DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
74

VIACHESLAV KOMISARENKO

Aligning Training Loss to Evaluation Metrics in Deep Learning

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
74

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
74

VIACHESLAV KOMISARENKO

Aligning Training Loss to Evaluation Metrics in Deep Learning

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on December 9, 2025 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Prof.	Meelis Kull University of Tartu Estonia

Opponents

Prof.	Jesús Cid-Sueiro Charles III University of Madrid Spain
Assoc. Prof.	Maurizio Filippone King Abdullah University of Science and Technology Kingdom of Saudi Arabia

The public defense will take place on January 13, 2026 at 10:00 in Narva Rd. 18-1019.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print) ISSN 2806-2345 (pdf)
ISBN 978-9908-57-097-6 (print) ISBN 978-9908-57-098-3 (pdf)

University of Tartu Press
http://www.tyk.ee/

To my family, friends, and those who defend my country

ABSTRACT

Over the past decade, advances in machine learning research have turned once-theoretical algorithms into practical tools whose performance surpasses human expertise across diverse tasks. These breakthroughs have catalysed the integration of machine learning components into nearly every layer of modern technology, reshaping sectors from healthcare and finance to transportation and entertainment.

In practice, achieving reliable performance requires substantial data and sensible optimisation choices, including tuned hyperparameters and appropriate loss functions. Among training losses, cross-entropy is often the default for classification. Similarly, upstream evaluation metrics such as the Brier score are commonly used for performance evaluation. However, for many practical applications, more task-specific downstream measures are needed, raising the question of how to select the optimal training loss accordingly. This becomes especially important in applications where misclassifications are costly, predicted probabilities need to be calibrated, or non-standard, task-specific downstream metrics are used.

The thesis presents three interconnected studies that establish theory-backed and empirically validated approaches to align deep learning training loss selection with application-driven downstream evaluation metrics.

The first study addresses cost-sensitive classification, where misclassifying different classes incurs unequal costs, and the goal is to minimise the total cost over all instances. Recognising that true cost values are often unavailable during training, we formalise cost uncertainty, modelling class costs as a probability distribution rather than a fixed scalar, approximated from domain experts' opinions, for example. By analysing how total cost behaves under existing and newly derived training losses in this uncertain setting, we derive practical guidelines for training and model selection in cost-sensitive applications.

The second study investigates theoretical reasons behind the empirical success of the focal loss for probability calibration, which assesses how well predicted probabilities align with actual class frequencies. Through a rigorous theoretical analysis using proper loss theory, we derive its decomposition into a proper loss component and a calibration map component, highlight its connection to temperature scaling, and introduce a novel calibration method inspired by these insights. Moreover, we extend these results to the class of separable losses, in which the loss depends only on the predicted probability assigned to the true class.

Finally, the third study addresses the evaluation mismatch between commonly used regression upstream metrics and task-specific downstream measures. We propose a data-driven method for learning a proxy function to align upstream and downstream measures, and we investigate theoretically when this alignment can preserve the measures' properness.

Taken together, these studies advance the field of machine learning by providing clear guidelines for selecting training loss functions that directly improve application-oriented metrics, such as expected total cost and calibration error.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC Accuracy. 21
AUC Area under the Receiver Operating Characteristic curve. 17, 23, 37

CDF Cumulative density function. 63
CRPS Continuous Ranked Probability Score. 17, 57

ECE Expected calibration error. 11, 17, 28, 47, 52
ERM Expected risk minimisation. 22

LLM Large language model. 17, 21, 62–64

MAE Mean absolute error. 21, 23, 58
MC Misclassification cost. 32
ML Machine learning. 17, 21, 32, 36, 62
MMCE Maximum Mean Calibration Error. 25
MSE Mean squared error. 21, 23, 37, 58

NN Neural network. 62

OOD Out-of-distribution. 54

PDF Probability density function. 33, 35

RLHF Reinforcement learning from human feedback. 63
RQ Research question. 18

SMOTE Synthetic Minority Over-sampling Technique. 28
SVM Support vector machines. 28

