.

To statistically assess whether MASURE provides a performance improvement over the baseline RL agents, we applied the Friedman test, a non-parametric test designed for comparing multiple algorithms over repeated evaluations. The null

TABLE 4.4: Average ranks of RL agents under the Friedman test (lower is better).

Agent	Average Rank
MASURE (Ours)	1.25
BootstrapDQN	2.38
DQN	2.75
SunriseDQN	4.13
IVDQN	4.50

hypothesis(H_0) is defined as:

H_0 : There is no significant difference in performance between MASURE and baseline RL agents.

The test was conducted using the balanced accuracy obtained across eight independent runs for each agent.

The Friedman statistic is computed using:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_{j=1}^k R_j^2 \right] - 3N(k+1),$$

where N is the number of runs, k is the number of algorithms, and R_j is the average rank of algorithm j , according to balanced accuracy. The ranks are presented in Table 4.4 Using the balanced accuracy results obtained across eight runs ($N = 8$ and $k = 5$), we obtain:

$$p = 0.00042.$$

Since $p < 0.05$, the null hypothesis is rejected with 95% confidence, indicating that the differences in performance among the agents are statistically significant.

The average ranks, reported in Table 4.4, show that MASURE achieves the lowest (best) rank, confirming its higher performance.

4.5 Conclusion

This work introduced MASURE, a Masksembles-based architecture for UE-based RL. Our results demonstrate that weighing updates with uncertainty provides clear added value: the proposed method not only stabilizes training under burst-style noise but also achieves superior predictive performance on a real-world churn prediction task. Compared to standard RL agents and state-of-the-art UE-based baselines such as IVDQN (Mai, Mani, and Paull, 2022), MASURE consistently delivers improved stability and robustness in both offline and online settings.

In simulated OpenAI Gym environments, baseline agents suffered substantial drop in performance when trained with observation noise, while MASURE maintained stable learning dynamics and achieved higher cumulative rewards. In the offline churn prediction task, MASURE attained the best BA, highlighting its practical value in industrial applications where feature noise is inherent. Importantly, by reducing the number of networks required relative to deep ensembles, MASURE addresses the significant computational overhead typically associated with ensemble-based u_e . The core contribution of MASURE is the formalization of *uncertainty-conscious Q-value updates* via variance across Masksemble heads, ensuring stable learning even under consecutive noisy states. Although we employ a Q-learning backbone for MASURE, the same principle of uncertainty-weighted updates can be extended to actor-critic methods such as SAC-N (An et al., 2021), where uncertainty is estimated from an ensemble.

Overall, the contributions of this work are threefold: (i) improved predictive performance in real-world offline environments, (ii) enhanced robustness in noisy online settings, and (iii) more efficient uncertainty quantification without the heavy compute of deep ensembles. By reducing the number of networks required relative to deep ensembles, MASURE directly addresses the substantial computational overhead typically associated with ensemble-based UE. As shown by the authors in Durasov et al. (2021), Masksembles deliver ensemble-level performance and UE at nearly the cost of a single network, avoiding the need to train and evaluate multiple independent models.

A key challenge remains to improve exploration in sparse-reward environments, where both MASURE and IVDQN underperform. Future work could also extend MASURE by explicitly disentangling aleatoric and epistemic uncertainty, allowing agents to treat state noise and model uncertainty differently. Scaling MASURE to high-dimensional continuous-control tasks (e.g., MuJoCo) would provide further evidence of robustness in more complex domains.

Operational implications for sequential decision-making under noise. Many business decision problems are sequential: actions taken today influence future observations, available options, and measured outcomes, and errors can compound through feedback loops. In such settings, uncertainty is useful not only as a reporting layer, but also as a mechanism to *stabilize learning* and reduce brittle policy updates when observations are noisy, perturbed, or subject to distributional shift. The proposed uncertainty-aware Bellman updates based on Masksembles provide a practical compromise for deployment: they approximate ensemble-style epistemic signals while requiring fewer trained parameters than classical ensembles, making repeated training and evaluation more feasible under industrial compute constraints. From an application perspective, this supports safer policy improvement by modulating updates and exploration according to uncertainty, and by enabling conservative operation when the agent encounters unfamiliar states. These considerations also highlight that real-world adoption depends on end-to-end design choices; data interfaces, monitoring, and decision thresholds; which motivates consolidating the practical guidance and limitations across chapters in the concluding chapter.

Chapter 5

Conclusion & Future works

This chapter brings together the managerial and scientific contributions of the thesis. It shows how uncertainty estimates can support practical decision-making through human-in-the-loop workflows and resource prioritisation. In addition, this chapter summarises the empirical findings across Belief Functions, Heterogeneous Ensemble for Uncertainty Quantification, and uncertainty-aware Reinforcement Learning. The chapter concludes by answering the research questions, highlighting limitations, and outlining future directions for advancing uncertainty-aware models in real-world decision environments.

Contents

5.1 Introduction	85
5.2 Answers to the Research Questions	86
5.3 Human-in-the-loop workflow for uncertainty-aware decision support	87
5.3.1 Sub-portfolio optimization: Churn management	88
5.4 Synthesis of Key Empirical Insights	92
5.5 Limitations	93
5.6 Future Work	93
5.6.1 From prediction to “predict-then-optimize” with uncertainty	94

5.1 Introduction

This thesis investigated how uncertainty quantification (UQ) can enhance the reliability, interpretability, and practical relevance of machine learning and reinforcement learning systems for real-world decision-making. It studied three approaches: Belief Functions and Conformal Prediction (Chapter 2), Heterogeneous Ensemble for Uncertainty Quantification (Chapter 3), and uncertainty-aware reinforcement learning via Masksembles (MASURE, Chapter 4). It also examined (Section 5.3) how epistemic and aleatoric uncertainty can be estimated, decomposed, and operationalised in business applications such as churn prediction, fraud detection, and portfolio prioritisation.

This conclusion chapter answers the research questions, synthesises the main empirical insights, and identifies limitations and avenues for future work.

5.2 Answers to the Research Questions

RQ1 (Belief Functions and Conformal Prediction)

For what kind of data do belief-function methods provide reliable and actionable uncertainty estimates, and how well do they scale to noisy, imbalanced business datasets such as operational churn prediction?

Our results show that the performance of BF and CP is constrained by the structure of the underlying data. On the churn dataset, which is severely imbalanced and shows substantial feature overlap between churn and non-churn profiles, BF produced highly diffuse mass assignments and CP generated large prediction sets that often contained both labels. Although conceptually distinct, both methods converged toward the same behaviour: broad, non-discriminative uncertainty estimates for real-world data. In contrast, CP achieved near-perfect coverage on the credit card fraud dataset, where class-separating features were stronger. These findings indicate that BF and CP remain theoretically sound but are limited in practice when discriminative evidence is weak. Thus, the primary bottleneck is not the uncertainty formalism itself but the insufficient discriminative features in the data.

RQ2 (Supervised ML and Heterogeneous Ensembles)

How can heterogeneous ensembles combining different learning paradigms improve epistemic uncertainty estimation compared to homogeneous ensembles, and what is the impact of feature transformations on the quality and stability of uncertainty estimates?

Chapter 3 demonstrated that HEUQ provides competitive or superior predictive accuracy relative to strong baseline classifiers while delivering substantially better calibration, as reflected in consistently lower negative log-likelihood and Brier scores. Importantly, HEUQ remained robust under PCA transformations and random Gaussian projections, highlighting the value of architectural diversity in absorbing changes in feature representation. The Ablation study and analyses confirmed that removing learners reduces calibration quality and increases predictive uncertainty, underscoring the central role of heterogeneity in capturing epistemic uncertainty that is meaningful. Homogeneous ensembles, by contrast, often produced epistemic uncertainty that was not informative. These findings support the proposition that heterogeneous ensembles broaden hypothesis coverage and lead to more informed uncertainty estimates.

RQ3 (Reinforcement Learning and MASURE)

How can we design lightweight, uncertainty-aware reinforcement learning architectures that are computationally efficient, robust to extreme observation noise, and applicable to real-world business decision problems such as churn management?

Chapter 4 introduced MASURE, the first RL framework integrating Masksembles to obtain epistemic uncertainty at nearly the cost of a single network. MASURE’s uncertainty-conscious Bellman updates yielded substantial robustness under burst-style and persistent observation noise. Across OpenAI Gym environments, MASURE maintained stable learning where baselines suffered pronounced degradation. In the offline churn prediction task, MASURE achieved the highest balanced accuracy, illustrating its suitability for industrial environments with noisy features. By reducing the number of networks required relative to deep ensembles, MASURE provides an efficient alternative for large-scale RL systems while retaining high-quality uncertainty estimates. The framework therefore constitutes a practical and computationally lightweight approach to uncertainty-aware reinforcement learning.

5.3 Human-in-the-loop workflow for uncertainty-aware decision support

This section explains why uncertainty quantification matters for practice and how it can be embedded in a human-in-the-loop workflow. The aim is to operationalise uncertainty as a review signal: it indicates when an automated prediction is reliable enough for direct action and when additional review is warranted. In this workflow, domain experts in the company (e.g., retention managers for churn, or fraud analysts for fraud) set an *uncertainty threshold* that matches their risk tolerance and review capacity. Given an input x , the model outputs a prediction and an uncertainty estimate; cases whose uncertainty exceeds the threshold are treated as *high-uncertainty* and are therefore flagged for further processing. This threshold acts as a practical filter: it separates routine cases from cases where the system is unsure. High-uncertainty cases can then be handled using additional processes, such as manual review, an expert rule-based check, or the collection of additional information.

We illustrate the workflow with two use cases (Figure 5.1). We keep the discussion here at the level of uncertainty-based routing. The detailed prescriptive formulation for portfolio selection is provided later in Section 5.3.1.

- **Prioritization for resources use case: sub-portfolio optimization.** Uncertainty flags cases where the prediction is least reliable. A decision layer can then prioritise human attention toward cases that are both *high-uncertainty* and *high-impact* (e.g., high customer value), under limited resources. Subsection 5.3.1 presents the full sub-portfolio optimisation formulation; the role of this section is only to motivate the review step.