LIST OF MATHEMATICAL SYMBOLS AND NOTATION

Symbol	Description
$\mathbf{x} \in \mathscr{X}$	Feature/input vector (bold lowercase - vectors).
$y \in \mathscr{Y}$	Label/target; for classification, $\mathscr{Y} = \{1, \ldots, \mathrm{K}\}$.
K	Number of classes (classification).
$\mathscr{D}$	A dataset; e.g., $\mathscr{D}_{\text{val}}$ validation set.
n, m	Sizes of the training and validation sets, respectively.
P_{XY}	Joint distribution of (X, Y).
$\mathscr{P}$	(Generic) true data-generating distribution (e.g., on outcomes or labels).
$\mathscr{Q}_\theta(\cdot \mid \mathbf{x})$	Predictive distribution produced by model f_θ.
$\Delta(\mathscr{Y})$	Probability simplex over $\mathscr{Y}$. If $\mathscr{Y} = \{1, \ldots, \mathrm{K}\}$: $\Delta^{\mathrm{K}-1} = \{\mathbf{p} \in [0,1]^{\mathrm{K}} : \sum_k p_k = 1\}$.
$\Delta_\circ^{\mathrm{K}-1}$	Open simplex $\{\mathbf{p} \in (0,1)^{\mathrm{K}} : \sum_k p_k = 1\}$.
$\mathbf{1}_{\mathrm{K}}$	All-ones vector in $\mathbb{R}^{\mathrm{K}}$.
$\mathbb{E}[\cdot]$	Expectation.
$R(f), \widehat{R}_n(f)$	Population and empirical risk.
$L(\cdot, \cdot)$	Training loss (pointwise), arguments: prediction, target; e.g., $L\big(f_\theta(\mathbf{x}), y\big)$.
$S(\mathscr{Q}, y)$	Scoring rule; proper if minimised in expectation at the truth.
S_{Brier}, $S_{\log}$	Brier score; cross-entropy (log loss).
∇_q	Gradient w.r.t. probability vector q on the simplex.
M, $\widehat{M}_{\text{val}}$	Evaluation metric and its empirical (validation) estimate.
ECE	Expected Calibration Error (confidence/top-label).
$\mathrm{acc}(b)$, $\mathrm{conf}(b)$	Empirical accuracy and confidence in bin b.
$t \in [0,1]$	Decision threshold (e.g., binary classification; used in $\hat{y} = \mathbb{I}\{p \geq t\}$).
$\phi : \Delta^{\mathrm{K}-1} \to \Delta^{\mathrm{K}-1}$	Calibration map (e.g., temperature scaling, focal calibration).
$k : (0,1] \to (0, \infty)$	Weight/shape function inducing ϕ via $\phi_i(\mathbf{p}) = k(p_i) / \sum_j k(p_j)$.
γ	Focal parameter of the focal loss.
T	Temperature parameter in temperature scaling.
$C = (c_{ij})$	Cost matrix; c_{ij} is cost of predicting j when i is true.
C_{tot}	Total (expected) misclassification cost.
$\mathbb{I}\{\cdot\}$	Indicator function.
$L^u(\mathscr{Q}, y)$, $L^d(\mathscr{Q}, y)$	Upstream metric (generic performance) and downstream utility (task-/decision-specific).

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

I. **Komisarenko**, **Viacheslav** and Kull, Meelis (2025). "Cost-sensitive classification with cost uncertainty: do we need surrogate losses?" In: *Machine Learning* 114.132, pp. 1–36. DOI: 10.1007/s10994-024-06634-8.

II. **Komisarenko**, **Viacheslav** and Kull, Meelis (2024). "Improving Calibration by Relating Focal Loss, Temperature Scaling, and Properness". In: *European Conference on Artificial Intelligence*. IOS Press, pp. 1535–1542. DOI: 10.3233/FAIA240658.

III. Shahroudi, Novin, **Komisarenko**, **Viacheslav**, and Kull, Meelis (2025). "Aligning the Evaluation of Probabilistic Predictions with Downstream Value". In: *European Conference on Artificial Intelligence*. IOS Press, pp. 1969–1976. DOI: 10.3233/FAIA251032.

Publications not included in the thesis

Other published work of the author

I. Burdun, Iuliia, Bechtold, Michel, Aurela, Mika, De Lannoy, Gabrielle, Desai, Ankur R, Humphreys, Elyn, Kareksela, Santtu, **Komisarenko**, **Viacheslav**, Liimatainen, Maarit, Marttila, Hannu, et al. (2023). "Hidden becomes clear: Optical remote sensing of vegetation reveals water table dynamics in northern peatlands". In: *Remote Sensing of Environment* 296. DOI: 10.1016/j.rse.2023.113736.

II. **Komisarenko**, **Viacheslav**, Voormansik, Kaupo, Elshawi, Radwa, and Sakr, Sherif (2022). "Exploiting time series of Sentinel-1 and Sentinel-2 to detect grassland mowing events using deep learning with reject region". In: *Scientific Reports* 12.1. DOI: 10.1038/s41598-022-04932-6.

III. Ingel, Anti, Shahroudi, Novin, Kängsepp, Markus, Tättar, Andre, **Komisarenko**, **Viacheslav**, and Kull, Meelis (2020). "Correlated daily time series and forecasting in the M4 competition". In: *International Journal of Forecasting* 36.1, pp. 121–128. DOI: 10.1016/j.ijforecast.2019.02.018.

IV. Burdun, Iuliia, Bechtold, Michel, Sagris, Valentina, **Komisarenko**, **Viacheslav**, De Lannoy, Gabrielle, and Mander, Ülo (2020). "A Comparison of Three Trapezoid Models Using Optical and Thermal Satellite Imagery for Water Table Depth Monitoring in Estonian Bogs". In: *Remote Sensing* 12.12. DOI: 10.5194/egusphere-egu21-4698.

V. Gubarev, Vyacheslav F, Boyun, Vitaliy P, Melnichuk, Sergey V, Salnikov, Nikolay N, Simakov, Vladimir A, Godunok, Leonid A, **Komisarenko**, **Vyacheslav** I, Dobrovolskyy, Victor Yu, Derkach, Sergey V, and Matviyenko, Sergey A (2016). "Using Vision Systems for Determining the Parameters of Relative Motion of Spacecrafts". In: *Journal of Automation and Information Sciences* 48.11. DOI: 10.1615/JAutomatInfScien.v48.i11.30.

Author's contribution to the publications

In Publication I and II, the author was responsible for all stages of the research, including formulating the idea and hypotheses, developing the methodology, conducting experiments, and writing the publication.

In Publication III, the author derived and proved the theoretical results that justify the proposed method and wrote Section 3 and Appendix A.1.

1. INTRODUCTION

Over the last decade, the volume of available data has surged, enabling practical machine learning (ML) deployments across domains and incentivising ever larger models with more parameters to learn (Hilbert and López 2011). Yet despite rapid architectural change, the training objective (loss functions) has remained fairly stable: cross-entropy (negative log-likelihood) is the de facto choice for classification and autoregressive language modelling (next-token prediction in large language models (LLMs)) (Brown et al. 2020; Ashish 2017). Even modern LLMs first use pure cross-entropy pre-training, and are only then fine-tuned with reinforcement-learning-style optimisation using learned reward models (Ouyang et al. 2022). For probabilistic evaluation, proper Brier score (for binary/multiclass probabilistic classification) and the Continuous Ranked Probability Score (CRPS, for probabilistic regression) are long-standing standards for evaluation (Gneiting and Raftery 2007; Hersbach 2000b).

At the same time, application domains mandate extending the standard set of losses and evaluation metrics beyond "accuracy-only" thinking. Different tasks value different mistakes: missing a disease can be worse than a false alarm, or an overoptimistic forecast can be riskier than a slightly pessimistic one. Metrics like accuracy, which weight all errors equally, give a baseline view but overlook the real-world context and consequences. This is why practitioners use metrics like precision/recall, AUC, calibration error, cost-sensitive scores, and even utility-based measures that better reflect real-world trade-offs.