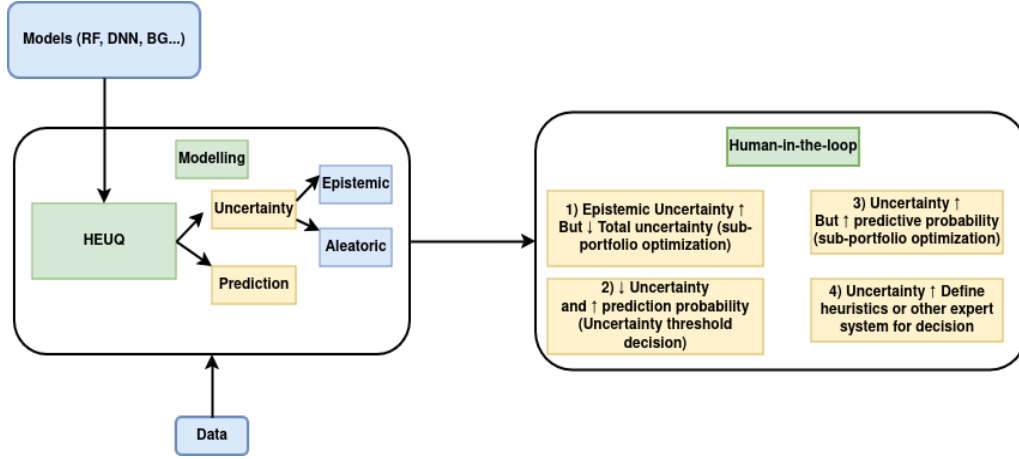


FIGURE 5.1: Human-in-the-loop workflow using uncertainty estimates (illustrated with HEUQ).

- **Anomaly Detection use case: Fraud detection.** In fraud detection, some cases may not be classified as fraud but still show high uncertainty. Such cases can be routed to further investigation to reduce missed fraud, while keeping review effort focused on the most ambiguous cases.

5.3.1 Sub-portfolio optimization: Churn management

We now present the management use case of our study of uncertainty quantification with ensembles in a practical setting of churn management, and customer relationship management in general.

Decision setting and constraints. In churn management on very large customer datasets, predictions are only useful if they support action under practical constraints. Marketing teams typically face limited contact capacity, limited budgets for incentives, and limited human resources for case-by-case review. As a result, the operational goal is not to act on every predicted churner, but to construct a sub-portfolio of customers for whom intervention is expected to be beneficial and sufficiently reliable. This makes uncertainty central: *wrong but highly confident* predictions waste budget (false positives), while missed churners can incur high opportunity costs (false negatives).

We will consider individual profiles x and examine the behaviour of total uncertainty, $u_t(x)$, epistemic uncertainty $u_e(x)$ and aleatoric uncertainty $u_a(x)$.

These three components were formally introduced and analysed in Chapter 3, and here we show how they can inform practical decision-making.

From uncertainty decomposition to actions: For each customer profile x , HEUQ provides a predicted churn probability $\hat{p}(x)$ and an uncertainty decomposition into $u_t(x)$, $u_e(x)$, and $u_a(x)$. We use these signals to define a simple operational policy with three actions: (i) *automate* a retention action when $\hat{p}(x)$ is high and total

uncertainty $u_t(x)$ is low, (ii) *defer to expert review* when $\hat{p}(x)$ is high but epistemic uncertainty $u_e(x)$ indicates strong model disagreement, and (iii) *monitor / no action* when predicted risk is low or when aleatoric uncertainty $u_a(x)$ suggests the outcome is dominated by irreducible noise. Total uncertainty $u_t(x)$ is used as a conservative criterion for automation, while the decomposition clarifies whether uncertainty is reducible (u_e) or irreducible (u_a), which supports different operational responses. This policy supports human-in-the-loop workflows under limited review capacity.

Optimised sub-portfolio: Precision metric optimization An important aspect of management is to predict, with reasonable precision, the probability of customers, x , liable to churn.

Precision score in function of (U_a, U_e)

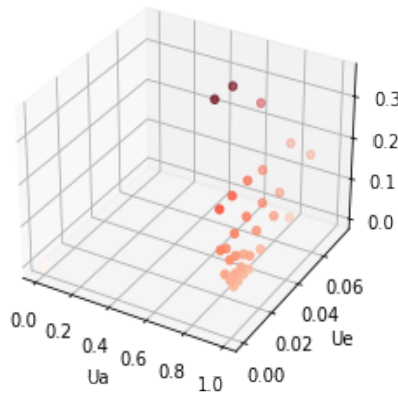


FIGURE 5.2: Precision score, in function of classes defined from crossing of quantile bins of the private data.

Given the customer features x and domain-specific business parameters, the marketing manager can assign a “profile value” $V(x)$ and choose an intervention strategy (e.g., offering an additional discount). To evaluate the cost and benefit of this strategy under limited budget and contact capacity, the manager selects a sub-portfolio of customers for which the predicted churn risk is sufficiently reliable. In this context, a natural objective is to maximise **precision** in the selected sub-portfolio:

$$Precision = \frac{TP}{TP + FP}.$$

In churn prediction, precision measures the share of contacted customers (predicted churners) who truly churn. High precision therefore reduces wasted interventions and helps control incentive costs.

We focus on precision here because the sub-portfolio decision is asymmetric: interventions are applied only to predicted churners, so false positives directly translate into avoidable costs. Recall can be analysed in the same framework, but it primarily reflects missed churners, which depends strongly on budget and contact constraints.

In the previous chapter 3, we reported negative log-likelihood (NLL), the Brier score (BS), and balanced accuracy (BA) to assess predictive performance, calibration, and classification quality.

Figure 5.2 reports the precision score across classes formed by crossing quantile bins of the private dataset, and shows how precision varies with the corresponding levels of epistemic uncertainty $u_e(x)$ and aleatoric uncertainty $u_a(x)$.

1. HEUQ should achieve a lower precision score for high values of u_t or u_a .
2. Very high epistemic uncertainty (Equation (3.4)) indicates strong disagreement across the learners in HEUQ. In such cases, marketing decisions based on the predictions should be treated with caution and may require additional review. Conversely, very low epistemic uncertainty can occur when several participants learn similar hypotheses. In our experiments, this effect was most visible when the participating models were all tree-based, which likely reduced diversity and led to similar hypotheses.

To study the interaction between uncertainty components, we group customers into sub-portfolios by crossing quantile bins of $u_a(x)$ and $u_e(x)$ (Figure 5.2). Within fixed bins of aleatoric uncertainty, precision often increases for low-to-moderate values of epistemic uncertainty and then decreases for higher values, resulting in a concave pattern. Operationally, this pattern can guide sub-portfolio selection by favouring groups that combine acceptable epistemic uncertainty with sufficient precision, under a fixed review and incentive budget.

Optimised sub-portfolio: balanced accuracy (BA) optimisation. The precision analysis above shows that uncertainty can help identify sub-portfolios where interventions are less likely to be wasted. However, precision focuses on the contacted group only (predicted churners) and does not account for performance on non-churners. We therefore also report balanced accuracy (BA), which evaluates both classes and is robust to class imbalance.¹ This section shows how the HEUQ uncertainty decomposition can be used to define sub-portfolios with higher BA on the proprietary churn dataset (Version 2.5.2).

To build these sub-portfolios, we *combine* uncertainty sources by using them jointly as selection criteria: we discretise $u_e(x)$ into quantile bins and $u_t(x)$ into quantile bins, and then evaluate BA in each crossed bin (Table 5.1). Across several rows (fixed ranges of u_t), BA shows a concave pattern as u_e increases: it first increases,

¹Recall: $BA = (TPR + TNR)/2$, where $TPR = TP/(TP + FN)$ and $TNR = TN/(TN + FP)$.

u_e increasing $\rightarrow$	Balanced accuracy								
				80.01	78.58	82.51	84.64	79	
	63.89	74.35	69.24	69.72	70.84	73.24	71.55	70.3	
u_t increasing $\downarrow$	60.25	68.08	64.41	67.73	67.41	65.45	65.39	65.72	
	58.42	63	62.64	58.92	62.54	63.2	64.59	58.38	
	55.27	57.26	57	57.6	55.92	60.71	53.72	56.78	
	53.15	52.66	50.45	53.32	52.5	53.22	49.18	52.31	

TABLE 5.1: Balanced accuracy (BA) computed on sub-portfolios obtained by crossing quantile bins of epistemic uncertainty u_e (8 bins) and total uncertainty u_t (5 bins) on the proprietary churn dataset (Version 2.5.2). Each cell reports the BA for the subset of customers whose (u_e, u_t) values fall in the corresponding bin ranges.

reaches a peak, and then decreases. In addition, some rows show a sharp increase in the last u_e bin (see, e.g., the first row of Table 5.1).

We also report a simpler one-dimensional selection strategy where a sub-portfolio is defined by keeping only customers with uncertainty above a chosen quantile threshold. Let $u(x)$ denote one uncertainty score (e.g., u_e , u_t , or u_a) and let $q \in (0, 1)$ be a quantile level. Define the sub-portfolio

$$S(u, q) = \{x : u(x) \geq Q_u(q)\},$$

where $Q_u(q)$ is the empirical q -quantile of $u(x)$ computed over the full portfolio. If $q = 0.7$, then $S(u, 0.7)$ is the set of customers whose uncertainty values are in the top 30%. In Table 5.2, the column *Unc Minimal Bound* corresponds to this threshold value $Q_u(q)$.

Operationally, Table 5.2 is computed as follows on the proprietary churn dataset (Version 2.5.2): for each choice of uncertainty score u and each quantile level q , we select $S(u, q)$ using the threshold $Q_u(q)$ and then compute BA on that selected sub-set.

Beyond empirical performance, this section demonstrated how uncertainty scores can be embedded into managerial workflows to prioritise high-value cases, detect ambiguous or anomalous profiles, and construct sub-portfolios with higher precision or balanced accuracy. The uncertainty decomposition offered by HEUQ supports risk-aware decision-making by distinguishing customers whose predicted churn probability is reliable from those requiring human validation, and channel-specific results further highlight the importance of incorporating uncertainty information when customer behaviour varies across channels.

By equipping models with per-customer uncertainty estimates, management can explicitly trade off expected profit against risk. High-value policyholders with a high predicted churn probability and low epistemic uncertainty can be targeted automatically, while customers exhibiting high epistemic uncertainty (strong model disagreement) can be escalated for expert review. This approach helps optimise retention costs and manage financial exposure by reducing the risk and cost of incorrect decisions.