Classical losses, most notably cross-entropy, arose from probabilistic modeling as the negative log-likelihood for classical small-scale models such as logistic regression, and became standard for classification (Bishop and Nasrabadi 2006). As architectures scaled to deep networks, cross-entropy remained the default objective and delivered strong accuracy on benchmarks (Krizhevsky, Sutskever, and G. E. Hinton 2012; K. He et al. 2016). However, follow-up work that analysed the suitability of predicted probabilities (i.e., how well model confidences reflect actual outcomes), which is a key requirement for downstream decision-making, found frequent overconfidence and related calibration issues (Guo et al. 2017). Part of this behaviour is in fact unavoidable: modern training and model selection procedures effectively perform a multiple hypothesis search over many parameter configurations and then choose the most favourable one, which naturally induces optimistic and overconfident predictions unless additional structural assumptions are imposed. Nevertheless, such findings motivate the use of metrics beyond accuracy and the development of new, or modifications of existing, loss functions that better account for these effects.

Consequently, loss modifications such as regularisation were proposed to mitigate overfitting and improve generalisation (Srivastava et al. 2014; Lienen and Hüllermeier 2021); reliability concerns create a need for calibration methods (e.g., post-hoc calibration and ECE monitoring (Platt et al. 1999)); and class imbalance

motivates cost-sensitive metrics and new losses, such as focal loss, which can outperform classical choices (Elkan 2001; T.-Y. Lin et al. 2017). Despite this progress, there is no unified framework to guide the choice of loss, its modifications/regularisation, and the evaluation metric across domains, architectures, or modalities. Practitioners still rely on trial-and-error or reported findings from often noncomparable settings, highlighting the need for practical guidelines on loss-metric alignment.

Every training loss and evaluation metric can be described by their mathematical properties, including, for example, convexity and properness. In theory, strictly proper losses have the favourable property of being uniquely minimised when predicting the true class-posterior probabilities. In practice, however, the loss rarely attains its global optimum due to finite samples, optimisation difficulties, and data and model imperfections, so properness alone does not guarantee superior performance away from that point. Nevertheless, largely because of their strong empirical performance, many widely used losses and metrics are proper, and most refinements, such as regularisation, post-hoc calibration, and loss-modulating factors, are layered on top of these proper losses, rather than introducing entirely new, improper families.

We believe that proper scoring rule theory could provide a principled basis for selecting or designing losses that align optimisation with the domain's evaluation metric, narrowing the gap between predictive performance and downstream utility.

This thesis comprises three primary investigations that collectively seek to advance the state of the art in loss-metric alignment in machine learning, with a particular focus on proper loss theory, cost-sensitive learning and calibration. To this end, we address the following research questions (RQ):

- Which loss function should be used to minimise cost-sensitive evaluation metrics (e.g., expected total cost), especially in the presence of class-cost uncertainty?
- Which proper loss families perform best in cost-sensitive settings when the evaluation metric exactly matches the training loss?
- Why does the (improper) focal loss yield better calibration than the commonly used (proper) cross-entropy loss as reported by (Mukhoti et al. 2020), and how does focal loss relate to properness?
- Can we decompose any separable loss into a proper component and a fixed calibration map component, and can some separable losses perform comparably to cross-entropy?
- Can upstream and downstream regression evaluation be aligned via a proxy function learned on validation data?
- Under what assumptions on the transformations do both the upstream and downstream metrics remain proper?

To explore these questions, in **Chapter 2**, we first survey the theory of proper

losses, calibration methods, and cost-sensitive learning, laying the groundwork for the contributions that follow.

Chapter 3 addresses the problem of choosing the classification loss function for minimising the expected class-dependent cost, a metric widely used in real-world applications, as it maps model errors directly to the funds required to mitigate them. It considers the practical scenario in which misclassification costs are unknown during training and are available only as probability distributions, constructed, for example, from domain experts' estimates. Exact cost values are realised at deployment by sampling from these distributions. In practice, these cost distributions are often most uncertain in the tail, where rare but costly errors occur. Our analysis is therefore conditional on the specified cost distributions and does not resolve the underlying uncertainty about such high-cost events. We then derive a family of loss functions that are mathematically equivalent to the expected total cost objective under class cost uncertainty and demonstrate, via extensive experiments, that optimising appropriately chosen members of this family consistently delivers lower realised total cost. [Paper 1 ***Cost-sensitive classification with cost uncertainty: do we need surrogate losses?*** (**Komisarenko** and Kull 2025)].

Chapter 4 investigates why certain *improper* classification losses, such as focal loss, often yield better calibrated probability estimates than the widely used proper losses (i.e., losses minimised by the true class-posterior probabilities), including Brier score and cross-entropy. We begin by establishing new theoretical properties of focal loss, showing that it can be decomposed into (i) a proper base loss and (ii) a fixed, rank-preserving transformation. This decomposition reveals an unexpected connection to temperature scaling and motivates a novel two-parameter family of post-hoc calibration maps. Extensive experiments demonstrate that the proposed maps consistently improve calibration while maintaining predictive accuracy. [Paper 2 ***Improving Calibration by Relating Focal Loss, Temperature Scaling, and Propernes*** (Komisarenko and Kull 2024)].

We extend the decomposition results to the class of separable losses, where the loss depends only on the predicted probability assigned to the true class. We propose several previously unstudied separable losses and derive corresponding families of calibration maps. Experiments indicate that these separable losses, especially when paired with the corresponding post-hoc calibration, can outperform standard temperature-scaled cross-entropy or focal-loss–trained models, in both predictive accuracy and calibration, provided the calibration map family and its hyperparameters are selected on a validation set.

Chapter 5 (third investigation) reframes regression metric choice as an *evaluation alignment* problem: prediction quality should be judged not only by standard evaluation metrics but, more importantly, in the context of downstream use. We investigate which transformations preserve properness and propose a data-driven proxy evaluation that aligns upstream assessment with downstream utility via learned task-specific mappings. This approach removes the need for hand-

crafted cost structures and long lists of task-specific metrics, enabling fast, scalable evaluation when the relevant weighting is complex or unknown a priori. Synthetic and real-data regression experiments demonstrate improved alignment between predictive evaluation and downstream utility. [Paper 3 ***Aligning the Evaluation of Probabilistic Predictions with Downstream Value (Publication III)*** (Shahroudi, **Komisarenko**, and Kull 2025)].

Chapter 6 summarises the thesis's main findings, acknowledges current limitations, and proposes potential directions for future research. Together, the three interconnected studies deepen our theoretical and empirical understanding of proper loss theory, cost-sensitive learning, and calibration methods, while offering practical guidelines that help practitioners build more reliable machine learning systems. In particular, we provide actionable guidance on which loss functions are most beneficial in practice when the objective is to optimise a specified evaluation metric (e.g., expected total cost or expected calibration error).

2. BACKGROUND

This chapter provides the background for the remainder of the thesis. We establish notation and review supervised learning, proper scoring rules, uncertainty in ML, model calibration, cost-sensitive learning, and the alignment between upstream and downstream measures.

2.1. Main definitions and notation

We establish notation and basic definitions used throughout. Unless stated otherwise, vectors are *bold lowercase* (e.g., $\mathbf{x}$), random variables are *uppercase italic* (e.g., Y), distributions are *calligraphic* (e.g., $\mathscr{D}$), fixed scalars are *upright (roman)* (e.g. K), and expectations are denoted by $\mathbb{E}[\cdot]$. We now review core concepts in supervised learning and the theory of proper scoring rules using this notation, concluding with training loss, upstream metric and downstream utility alignment.

2.1.1. Supervised learning

Supervised learning, the task of inferring a predictive mapping from labeled input-output examples so that it generalises to unseen data, has a long and rich research history, with conceptual roots in 19th-century work on least-squares regression and early studies of correlation by Galton (Galton 1886) and Pearson (Pearson 1896). Key milestones that have accelerated its development and performance include Rosenblatt's perceptron (Rosenblatt 1958), the advent of back-propagation (Rumelhart, G. E. Hinton, and Williams 1986), the rise of statistical learning theory and support-vector machines (Cortes and Vapnik 1995), the breakthroughs of AlexNet (Krizhevsky, Sutskever, and G. E. Hinton 2012), and, more recently, the Transformer architecture (Ashish 2017), fundamentally self-supervised yet could be fine-tuned for supervised tasks, along with its large language model (LLM) successors (Radford et al. 2018).