Uncertainty type	Datasize in %	Quantile	Unc Minimal Bound	BA
u_e	100%	0.00	0.0000	64.95
u_e	90%	0.10	0.0035	66.01
u_e	80%	0.20	0.0062	66.96
u_e	70%	0.30	0.0095	68.07
u_e	60%	0.40	0.0139	69.67
u_e	55%	0.45	0.0167	70.34
u_e	50%	0.50	0.0201	71.26
u_e	45%	0.55	0.0242	71.98
u_e	40%	0.60	0.0292	72.96
u_e	30%	0.70	0.0405	74.34
u_t	100%	0.00	0.3678	64.95
u_t	90%	0.10	0.7246	62.81
u_t	80%	0.20	0.8114	61.04
u_t	70%	0.30	0.8692	59.59
u_t	60%	0.40	0.9132	58.40
u_t	55%	0.45	0.9300	57.63
u_t	50%	0.50	0.9444	56.84
u_t	45%	0.55	0.9563	56.11
u_t	40%	0.60	0.9662	55.66
u_t	30%	0.70	0.9818	53.98
u_a	100%	0.00	0.3164	64.95
u_a	90%	0.10	0.6690	62.85
u_a	80%	0.20	0.7603	61.15
u_a	70%	0.30	0.8283	59.72
u_a	60%	0.40	0.8829	58.47
u_a	55%	0.45	0.9047	57.84
u_a	50%	0.50	0.9225	57.03
u_a	45%	0.55	0.9373	56.47
u_a	40%	0.60	0.9493	55.76
u_a	30%	0.70	0.9677	54.44

TABLE 5.2: Evolution of the BA score with increasing level of uncertainty (u_t , u_e , u_a)

This managerial layer complements the theoretical and empirical contributions of this thesis, demonstrating how uncertainty-aware models can be operationalised in real-world decision environments.

5.4 Synthesis of Key Empirical Insights

Across the thesis, three broad conclusions emerge.

First, uncertainty methods behave predictably in relation to data quality. BF and CP produce reliable uncertainty estimates when class separability is strong but revert to wide, non-informative uncertainty intervals under extreme overlap and imbalance. This highlights structural limitations inherent to the data rather than methodological flaws.

Second, HEUQ demonstrates that heterogeneous ensembles provide more trustworthy epistemic uncertainty than homogeneous ensembles. Their diversity mitigates representation distortions, supports robustness under feature transformations, and yields consistently superior calibration. The conversion prediction experiments further illustrate HEUQ's practical relevance by revealing systematic behavioural differences between digital and physical customer channels.

Third, in sequential decision-making settings, MASURE shows that modern uncertainty-aware RL methods can be made significantly more robust without incurring the computational costs typical of deep ensembles. By integrating Masksembles

into the Q-learning update rule, MASURE achieves a favourable balance between robustness, and efficiency.

5.5 Limitations

The contributions of this thesis must be interpreted in light of several limitations.

For the ancillary investigation of BF and CP, an important limitation is their sensitivity to the underlying data conditions. On noisy, weakly separable, and strongly imbalanced industrial data, BF tends to produce broad sets as outputs, while conformal prediction may achieve nominal coverage yet still return large or weakly informative prediction sets. Their practical usefulness improves when the underlying features provide stronger class separability.

For HEUQ, only two aggregation mechanisms, the Basic Combination Rule and a convex combination, were examined, leaving open the exploration of other optimization based strategies. The heterogeneous ensemble also requires occasional model-specific hyperparameter tuning, which is time-consuming and does not always yield optimal ensemble-level behaviour.

For MASURE, the method has not yet been evaluated on continuous-control environments, and its exploration capability remains limited, particularly in sparse-reward settings where both MASURE and State-of-art IVDQN underperform. Moreover, MASURE does not explicitly disentangle aleatoric from epistemic uncertainty within the update rule, limiting its ability to treat state noise and model uncertainty differently.

These constraints define the boundaries of the contributions and motivate several directions for continued research.

5.6 Future Work

Future work may extend this thesis in several promising directions. For heterogeneous ensembles, alternative UQ frameworks including evidential methods, distribution free approaches, and hybrid techniques, may offer richer decompositions of epistemic uncertainty, particularly in high-dimensional or imbalanced settings. Expanding HEUQ to broader operations-research contexts, such as sequential decision tasks or multi-objective optimization, may further enhance its managerial applicability. More flexible aggregation strategies, including genetic algorithms, particle swarm optimization, and differential evolution, represent a natural extension of the ensemble-design space.

For reinforcement learning, future work could disentangle aleatoric and epistemic uncertainty within MASURE, enabling agents to respond differently to state noise versus model uncertainty. Scaling MASURE to high-dimensional continuous control tasks would provide a more comprehensive assessment of its robustness. We

next highlight a decision-centric direction that is especially relevant for industrial deployment: “predict–then–optimize” with uncertainty.

5.6.1 From prediction to “predict–then–optimize” with uncertainty

A promising future direction is to move from uncertainty-aware prediction toward uncertainty-aware prescription, where models are evaluated and trained based on the decision quality they induce when embedded in an optimization problem (Bertsimas and Kallus, 2020; Elmachoub and Grigas, 2022; Mandi et al., 2024). This perspective is especially relevant in operational settings because decision constraints (budget, contact capacity, and asymmetric costs) often dominate the value of marginal accuracy improvements.

In our context, HEUQ can be extended to this paradigm by defining a cost-sensitive objective for portfolio selection, where false positives (contacting non-churners) and false negatives (missing true churners) carry different business costs. Uncertainty estimates can then be incorporated into the optimization layer, for example as a penalty that discourages actions on unreliable profiles, or as a robustness constraint that restricts actions to cases whose expected benefit remains positive under plausible model error. MASURE can be extended in the same spirit in sequential decision problems by defining rewards in business utility terms (retained value minus intervention costs) and by using epistemic uncertainty as a safety signal that encourages conservative actions or selective deferral when the agent is unsure.

Concretely, this line of work can be developed along three complementary axes:

- **Decision formulation (prescription layer).** Define an optimization problem that selects an intervention portfolio under operational constraints (e.g., contact capacity and campaign budget), using predicted churn risk and customer value as inputs.
- **Uncertainty-aware optimization.** Incorporate uncertainty into the optimization step, for example by penalising high-uncertainty profiles, imposing risk constraints, or enabling selective deferral based on uncertainty (Bertsimas and Kallus, 2020).
- **Decision-centric evaluation and deployment.** Evaluate the end-to-end pipeline using decision-focused metrics (e.g., expected profit uplift, cost per retained customer, regret relative to an oracle, and stability under time-based shifts) rather than only predictive metrics (Bertsimas and Kallus, 2020; Mandi et al., 2024).

In summary, predict–then–optimize provides a direct way to translate calibrated predictions and uncertainty into operational choices under constraints (Bertsimas and Kallus, 2020; Elmachoub and Grigas, 2022). We now make this link concrete through a simple uncertainty-aware sub-portfolio formulation.

Sub-portfolio selection as a prescriptive optimization problem: A natural next step is to make the sub-portfolio decision explicit. Let x denote an individual customer profile. Let $V(x)$ denote the managerial value of retaining customer x (e.g., expected margin or lifetime value), and let C denote the per-customer intervention cost (e.g., discount, call-centre time, or campaign cost). Let $\hat{p}(x)$ be the predicted probability of churn produced by the predictive model. To account for decision risk induced by predictive uncertainty, we define a utility score

$$S(x) = V(x) \hat{p}(x) - C - \lambda g(u_t(x), u_e(x)),$$

where $u_t(x)$ and $u_e(x)$ are the total and epistemic uncertainty estimates for x (Chapter 3), $g(\cdot)$ is a nonnegative penalty function that maps uncertainty to a risk surcharge, and $\lambda \geq 0$ controls how strongly uncertainty reduces the decision score. For example, one can set $g(u_t, u_e) = u_t$ to penalise overall unreliability, set $g(u_t, u_e) = u_e$ to penalise model disagreement only, or use a weighted combination $g(u_t, u_e) = \alpha u_t + (1 - \alpha) u_e$ with $\alpha \in [0, 1]$.

Given $S(x)$, one can select a sub-portfolio $\mathcal{S}$ that maximises $\sum_{x \in \mathcal{S}} S(x)$ under capacity and budget constraints (e.g., $|\mathcal{S}| \leq K$ and $\sum_{x \in \mathcal{S}} C \leq B$). This formulation makes explicit how uncertainty-aware predictions translate into an operational portfolio selection problem.

Overall, the thesis establishes a foundation for uncertainty-aware machine learning and reinforcement learning in real-world decision environments. Theoretical contributions such as the formalisation of uncertainty for heterogeneous ensembles, combined with practical frameworks such as MASURE, demonstrate how uncertainty quantification can bridge machine learning, reinforcement learning, and managerial decision science. Future research building on these ideas may further strengthen the reliability, transparency, and safety of AI-driven decision systems.

Appendix A

Basic vs convex combination rule in HEUQ

Equation (3.2) expressing BCR is a simple averaging formula. At first sight, it seems obvious that we should be able to improve BCR results by adapting to a more general convex approach.

Strangely, it is not so easy to improve **substantially** HEUQ, as the following will show.

A.1 Convex Combination Rule

In that spirit, we define a **Convex Composition Rule (CCR(α))** as a combination of the set of predictive probabilities $p(\cdot|h_j, \mathbf{x})$ in the following way :

$$p(y_k|CCR(\alpha), \mathbf{x}) = \sum_{j=1}^M \alpha(j, \mathbf{x}) p(y_k|h_j, \mathbf{x})$$

Where for any profile $\mathbf{x}$, $(\alpha(1, \mathbf{x}), \dots, \alpha(M, \mathbf{x}))$ is a set of positive values adding to 1.

Adaptation of Uncertainties

The definitions of $u_t(\mathbf{x})$, $u_a(\mathbf{x})$, $u_e(\mathbf{x})$ could be logically adapted as follows :

1. $u_t(\mathbf{x})$ is the classical Shannon entropy (in logarithm base 2) of predictive probability $p(\cdot|\alpha), \mathbf{x}$:

$$u_t(\mathbf{x}) = - \sum_{k=0}^1 p(y_k|CCR, \mathbf{x}) \log_2(p(y_k|CCR, \mathbf{x}))$$

2. $u_e(\mathbf{x})$ is the Jensen-Shannon divergence of predictive probability distributions $p(\cdot|h_1, \mathbf{x}), \dots, p(\cdot|h_M, \mathbf{x})$.

$$u_e(\mathbf{x}) = \sum_{j=1}^M \alpha(j, \mathbf{x}) KL[p(\cdot|h_j, \mathbf{x}), p(\cdot|BCR, \mathbf{x})]$$

3. $u_a(\mathbf{x}) = u_t(\mathbf{x}) - u_e(\mathbf{x})$ which gives us

$$u_a(\mathbf{x}) = - \sum_{k=0}^1 \sum_{j=1}^M \alpha(j, \mathbf{x}) p(y_k | h_j, \mathbf{x}) \log_2(p(y_k | h_j, \mathbf{x}))$$

We of course see $\alpha(j, \mathbf{x})$ as an *a priori* choice of hypothesis j , given the profile $\mathbf{x}$.

A.1.1 Comparison of BCR results with some specific CCR(α)

We of course only considered CCR rules being build on the information obtained after running the 6 learners on the training set.

One of the key ingredients guiding our choices is the **Expected Negative Log Likelihood** (or entropy) $ENLL(j, \mathbf{x})$ where $1 \leq j \leq 6$ denotes one of the 6 learners.

$$ENLL(j, \mathbf{x}) = - \sum_{k=0}^1 p(y_k | h_j, \mathbf{x}) \log(p(y_k | h_j, \mathbf{x}))$$

We consider the following CCR's :

1. HEUQ

$$\alpha(j, \mathbf{x}) = \frac{1}{6} \quad \forall \mathbf{x}$$

2. CCR1

$$\alpha(j, \mathbf{x}) = 1_{ENLL(j, \mathbf{x}) = \min_b(ENLL(b, \mathbf{x}))} \quad \forall \mathbf{x}$$

3. CCR2

$$\alpha(j, \mathbf{x}) = \frac{1}{5} * 1_{j < 4} \quad \forall \mathbf{x}$$

(HEUQ without CATBoost)

4. CCR3

$$\alpha(j, \mathbf{x}) = \frac{ENLL(j, \mathbf{x})^K}{\sum_h ENLL(h, \mathbf{x})^K} \quad \forall \mathbf{x}$$

for $K = -1, 1, -0.5$

5. CCR4

$$\alpha(j, \mathbf{x}) = \frac{1}{2} 1_{j \in \{1, 2\}} \quad \forall \mathbf{x}$$

(HEUQ only on LR and RF)

6. CCR5

$$\alpha(j, \mathbf{x}) = \frac{1}{4} 1_{ENLL(j, \mathbf{x}) < \min_b(ENLL(b, \mathbf{x}))} 1_{ENLL(j, \mathbf{x}) < \max_b(ENLL(b, \mathbf{x}))} \quad \forall \mathbf{x}$$

And the results are in table [A.1](#)

Model	BA $\uparrow$	Brier $\downarrow$	NLL $\downarrow$
HEUQ	64.95	0.2189	0.6255
CCR1	64.98	0.2480	0.7600
CCR2	64.85	0.2200	0.6281
CCR3(K = -1)	65.00	0.2197	0.6268
CCR3(K = 1)	64.99	0.2185	0.6247
CR3(K = -0.5)	64.94	0.2193	0.6260
CCR4	64.47	0.2153	0.6189
CCR5	64.74	0.2186	0.6252

TABLE A.1: Performance comparison of different convex combination rule with HEUQ

Embedding size	Distortion	NLL $\downarrow$	BS $\downarrow$	Balanced Accuracy $\uparrow$
10	0.1	0.654	0.232	62.35
10	0.3	0.6572	0.233	61.99
10	0.5	0.652	0.231	62.35
10	0.9	0.651	0.232	62.11
20	0.1	0.648	0.229	62.51
20	0.3	0.647	0.227	62.55
20	0.5	0.649	0.229	62.64
20	0.9	0.648	0.228	62.79
30	0.1	0.645	0.228	62.96
30	0.3	0.642	0.228	62.8
30	0.5	0.62	0.219	63.05
30	0.9	0.612	0.215	63.58
40	0.1	0.62	0.22	63.8
40	0.3	0.616	0.212	64.2
40	0.5	0.6	0.204	64.46
40	0.9	0.611	0.202	64.11

TABLE A.2: Performance HEUQ with RG

A.2 HEUQ and Data transformations

We limit the study of PCA with embedding sizes of 20, 25, 35, and 40. We observed all methods started overfitting for embedding sizes under 35 with a considerable drop in Balanced Accuracy and an increase in uncertainty.

Table.A.2 presents a comprehensive overview of the performance of HEUQ with varying embedding sizes and distortion rates. It is noticeable that an increased distortion rate leads (lower value) to reduced performance, while higher embedding size generally enhances the performance. When combined with the empirical results in Table.A.2, Table.3.6, and Table.3.3, the results are indicative that uncertainty-aware heterogeneous ensembles exhibit greater resilience to data transformations that introduce distortion.

A.3 Churn literature study

Given the focus of chapter 3, we provide a condensed overview of churn prediction literature. We categorize reviewed methods based on ensemble usage and address the question of uncertainty quantification following Bogaert and Delaere (2023).

Study	Models	IsEnsemble	IsHetrogeneous	Uncertainty
Khan et al., (2019)	NN	×	×	×
Kassem et al., 2020	NN,LR,NaiveBayes	×	×	×
Ahmad, Jafar, and Aljoumaa, 2019	DecisionTree,RF,GB,XGB	✓	×	×
Ullah et al., 2019	RF	✓	×	×
Xu, Ma, and Kim, 2021	XGB,LR,DecisionTree,NaiveBayes	✓	✓	×
He, Xiong, and Tsai, 2020	RandomisedTree,BG	✓	×	×
Mauritsius et al., 2020	DecisionTree,NaiveBayes,NN	×	×	×
Jain, Yadav, and Manoov, 2020	LR,RF,SVM,XGB	✓	×	×
Matuszelański and Kopczewska, 2022	XGB,LR	✓	×	×
Buckinx and Poel, 2005	LR,NN,RF	✓	×	×
Lemmens and Croux, 2006	Bagging,Boosting	✓	×	×
Burez and Poel, 2007	LR	×	×	×
Coussement and Poel, 2008	SVM,LR,RF	✓	×	×
Burez and Poel, 2009	LR,RF	✓	×	×
Tsai and Lu, 2009	NN	×	×	×
Xie et al., 2009	Balanced RF,NN,Decision Tree,SVM	✓	✓	×
De Bock and Poel, 2011	Rotation Forest,RotBoost,Bagging,RF,C4.5,CART	✓	×	×
Ballings and Poel, 2012	LR,Classification Tree,Bagging	✓	×	×
De Bock and Poel, 2012	GAMensPlus, Bagging, RF, RSM, LR, GAM	✓	×	×
Kim et al., 2012	Uniform subsampled ensemble of SVM,Bagging,Boosting	✓	×	×
Bogaert and Delaere, 2023	LR, Decision Tree, Naive Bayes, SVM, NN, Bagging, Uniform subsample, Random subspace, RF, AdaBoost, Stochastic Gradient Boost, XGB, LightGBM, CatBoost, Logistic Model Tree, Rotation Forest, Stacking, Genetic Algorithm, Hill climb, Likelihood, Nonnegative least square, Differential evolution, Particle swarm, Simulated annealing, Memetic Algorithm, Self organizing migration algorithm.	✓	✓	×
Coussement and De Bock, 2013	CART,RF,GAMens	✓	×	×
Idris, Rizwan, and Khan, 2012	RF,KNN	✓	×	×
Vafeiadis et al., 2015	AdaBoost,NN,SVM,DecisionTree,NaiveBayes,LR	✓	×	×
Idris and Khan, 2017	RF,Rotationforest,SVM	✓	✓	×
Óskarsdóttir et al., 2017	Relational Classifier,weighted vote relational classifier, class distribution relational classifier, Network-only link-based classifier, Reading activation relational classifier, LR,RF,NN	✓	×	×
Zhu et al., 2017	RF,Bagging, UnderBagging, OverBagging,SMOTEBagging, RUSBoost,SMOTEBoost, AdaBoost,C4.5	✓	×	×
De Caigny, Coussement, and De Bock, 2018	Logit Leaf model,Decision Tree, RF, Logistic Model Tree	✓	×	×
Zhu et al., 2018	LR,C4.5,RF,SVM	✓	×	×
Ullah et al., 2019	RF,AttributeslectedClassifier, J48,Random Tree,AdaBoost,Bagging, NaiveBayes,NN,LR,IBK,LWL	✓	×	×
De Bock and De Caigny, 2021	Splieruleensemble,LR,DecisionTree, RF,Ruleensemble,GAM,MARS	✓	×	×
Janssens et al., 2022	B2Boost,XGB,VerbrakenBoost,ProfLogit,ProfTree,RF,LASSO	✓	×	×
Lessmann et al., 2021	LR,DecisionTree,Discriminantanalysis, RF,SVM,GB	✓	✓	×
Sagala and Permai, 2021	XGB,LightGB,CatBoost	✓	×	×
Wu et al., 2021	AdaBoost,RandomForest,NN,BayesianLR,K-Means	✓	×	×
Gattermann-Itschert and Thonemann, 2021	LR,SVM,RF	✓	×	×
Mirkovic et al., 2022	LR,RF,SVM	✓	×	×
Our Study	LR,RF,Bagging,CatBoost,XGB,NN	✓	✓	✓

TABLE A.3: Existing literature explores ensemble methods in the use case of churn prediction, yet none address uncertainty estimation.