Formally, supervised learning considers an input space $\mathscr{X}$ and a label or target space $\mathscr{Y}$. In *classification*, $\mathscr{Y}$ is a finite set of discrete labels, e.g., $\{1,\dots,\mathrm{K}\}$. The learner assigns each input to one of several categories, and performance is often evaluated by accuracy (ACC), the frequency of correct predictions. In *regression*, $\mathscr{Y} \subseteq \mathbb{R}^d$ is (typically) a real-valued target space rather than a finite set of labels, and the task is to predict a scalar or vector in $\mathbb{R}^d$. We usually assume that Y has finite moments (e.g., $\mathbb{E}\|Y\|^2 < \infty$) so that standard error measures such as mean squared error (MSE) or mean absolute error (MAE) are well defined. Typically, no additional smoothness or continuity assumptions on the distribution of Y (such as the absence of gaps or the existence of a density) are required at this level of generality; whenever stronger assumptions are needed in later sections, we state them explicitly.

We observe an i.i.d. sample $\big((\mathbf{x}_i, y_i)\big)_{i=1}^{n}$, usually called the training dataset,

drawn from an unknown joint distribution P_{XY} over $\mathscr{X} \times \mathscr{Y}$. We seek a hypothesis $f \in \mathscr{F}$, where $\mathscr{F} \subseteq \{f : \mathscr{X} \to \mathscr{A}\}$ is a class of functions mapping inputs to a prediction space $\mathscr{A}$. In the simplest settings we take $\mathscr{A} = \mathscr{Y}$ (e.g., regression or hard-label classification), while for probabilistic classification we may use $\mathscr{A} = \Delta(\mathscr{Y})$, the probability simplex over $\mathscr{Y}$, i.e., the set of all probability distributions on $\mathscr{Y}$. Concretely, if $\mathscr{Y} = \{1, \ldots, \mathrm{K}\}$, then

$$\Delta(\mathscr{Y}) = \big\{\mathbf{p} = (p_1, \ldots, p_{\mathrm{K}}) \in [0,1]^{\mathrm{K}} : \textstyle\sum_{k=1}^{\mathrm{K}} p_k = 1\big\}.$$

The set $\mathscr{F}$, known as the *hypothesis space*, is determined a priori by the chosen family of functions, often called the *model class* or *architecture* (e.g., linear models, decision trees, neural networks).

In addition, we assume access to a *validation dataset* $\mathscr{D}_{\mathrm{val}} = \big((\mathbf{x}_j^{\mathrm{val}}, y_j^{\mathrm{val}})\big)_{j=1}^{m}$, consisting of examples that are not used during training (for example, a held-out subset of the original sample or an independent i.i.d. sample from P_{XY}).

We adopt the standard empirical risk minimisation (ERM) perspective, where the goal of the *training process* is to minimise the **expected** (population) **risk**.

$$R(f) \;=\; \mathbb{E}_{(X,Y)\sim P_{XY}}\big[L\big(f(X), Y\big)\big],$$

where $L : \mathscr{A} \times \mathscr{Y} \to \mathbb{R}$ is a user-chosen (task-specific) loss that encodes the cost of predicting $f(\mathbf{x})$ when the true label is y.

Because P_{XY} is unknown, we instead minimise the empirical risk

$$\widehat{R}_n(f) \;=\; \frac{1}{n}\sum_{i=1}^{n} L\big(f(\mathbf{x}_i), y_i\big),$$

often augmented with a regulariser to control model capacity and promote generalisation.

While the loss function drives optimisation by determining the hypothesis f (model parameters) that minimises empirical risk on the training set, in practice, we ultimately care about our *evaluation metric of interest* M, measured on data that were not used for training, typically the validation dataset $\mathscr{D}_{\mathrm{val}}$. We evaluate

$$\widehat{M}_{\mathrm{val}}(f) = M\big((f(\mathbf{x}_j^{\mathrm{val}}), y_j^{\mathrm{val}})_{j=1}^{m}\big),$$

or, for decomposable metrics M, we use its corresponding pointwise form M_{pt} and define

$$\widehat{M}_{\mathrm{val}}(f) = \frac{1}{m}\sum_{j=1}^{m} \mathsf{M}_{\mathrm{pt}}\big(f(\mathbf{x}_j^{\mathrm{val}}), y_j^{\mathrm{val}}\big).$$

We then select a model $\hat{f}$ that optimises $\widehat{M}_{\mathrm{val}}$, typically maximising it when larger values indicate better performance and minimising it when smaller values indicate better performance, according to the convention of the chosen metric.