Bibliography

- Abdar, Moloud, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. (2021). "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". In: *Information Fusion* 76, pp. 243–297. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2021.05.008](https://doi.org/10.1016/j.inffus.2021.05.008).
- Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. (2023). "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774*.
- Afroogh, Saleh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi (2024). "Trust in AI: progress, challenges, and future directions". In: *Humanities and Social Sciences Communications* 11.1, p. 1568.
- Ahmad, Abdelrahim Kasem, Assef Jafar, and Kadan Aljoumaa (2019). "Customer churn prediction in telecom using machine learning in big data platform". In: *Journal of Big Data* 6.1, pp. 1–24. DOI: [10.1186/s40537-019-0191-6](https://doi.org/10.1186/s40537-019-0191-6).
- Ahn, Jaehyun, Junsik Hwang, Doyoung Kim, Hyukgeun Choi, and Shinjin Kang (2020). "A survey on churn analysis in various business domains". In: *IEEE Access* 8, pp. 220816–220839. DOI: [10.1109/ACCESS.2020.3042657](https://doi.org/10.1109/ACCESS.2020.3042657).
- An, Gaon, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song (2021). "Uncertainty-based offline reinforcement learning with diversified q-ensemble". In: *Advances in neural information processing systems* 34, pp. 7436–7447.
- Aregui, Astride and Thierry Denœux (2008). "Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities". In: *International Journal of Approximate Reasoning* 49.3, pp. 575–594. ISSN: 0888-613X. DOI: <https://doi.org/10.1016/j.ijar.2008.06.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0888613X0800087X>.
- Bag, Surajit, Gautam Srivastava, Md Mamoon Al Bashir, Sushma Kumari, Mihalios Giannakis, and Abdul Hannan Chowdhury (2022). "Journey of customers in this digital era: Understanding the role of artificial intelligence technologies in user engagement and conversion". In: *Benchmarking: An International Journal* 29.7, pp. 2074–2098. DOI: [10.1108/BIJ-07-2021-0415](https://doi.org/10.1108/BIJ-07-2021-0415).
- Bagot, Louis, Kevin Mets, and Steven Latré (2020). "Learning Intrinsically Motivated Options to Stimulate Policy Exploration". In: *4th Lifelong Machine Learning Workshop at ICML 2020*. URL: <https://openreview.net/forum?id=Vcf1fDmBYJk>.

- Bai, Xiao, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim (2021). "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments". In: *Pattern Recognition* 120, p. 108102.
- Ballings, Michel and Dirk Van den Poel (2012). "Customer event history for churn prediction: How long is long enough?" In: *Expert Systems with Applications* 39.18, pp. 13517–13522.
- Barber, Rina Foygel, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani (2023). "Conformal prediction beyond exchangeability". In: *The Annals of Statistics* 51.2, pp. 816–845.
- Batista, Gustavo EAPA, Ronaldo C Prati, and Maria Carolina Monard (2004). "A study of the behavior of several methods for balancing machine learning training data". In: *ACM SIGKDD explorations newsletter* 6.1, pp. 20–29.
- Ben-Tal, Aharon, Arkadi Nemirovski, and Laurent El Ghaoui (2009). *Robust optimization*. Princeton university press.
- Bertsimas, Dimitris and Nathan Kallus (2020). "From predictive to prescriptive analytics". In: *Management Science* 66.3, pp. 1025–1044.
- Bharathi S, Vijayakumar, Dhanya Pramod, and Ramakrishnan Raman (2022). "An ensemble model for predicting retail banking churn in the youth segment of customers". In: *Data* 7.5, p. 61.
- Bilkhoo, Aman, Mehran Hosseini, Milad Kazemi, and Nicola Paoletti (2025). "CONFEX: Uncertainty-Aware Counterfactual Explanations with Conformal Guarantees". In: *arXiv preprint arXiv:2510.19754*.
- Black, Fischer and Robert Litterman (1992). "Global portfolio optimization". In: *Financial analysts journal* 48.5, pp. 28–43.
- Bogaert, Matthias and Lex Delaere (2023). "Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art". In: *Mathematics* 11.5, p. 1137.
- Bolancé, Catalina, Montserrat Guillen, and Alemar E Padilla-Barreto (2016). "Predicting probability of customer churn in insurance". In: *Modeling and Simulation in Engineering, Economics and Management: International Conference, MS 2016, Teruel, Spain, July 4-5, 2016, Proceedings*. Springer, pp. 82–91.
- Bradley, Andrew P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern Recognition* 30.7, pp. 1145–1159. ISSN: 0031-3203. DOI: [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2).
- Breiman, Leo (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.
- Buckinx, Wouter and Dirk Van den Poel (2005). "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting". In: *European journal of operational research* 164.1, pp. 252–268.
- Burez, Jonathan and Dirk Van den Poel (2007). "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services". In: *Expert Systems with Applications* 32.2, pp. 277–288.

- (2009). “Handling class imbalance in customer churn prediction”. In: *Expert Systems with Applications* 36.3, pp. 4626–4636.
- Caldeira, João and Brian Nord (2020). “Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms”. In: *Machine Learning: Science and Technology* 2.1, p. 015002.
- Candes, Emmanuel, Lihua Lei, and Zhimei Ren (2023). “Conformalized survival analysis”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85.1, pp. 24–45.
- Cauchois, Maxime, Suyash Gupta, and John C. Duchi (2021). “Knowing what You Know: valid and validated confidence sets in multiclass and multilabel prediction”. In: *Journal of Machine Learning Research* 22.81, pp. 1–42. URL: <http://jmlr.org/papers/v22/20-753.html>.
- Chander, Bhanu, Chinju John, Lekha Warriar, and Kumaravelan Gopalakrishnan (2025). “Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness”. In: *ACM Computing Surveys* 57.6, pp. 1–49.
- Charpentier, Bertrand, Ransalu Senanayake, Mykel Kochenderfer, and Stephan Günnemann (2022). “Disentangling epistemic and aleatoric uncertainty in reinforcement learning”. In: *arXiv preprint arXiv:2206.01558*.
- Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002). “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16, pp. 321–357.
- Chen, Qi, Amanda Whitbrook, Uwe Aickelin, and Chris Roadknight (2014). “Data classification using the Dempster–Shafer method”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 26.4, pp. 493–517. DOI: [10.1080/0952813X.2014.886301](https://doi.org/10.1080/0952813X.2014.886301). URL: <https://doi.org/10.1080/0952813X.2014.886301>.
- Chen, Tianqi and Carlos Guestrin (2016). “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.
- Chen, Xinning, Xuan Liu, Canhui Luo, and Jiangjin Yin (2022). “Robust multi-agent reinforcement learning for noisy environments”. In: *Peer-to-Peer Networking and Applications* 15.2, pp. 1045–1056.
- Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2021). “Distributional conformal prediction”. In: *Proceedings of the National Academy of Sciences* 118.48, e2107794118.
- Christodoulou, Leonidas and Chang Sun (2026). “The Impact of Machine Learning Uncertainty on the Robustness of Counterfactual Explanations”. In: *Expert Systems with Applications*, p. 131198.
- Coussement, Kristof and Koen W De Bock (2013). “Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning”. In: *Journal of Business Research* 66.9, pp. 1629–1636.
- Coussement, Kristof and Dirk Van den Poel (2008). “Churn prediction in subscription services: An application of support vector machines while comparing two

- parameter-selection techniques". In: *Expert systems with applications* 34.1, pp. 313–327.
- Dal Pozzolo, Andrea, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi (2015). "Calibrating probability with undersampling for unbalanced classification". In: *2015 IEEE symposium series on computational intelligence*. IEEE, pp. 159–166.
- Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen (2020). "A survey of the state of explainable AI for natural language processing". In: *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pp. 447–459.
- De, Soumi and P Prabu (2022). "Predicting customer churn: A systematic literature review". In: *Journal of Discrete Mathematical Sciences and Cryptography* 25.7, pp. 1965–1985.
- De Bock, Koen W and Arno De Caigny (2021). "Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling". In: *Decision Support Systems* 150, p. 113523.
- De Bock, Koen W and Dirk Van den Poel (2011). "An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction". In: *Expert Systems with Applications* 38.10, pp. 12293–12301.
- (2012). "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models". In: *Expert Systems with Applications* 39.8, pp. 6816–6826.
- De Caigny, Arno, Kristof Coussement, and Koen W De Bock (2018). "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees". In: *European Journal of Operational Research* 269.2, pp. 760–772.
- Delage, Erick and Yinyu Ye (2010). "Distributionally robust optimization under moment uncertainty with application to data-driven problems". In: *Operations research* 58.3, pp. 595–612.
- Dempster, Arthur P. (1967). "Upper and Lower Probabilities Induced by a Multivalued Mapping. Annals of Mathematical Statistics. Vol. 38. P. 325-339". In.
- (2008). "Upper and lower probabilities induced by a multivalued mapping". In: *Classic works of the Dempster-Shafer theory of belief functions*. Springer, pp. 57–72.
- Denoeux, T. (2000). "A neural network classifier based on Dempster-Shafer theory". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30.2, pp. 131–150. DOI: [10.1109/3468.833094](https://doi.org/10.1109/3468.833094).
- Denoeux, Thierry (1995). "A k-nearest neighbor classification rule based on Dempster-Shafer theory". In: vol. 25. 5. IEEE, pp. 804–813.
- Depeweg, Stefan, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft (2018). "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning". In: *International Conference on Machine Learning*. PMLR, pp. 1184–1193.

- Ding, Tiffany, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani (2023). "Class-Conditional Conformal Prediction with Many Classes". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., pp. 64555–64576. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/cb931eddd563f8d473c355518ce8601c-Paper-Conference.pdf.
- Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*.
- Dubois, Didier and Henri Prade (1988). "Representation and combination of uncertainty with belief functions and possibility measures". In: *Computational intelligence* 4.3, pp. 244–264.
- Durasov, Nikita, Timur Bagautdinov, Pierre Baque, and Pascal Fua (June 2021). "Masksembles for Uncertainty Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13539–13548.
- Ekbal, Asif and Sriparna Saha (2011). "A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies". In: *Expert Systems with Applications* 38.12, pp. 14760–14772.
- Elmachtoub, Adam N and Paul Grigas (2022). "Smart "predict, then optimize"". In: *Management Science* 68.1, pp. 9–26.
- Fan, Hongjian, Ming Fan, Kotagiri Ramamohanarao, and Mengxu Liu (2006). "Further improving emerging pattern based classifiers via bagging". In: *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, 2006. Proceedings* 10. Springer, pp. 91–96.
- Fannjiang, Clara, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan (2022). "Conformal prediction under feedback covariate shift for biomolecular design". In: *Proceedings of the National Academy of Sciences* 119.43, e2204569119.
- Feng, Rui, Ankun Ma, Zhijia Jing, Xiaoning Gu, Pengfei Dang, and Baozhen Yao (2022). "Understanding the uncertainty of traffic time prediction impacts on parking lot reservation in logistics centers". In: *Annals of Operations Research*, pp. 1–23.
- Flet-Berliac, Yannis and Philippe Preux (July 2020). "Only Relevant Information Matters: Filtering Out Noisy Samples To Boost RL". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. Main track. International Joint Conferences on Artificial Intelligence Organization, pp. 2711–2717. DOI: [10.24963/ijcai.2020/376](https://doi.org/10.24963/ijcai.2020/376). URL: <https://doi.org/10.24963/ijcai.2020/376>.
- Fontana, Matteo, Gianluca Zeni, and Simone Vantini (2023). "Conformal prediction: a unified review of theory and new challenges". In: *Bernoulli* 29.1, pp. 1–23.

- Fortunato, Meire, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg (2018). "Noisy Networks For Exploration". In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=rywHCPkAW>.
- Fox, Roy, Ari Pakman, and Naftali Tishby (2016). "Taming the noise in reinforcement learning via soft updates". In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'16. Jersey City, New Jersey, USA: AUAI Press, 202–211. ISBN: 9780996643115.
- Fuglede, B. and F. Topsøe (2004). "Jensen-Shannon divergence and Hilbert space embedding". In: *International Symposium on Information Theory, 2004. ISIT 2004*. pp. 31–. DOI: [10.1109/ISIT.2004.1365067](https://doi.org/10.1109/ISIT.2004.1365067).
- Gadgil, Soham, Yunfeng Xin, and Chengzhe Xu (2020). "Solving the lunar lander problem under uncertainty using reinforcement learning". In: *2020 SoutheastCon*. Vol. 2. IEEE, pp. 1–8.
- Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR, pp. 1050–1059.
- Gal, Yarin, Rowan McAllister, and Carl Edward Rasmussen (2016). "Improving PILCO with Bayesian neural network dynamics models". In: *Data-efficient machine learning workshop, ICML*. Vol. 4. 34, p. 25.
- Garcin, Matthieu and Samuel Stéphan (2021). "Credit scoring using neural networks and SURE posterior probability calibration". In: *arXiv preprint arXiv:2107.07206*.
- Gattermann-Itschert, Theresa and Ulrich W. Thonemann (2021). "How training on multiple time slices improves performance in churn prediction". In: *European Journal of Operational Research* 295.2, pp. 664–674. ISSN: 0377-2217. DOI: <https://doi.org/10.1016/j.ejor.2021.05.035>. URL: <https://www.sciencedirect.com/science/article/pii/S037722172100463X>.
- Geiler, Louis, Séverine Affeldt, and Mohamed Nadif (2022). "A survey on machine learning methods for churn prediction". In: *International Journal of Data Science and Analytics* 14.3, pp. 217–242.
- Gu, Zhaoyuan, Zhenzhong Jia, and Howie Choset (2019). "Adversary a3c for robust reinforcement learning". In: *arXiv preprint arXiv:1912.00330*.
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger (2017). "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR, pp. 1321–1330.
- Gustafsson, Fredrik K, Martin Danelljan, and Thomas B Schon (2020). "Evaluating scalable bayesian deep learning methods for robust computer vision". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 318–319.
- Hansen, Lars Peter and Thomas J Sargent (2001). "Robust control and model uncertainty". In: *American Economic Review* 91.2, pp. 60–66.

- Hashemi, Mahdi and Hassan Karimi (2018). "Weighted machine learning". In: *Statistics, Optimization and Information Computing* 6.4, pp. 497–525.
- He, Yunxuan, Ying Xiong, and Yiting Tsai (2020). "Machine learning based approaches to predict customer churn for an insurance company". In: *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, pp. 1–6.
- Hoffmann, Lara, Ines Fortmeier, and Clemens Elster (2021). "Uncertainty quantification by ensemble learning for computational optical form measurements". In: *Machine Learning: Science and Technology* 2.3, p. 035030.
- Hüllermeier, Eyke, Sébastien Destercke, and Mohammad Hossein Shaker (2022). "Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 548–557.
- Hüllermeier, Eyke and Willem Waegeman (2021). "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110, pp. 457–506.
- Idris, Adnan and Asifullah Khan (2017). "Churn prediction system for telecom using filter-wrapper and ensemble classification". In: *The Computer Journal* 60.3, pp. 410–430.
- Idris, Adnan, Muhammad Rizwan, and Asifullah Khan (2012). "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies". In: *Computers & Electrical Engineering* 38.6, pp. 1808–1819.
- Inagaki, Toshiyuki (2002). "Interdependence between safety-control policy and multiple-sensor schemes via Dempster-Shafer theory". In: *IEEE Transactions on Reliability* 40.2, pp. 182–188.
- Jain, Himani, Garima Yadav, and R Manoov (2020). "Churn prediction and retention in banking, telecom and IT sectors using machine learning techniques". In: *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019*. Springer, pp. 137–156.
- Janssens, Bram, Matthias Bogaert, Astrid Bagué, and Dirk Van den Poel (2022). "B2Boost: Instance-dependent profit-driven modelling of B2B churn". In: *Annals of Operations Research*, pp. 1–27.
- Jolliffe, Ian T and Jorge Cadima (2016). "Principal component analysis: a review and recent developments". In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065, p. 20150202.
- Jurado, Kyle, Sydney C Ludvigson, and Serena Ng (2015). "Measuring uncertainty". In: *American Economic Review* 105.3, pp. 1177–1216.
- Kassem, Essam Abou el, Shereen Ali Hussein, Alaa Mostafa Abdelrahman, and Fahad Kamal Alsheref (2020). "Customer churn prediction model and identifying features to increase customer retention based on user generated content". In: *International Journal of Advanced Computer Science and Applications* 11.5. DOI: [10.14569/IJACSA.2020.0110567](https://doi.org/10.14569/IJACSA.2020.0110567).

- Khan, Yasser, Shahryar Shafiq, Abid Naeem, Sheeraz Ahmed, Nadeem Safwan, and Sabir Hussain (2019). "Customers churn prediction using artificial neural networks (ANN) in telecom industry". In: *International journal of advanced computer science and applications* 10.9.
- Khraishi, Raad and Ramin Okhrati (2023). "Simple noisy environment augmentation for reinforcement learning". In: *arXiv preprint arXiv:2305.02882*.
- Kim, Namhyoung, Kyu-Hwan Jung, Yong Seog Kim, and Jaewook Lee (2012). "Uniformly subsampled ensemble (USE) for churn management: Theory and implementation". In: *Expert Systems with Applications* 39.15, pp. 11839–11845.
- Kiyani, Shayan, George J. Pappas, and Hamed Hassani (July 2024). "Conformal Prediction with Learned Features". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 24749–24769. URL: <https://proceedings.mlr.press/v235/kiyani24a.html>.
- Kläs, Michael and Anna Maria Vollmer (2018). "Uncertainty in machine learning applications: A practice-driven classification of uncertainty". In: *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings* 37. Springer, pp. 431–438.
- Kline, S. J. (June 1985). "The Purposes of Uncertainty Analysis". In: *Journal of Fluids Engineering* 107.2, pp. 153–160. ISSN: 0098-2202. DOI: [10.1115/1.3242449](https://doi.org/10.1115/1.3242449). eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/107/2/153/5596058/153_1.pdf. URL: <https://doi.org/10.1115/1.3242449>.
- Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30.
- Lalwani, Praveen, Manas Kumar Mishra, Jasroop Singh Chadha, and Pratyush Sethi (2022). "Customer churn prediction system: a machine learning approach". In: *Computing*, pp. 1–24.
- Lee, Kimin, Michael Laskin, Aravind Srinivas, and Pieter Abbeel (2021). "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning". In: *International conference on machine learning*. PMLR, pp. 6131–6141.
- Lefevre, Eric, Patrick Vannooenberghe, and Olivier Colot (1999). "Basic belief assignment in the framework of information theory". In: *7th European Congress on Intelligent Techniques and Soft Computing (EUFIT'99), Aachen*.
- Lehman, Joel, Elliot Meyerson, Tarek El-Gaaly, Kenneth O Stanley, and Tarin Ziyadee (2025). "Evolution and the knightian blindspot of machine learning". In: *arXiv preprint arXiv:2501.13075*.

- Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman (2018). "Distribution-free predictive inference for regression". In: *Journal of the American Statistical Association* 113.523, pp. 1094–1111.
- Lei, Lihua and Emmanuel J Candès (2021). "Conformal inference of counterfactuals and individual treatment effects". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.5, pp. 911–938.
- Lemmens, Aurélie and Christophe Croux (2006). "Bagging and boosting classification trees to predict churn". In: *Journal of Marketing Research* 43.2, pp. 276–286.
- Lessmann, Stefan, Johannes Haupt, Kristof Coussement, and Koen W De Bock (2021). "Targeting customers for profit: An ensemble learning framework to support marketing decision-making". In: *Information Sciences* 557, pp. 286–301.
- Li, Andrew, Zizhao Chen, Toryn Klassen, Pashootan Vaezipoor, Rodrigo Toro Icarte, and Sheila McIlraith (2024). "Reward machines for deep RL in noisy and uncertain environments". In: *Advances in Neural Information Processing Systems* 37, pp. 110341–110368.
- Liao, Yuan, Xinjie Ma, Andreas Neuhierl, and Linda Schilling (2025). "The uncertainty of machine learning predictions in asset pricing". In: *arXiv preprint arXiv:2503.00549*.
- Lin, Enlu, Qiong Chen, and Xiaoming Qi (2020). "Deep reinforcement learning for imbalanced classification". In: *Applied Intelligence* 50.8, pp. 2488–2502.
- Lindenstrauss, W Johnson J (1984). "Extensions of Lipschitz maps into a Hilbert space". In: *Contemp. Math* 26.189–206, p. 2.
- Liu, Qi, Yanjie Li, Yuecheng Liu, Meiling Chen, Shaohua Lv, and Yunhong Xu (2021). "Exploration via distributional reinforcement learning with epistemic and aleatoric uncertainty estimation". In: *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. IEEE, pp. 2256–2261.
- Liu, Zhun ga, Quan Pan, and Jean Dezert (2013). "A new belief-based K-nearest neighbor classification method". In: *Pattern Recognition* 46.3, pp. 834–844. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2012.10.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320312004293>.
- Lockwood, Owen and Mei Si (2022). "A review of uncertainty for deep reinforcement learning". In: *Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. AIIDE'22. Pomona, CA, USA: AAAI Press. ISBN: 978-1-57735-877-0. DOI: [10.1609/aiide.v18i1.21959](https://doi.org/10.1609/aiide.v18i1.21959).
- Luis, Carlos E, Alessandro G Bottero, Julia Vinogradskaya, Felix Berkenkamp, and Jan Peters (2023). "Model-based uncertainty in value functions". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 8029–8052.
- Maddox, Wesley J, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson (2019). "A simple baseline for bayesian uncertainty in deep learning". In: *Advances in neural information processing systems* 32.
- Mai, Vincent, Kaustubh Mani, and Liam Paull (2022). "Sample Efficient Deep Reinforcement Learning via Uncertainty Estimation". In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=vrW3tvDf0JQ>.

- Malinin, Andrey and Mark Gales (2018). "Predictive uncertainty estimation via prior networks". In: *Advances in neural information processing systems* 31.
- Mandi, Jayanta, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto (2024). "Decision-focused learning: Foundations, state of the art, benchmark and future opportunities". In: *Journal of Artificial Intelligence Research* 80, pp. 1623–1701.
- Martínez, Andrés, Claudia Schmuck, Sergiy Pereverzyev, Clemens Pirker, and Markus Haltmeier (2020). "A machine learning framework for customer purchase prediction in the non-contractual setting". In: *Eur. J. Oper. Res.* 281, pp. 588–596. DOI: [10.1016/J.EJOR.2018.04.034](https://doi.org/10.1016/J.EJOR.2018.04.034).
- Matuszelański, Kamil and Katarzyna Kopczewska (2022). "Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach". In: *Journal of Theoretical and Applied Electronic Commerce Research* 17.1, pp. 165–198.
- Mauritsius, Tuga, Reagen Y Sayoga, Natanael Alamas, Marcelina Anggraeni, Faisal Binsar, et al. (2020). "Customer churn prediction models for PT. XYZ insurance". In: *2020 8th international conference on orange technology (ICOT)*. IEEE, pp. 1–6.
- Mavor-Parker, Augustine, Kimberly Young, Caswell Barry, and Lewis Griffin (2022). "How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation". In: *International conference on machine learning*. PMLR, pp. 15220–15240.
- Mehdiyev, Nijat, Maxim Majlatow, and Peter Fettke (2023). "Explainable Artificial Intelligence Meets Uncertainty Quantification for Predictive Process Monitoring." In: *PMAI@IJCAI*, pp. 29–32.
- Michaud, Richard O (1989). "The Markowitz optimization enigma: Is 'optimized' optimal?" In: *Financial analysts journal* 45.1, pp. 31–42.
- Mirkovic, Milan, Teodora Lolic, Darko Stefanovic, Andras Anderla, and Danijela Gracanin (2022). "Customer Churn Prediction in B2B Non-Contractual Business Settings Using Invoice Data". In: *Applied Sciences* 12.10. ISSN: 2076-3417. DOI: [10.3390/app12105001](https://doi.org/10.3390/app12105001).
- Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. (2015). "Human-level control through deep reinforcement learning". In: *nature* 518.7540, pp. 529–533.
- Mohammed, Ammar and Rania Kora (2023). "A comprehensive review on ensemble deep learning: Opportunities and challenges". In: *Journal of King Saud University-Computer and Information Sciences*.
- Mukhopadhyay, Debajyoti, Aarati Malusare, Anagha Nandanwar, and Shriya Sakshi (2021). "An approach to mitigate the risk of customer churn using machine learning algorithms". In: *Machine Learning for Predictive Analysis: Proceedings of ICTIS 2020*. Springer, pp. 133–142.
- Nicora, Giovanna, Miguel Rios, Ameen Abu-Hanna, and Riccardo Bellazzi (2022). "Evaluating pointwise reliability of machine learning prediction". In: *Journal of Biomedical Informatics* 127, p. 103996.

- Ning, Chao and Fengqi You (2019). "Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming". In: *Computers & Chemical Engineering* 125, pp. 434–448.
- Osband, Ian, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy (2016). "Deep exploration via bootstrapped DQN". In: *Advances in neural information processing systems* 29.
- Óskarsdóttir, María, Cristián Bravo, Wouter Verbeke, Carlos Sarraute, Bart Baesens, and Jan Vanthienen (2017). "Social network analytics for churn prediction in telco: Model building, evaluation and network architecture". In: *Expert Systems with Applications* 85, pp. 204–220.
- Ovadia, Yaniv, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek (2019). "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift". In: *Advances in neural information processing systems* 32.
- Papadopoulos, Harris, Alex Gammerman, and Volodya Vovk (2009). "Reliable diagnosis of acute abdominal pain with conformal prediction". In: *Engineering Intelligent Systems* 17.2, p. 127.
- Papadopoulos, Harris, Vladimir Vovk, and Alex Gammerman (2011). "Regression conformal prediction with nearest neighbours". In: *Journal of Artificial Intelligence Research* 40, pp. 815–840.
- Papadopoulos, Harris, Volodya Vovk, and Alex Gammerman (2007). "Conformal prediction with neural networks". In: *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*. Vol. 2. IEEE, pp. 388–395.
- Pereira, Telma, Sandra Cardoso, Manuela Guerreiro, Sara C Madeira, Alzheimer's Disease Neuroimaging Initiative, et al. (2020). "Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, Venn-ABERS, and Conformal Predictors: A case study in AD". In: *Journal of biomedical informatics* 101, p. 103350.
- Petrakova, Aleksandra, Michael Affenzeller, and Galina Merkurjeva (2015). "Heterogeneous versus homogeneous machine learning ensembles". In: *Information Technology and Management Science* 18.1, pp. 135–140.
- Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin (2018). "CatBoost: unbiased boosting with categorical features". In: *Advances in neural information processing systems* 31.
- Prusa, Joseph, Taghi M Khoshgoftaar, and Daivd J Dittman (2015). "Using ensemble learners to improve classifier performance on tweet sentiment data". In: *2015 IEEE international conference on information reuse and integration*. IEEE, pp. 252–257.
- Reineking, Thomas (2014). "Belief functions: theory and algorithms". PhD thesis. Universität Bremen.
- Rockafellar, R. Tyrrell (1970). *Convex Analysis*. Vol. 28. Princeton, N.J: Princeton University Press.

- Sachdeva, Rewat, Raghav Gakhar, Sharad Awasthi, Kavinder Singh, Ashutosh Pandey, and Anil Singh Parihar (2024). "Uncertainty and Noise Aware Decision Making for Autonomous Vehicles-A Bayesian Approach". In: *IEEE Transactions on Vehicular Technology*.
- Sagala, Noviyanti Tri Maretta and Syarifah Diana Permai (2021). "Enhanced Churn Prediction Model with Boosted Trees Algorithms in The Banking Sector". In: *2021 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, pp. 240–245.
- Salvi, Massimo, Silvia Seoni, Andrea Campagner, Arkadiusz Gertych, U Rajendra Acharya, Filippo Molinari, and Federico Cabitza (2025). "Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare". In: *International Journal of Medical Informatics* 197, p. 105846.
- Schwalbe, Gesina and Bettina Finzel (2024). "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts". In: *Data Mining and Knowledge Discovery* 38.5, pp. 3043–3101.
- Scriney, Michael, Dongyun Nie, and Mark Roantree (2020). "Predicting customer churn for insurance data". In: *Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings* 22. Springer, pp. 256–265.
- Semih, Bulut, Avci Emin, and Bulut Ahmet (2023). "Forecasting Conversion Rate for Real Time CPC Bidding With Target ROAS". In: *IEEE Access* 11, pp. 134908–134916. DOI: [10.1109/ACCESS.2023.3338022](https://doi.org/10.1109/ACCESS.2023.3338022).
- Sesmero, M Paz, Agapito I Ledezma, and Araceli Sanchis (2015). "Generating ensembles of heterogeneous classifiers using stacked generalization". In: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 5.1, pp. 21–34.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press. ISBN: 9780691081755. URL: <http://www.jstor.org/stable/j.ctv10vm1qb>.
- Shahzad, Raja Khurram and Niklas Lavesson (2013). "Comparative analysis of voting schemes for ensemble-based malware detection". In: *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications* 4.1, pp. 98–117.
- Shaker, Mohammad Hossein and Eyke Hüllermeier (2020). "Aleatoric and epistemic uncertainty with random forests". In: *Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings* 18. Springer, pp. 444–456.
- Shakya, Ashish Kumar, Gopinatha Pillai, and Sohom Chakrabarty (2023). "Reinforcement learning algorithms: A brief survey". In: *Expert Systems with Applications* 231, p. 120495.
- Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587, pp. 484–489.

- Singh, Akash, Ashwin Ittoo, and Pierre Ars (2024). "Beyond Yes or No: Making Reliable Decisions and improving personalized customer service". In: URL: <https://hdl.handle.net/2268/328138>.
- Singh, Akash, Ashwin Ittoo, Pierre Ars, and Elise Vandomme (2025a). "Beyond Yes or No: Making Reliable Decisions and improving personalized customer service (Poster)". In: *Research day*.
- (2025b). "Conformal Prediction: Calibrated Decision-Making". In: *Joint ORBEL-NGB conference on Operations Research*.
- (2025c). "Multi-channel multi-model customer conversion prediction in the insurance domain". In: *9th International Conference on Applied Economics and Business*. Springer Nature Computer Science journal., Switzerland.
- Singh, Akash, Ashwin Ittoo, Elise Vandomme, and Pierre Ars (2025d). "Uncertainty-Aware Reinforcement Learning Agents for Noisy Environments". In: *IEEE (ITT 2025)*.
- Smets, Philippe (2002). "The combination of evidence in the transferable belief model". In: *IEEE Transactions on pattern analysis and machine intelligence* 12.5, pp. 447–458.
- Smets, Philippe and Robert Kennes (1994). "The transferable belief model". In: *Artificial intelligence* 66.2, pp. 191–234.
- Soni, Vishvesh (2022). "Role of temporal, demographic, and behavioral factors in customer conversion through dynamic creative optimization in the consumer-packaged goods setting". In: *International Journal of Business Intelligence and Big Data Analytics* 5.1, pp. 46–56.
- Spiteri, Maria and George Azzopardi (2018). "Customer churn prediction for a motor insurance company". In: *2018 Thirteenth international conference on digital information management (ICDIM)*. IEEE, pp. 173–178.
- Straitouri, Eleni, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez (2023). "Improving expert predictions with conformal prediction". In: *International Conference on Machine Learning*. PMLR, pp. 32633–32653.
- Sun, Ke, Yi Liu, Yingnan Zhao, Hengshuai Yao, SHANGLING JUI, and Linglong Kong (2022). *Exploring the Robustness of Distributional Reinforcement Learning against Noisy State Observations*. URL: <https://openreview.net/forum?id=z2zmSDKONK>.
- Thölke, Philipp, Yorguin-Jose Mantilla-Ramos, Hamza Abdelhedi, Charlotte Maschke, Arthur Dehgan, Yann Harel, Anirudha Kemtur, Loubna Mekki Berrada, Myriam Sahraoui, Tammy Young, et al. (2023). "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data". In: *NeuroImage* 277, p. 120253.
- Thuy, Arthur and Dries F Benoit (2023). "Explainability through uncertainty: Trustworthy decision-making with neural networks". In: *European Journal of Operational Research*.
- (2024). "Explainability through uncertainty: Trustworthy decision-making with neural networks". In: *European Journal of Operational Research* 317.2, pp. 330–340.

- Tibshirani, Ryan J, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas (2019). "Conformal Prediction Under Covariate Shift". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.
- Tjoa, Erico and Cuntai Guan (2020). "A survey on explainable artificial intelligence (xai): Toward medical xai". In: *IEEE transactions on neural networks and learning systems* 32.11, pp. 4793–4813.
- Tokuç, A. Aylin and Tamer Dağ (2024). "Customer Purchase Intent Prediction using Feature Aggregation on E-Commerce Clickstream Data". In: *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5. DOI: [10.1109/IDAP64064.2024.10711144](https://doi.org/10.1109/IDAP64064.2024.10711144).
- Tsai, Chih-Fong and Yu-Hsin Lu (2009). "Customer churn prediction by hybrid neural networks". In: *Expert Systems with Applications* 36.10, pp. 12547–12553.
- Ullah, Irfan, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul Islam, and Sung Won Kim (2019). "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector". In: *IEEE access* 7, pp. 60134–60149.
- Vafeiadis, Thanasis, Konstantinos I Diamantaras, George Sarigiannidis, and K Ch Chatzisavvas (2015). "A comparison of machine learning techniques for customer churn prediction". In: *Simulation Modelling Practice and Theory* 55, pp. 1–9.
- Vazquez, Janette and Julio C Facelli (2022). "Conformal prediction in clinical medical sciences". In: *Journal of Healthcare Informatics Research* 6.3, pp. 241–252. DOI: <https://doi.org/10.1007/s41666-021-00113-8>.
- Verbraken, Thomas, Wouter Verbeke, and Bart Baesens (2012). "A novel profit maximizing metric for measuring classification performance of customer churn prediction models". In: *IEEE transactions on knowledge and data engineering* 25.5, pp. 961–973.
- Vovk, Vladimir, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman (June 2018). "Cross-conformal predictive distributions". In: *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*. Ed. by Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Ralf Peeters. Vol. 91. Proceedings of Machine Learning Research. PMLR, pp. 37–51. URL: <https://proceedings.mlr.press/v91/vovk18a.html>.
- Wallace, Byron C and Issa J Dahabreh (2012). "Class probability estimates are unreliable for imbalanced data (and how to fix them)". In: *2012 IEEE 12th international conference on data mining*. IEEE, pp. 695–704.
- Walley, Peter (1991). *Statistical reasoning with imprecise probabilities*. Vol. 42. London: Chapman and Hall: Springer.
- Wang, Cheng, Carolin Lawrence, and Mathias Niepert (2020). "Uncertainty estimation and calibration with finite-state probabilistic RNNs". In: *arXiv preprint arXiv:2011.12010*.

- Wang, Guotai, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren (2019). "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks". In: *Neurocomputing* 338, pp. 34–45.
- Wang, Hui, Jun Liu, and Juan Carlos Augusto (2010). "Mass function derivation and combination in multivariate data spaces". In: *Information Sciences* 180.6, pp. 813–819.
- Watson-Daniels, Jamelle, Flavio du Pin Calmon, Alexander D'Amour, Carol Long, David C Parkes, and Berk Ustun (2024). "Predictive Churn with the Set of Good Models". In: *arXiv preprint arXiv:2402.07745*.
- Wei, Yang and Fu Yaowen (2013). "Constructing basic belief assignment from feature data". In: *2013 Chinese Automation Congress*. IEEE, pp. 605–610.
- Wendi, Ji, Wang Xiaoling, and Zhu Feida (2017). "Time-aware conversion prediction". In: *Frontiers of Computer Science* 11, pp. 702–716. DOI: [10.1007/s11704-016-5546-y](https://doi.org/10.1007/s11704-016-5546-y).
- Woźniak, Michał, Manuel Grana, and Emilio Corchado (2014). "A survey of multiple classifier systems as hybrid systems". In: *Information Fusion* 16, pp. 3–17.
- Wu, Shuli, Wei-Chuen Yau, Thian-Song Ong, and Siew-Chin Chong (2021). "Integrated churn prediction and customer segmentation framework for telco business". In: *IEEE Access* 9, pp. 62118–62136.
- Xie, Yaya, Xiu Li, EWT Ngai, and Weiyun Ying (2009). "Customer churn prediction using improved balanced random forests". In: *Expert Systems with Applications* 36.3, pp. 5445–5449.
- Xu, Chen and Yao Xie (2023). "Conformal Prediction for Time Series". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10, pp. 11575–11587. DOI: [10.1109/TPAMI.2023.3272339](https://doi.org/10.1109/TPAMI.2023.3272339).
- Xu, Tianpei, Ying Ma, and Kangchul Kim (2021). "Telecom churn prediction system based on ensemble learning using feature grouping". In: *Applied Sciences* 11.11, p. 4742.
- Yager, Ronald R (1987). "On the Dempster-Shafer framework and new combination rules". In: *Information sciences* 41.2, pp. 93–137.
- Yager, Ronald R. and Liping Liu (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions*. Vol. 219. Studies in Fuzziness and Soft Computing. Springer.
- Yeo, Jinyoung, Seung-won Hwang, sungchul kim, Eunye Koh, and Nedim Lipka (2020). "Conversion Prediction from Clickstream: Modeling Market Prediction and Customer Predictability". In: *IEEE Transactions on Knowledge and Data Engineering* 32.2, pp. 246–259. DOI: [10.1109/TKDE.2018.2884467](https://doi.org/10.1109/TKDE.2018.2884467).
- Yong, Deng, Shi WenKang, Zhu ZhenFu, and Liu Qi (2004). "Combining belief functions based on distance of evidence". In: *Decision support systems* 38.3, pp. 489–493.
- Zadeh, Lotfi A (1984). "Review of a mathematical theory of evidence". In: *AI magazine* 5.3, pp. 81–81.

- Zhang, Huan, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh (2020). "Robust deep reinforcement learning against adversarial perturbations on state observations". In: *Advances in neural information processing systems* 33, pp. 21024–21037.
- Zhang, Rong, Weiping Li, Wei Tan, and Tong Mo (2017). "Deep and shallow model for insurance churn prediction service". In: *2017 IEEE International Conference on services computing (SCC)*. IEEE, pp. 346–353.
- Zhou, Qianli, Guojing Tian, and Yong Deng (2023). "BF-QC: Belief functions on quantum circuits". In: *Expert Systems with Applications* 223, p. 119885. ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2023.119885>. URL: <https://www.sciencedirect.com/science/article/pii/S095741742300386X>.
- Zhou, Xiaofan, Baiting Chen, Yu Gui, and Lu Cheng (2025). "Conformal prediction: A data perspective". In: *ACM Computing Surveys*.
- Zhu, Bing, Bart Baesens, Aimée Backiel, and Seppe KLM Vanden Broucke (2018). "Benchmarking sampling techniques for imbalance learning in churn prediction". In: *Journal of the Operational Research Society* 69.1, pp. 49–65.
- Zhu, Bing, Seppe Broucke, Bart Baesens, and Sebastián Maldonado (2017). "Improving Resampling-based Ensemble in Churn Prediction". In: *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR, pp. 79–91.
- Zhu, Yi, Jing Dong, and Henry Lam (2024). "Uncertainty quantification and exploration for reinforcement learning". In: *Operations Research* 72.4, pp. 1689–1709.
- Zouhal, Lalla Meriem and Thierry Denoeux (1998). "An evidence-theoretic k-NN rule with parameter optimization". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 28.2, pp. 263–271.

Index

- Belief Functions, 18
- Basic Belief Assignment, 18
- Belief and Plausibility, 18
- Belief Functions, 12
 - Basic Belief Assignment, 24
- Belief functions, 12
- Bellman Equation, 74
- Churn prediction, 12
- Conformal
 - Prediction, 12, 25, 31
 - Threshold, 22
- Dempster–Shafer Theory, 12
 - Combination Rule, 19, 24
- Focal Elements, 18
- Frame of Discernment, 18
- Fraud detection, 12
- HEUQ, 37
- Iris dataset, 12
- Masksembles, 71, 73
- MASURE, 71
- Nonconformity Score, 22
- Power Set, 18
- Q-learning, 74
- Score
 - Balanced accuracy, 56
 - Brier, 56
 - F1, 56
 - Negative Log-Likelihood, 57
- Set-valued prediction, 12
- Transformation
 - Gaussian Random Projections, 53
 - Principal Component Analysis, 53
- Uncertainty
 - Aleatoric, 3, 43, 47
 - Epistemic, 3, 12, 43, 47
 - Quantification (UQ), 2, 12
 - Total, 44, 46