2.1.2. Aligning Upstream Metrics with Downstream Utility

In practice, L and M usually differ. The metric M is dictated by task requirements, whereas the training loss L must first have good optimisation properties, including differentiability (often with well-behaved gradients), numerically stability, and suitability to stochastic gradient-based methods (Goodfellow, Bengio, and Courville 2016). Moreover, among evaluation metrics M it is useful to distinguish: *upstream* measures, which summarise model performance in general (e.g., MSE/MAE for regression, accuracy/AUC/log loss for classification), and *downstream* utilities, which are application-specific objectives computed through the decision pipeline (e.g., expected monetary cost, policy reward, service-level penalties). The latter often depend on context variables or constraints not present at training time. While upstream metrics provide a useful high-level summary of model performance, choosing between models with similar upstream scores for task-specific needs requires measuring their downstream utilities - the quantities that ultimately matter for the application. This distinction motivates explicit alignment between the training loss, common upstream metrics, and the downstream utility. By alignment we mean choosing or designing the training objective (and associated decision rule/post-processing) so that improvements in the optimised loss reliably translate into gains in the target upstream metric and, ultimately, the downstream utility.

2.1.3. Proper scoring rules

Let a model f_θ produce, for each input $\mathbf{x} \in \mathscr{X}$, a predictive distribution over outcomes, denoted $\mathscr{Q}_\theta(\cdot \mid \mathbf{x})$ (for classification, a categorical distribution over classes; for regression, a distribution on $\mathscr{Y} \subseteq \mathbb{R}^d$). A (probabilistic) scoring rule is a function $S(\mathscr{Q}, y)$ that assigns a numerical score to a predictive distribution $\mathscr{Q}$ when the outcome is y. It is *proper* if its expected score is minimised when $\mathscr{Q}$ equals the true data-generating distribution $\mathscr{P}$, and *strictly proper* if this minimiser is unique (Gneiting and Raftery 2007):

$$\mathbb{E}_{Y\sim\mathscr{P}}\big[S(\mathscr{Q}, Y)\big] \;\geq\; \mathbb{E}_{Y\sim\mathscr{P}}\big[S(\mathscr{P}, Y)\big] \quad \text{for all predictive } \mathscr{Q},$$

with equality iff $\mathscr{Q} = \mathscr{P}$. In our notation, training with a proper rule corresponds to minimising the (empirical) average of $S\big(\mathscr{Q}_\theta(\cdot \mid \mathbf{x}_i), y_i\big)$, i.e., using the loss $L_S(\theta; \mathbf{x}, y) := S(\mathscr{Q}_\theta(\cdot \mid \mathbf{x}), y)$ to find optimal model parameters $\hat{\theta}$:

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n} \cdot \sum_{i=1}^{n} S\big(\mathscr{Q}_\theta(\cdot \mid \mathbf{x}_i), y_i\big).$$

In our notation, training with a proper rule corresponds to minimising the (empirical) average of $S\big(\mathscr{Q}_\theta(\cdot \mid \mathbf{x}_i), y_i\big)$, i.e., using the loss $L_S(\theta; \mathbf{x}, y) := S(\mathscr{Q}_\theta(\cdot \mid \mathbf{x}), y)$ to find optimal model parameters $\hat{\theta}$:

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} S\big(\mathscr{Q}_\theta(\cdot \mid \mathbf{x}_i), y_i\big).$$

Under mild conditions (differentiability of the scoring rule and convexity of the corresponding Bayes risk in the prediction argument), properness follows from a first-order optimality condition on the probability simplex. For each true distribution $\mathscr{P}$ on a label set$\{1,\ldots,\mathrm{K}\}$,

$$\exists\,\lambda\in\mathbb{R}:\quad \nabla_q\left(\mathbb{E}_{Y\sim\mathscr{P}}\left[S(q,Y)\right]\right)\Big|_{q=\mathscr{P}}=\lambda\,\mathbf{1}_{\mathrm{K}},$$

where $\nabla_q=(\partial/\partial q_1,\ldots,\partial/\partial q_{\mathrm{K}})$ is the gradient with respect to the probability vector q, and $\mathbf{1}_{\mathrm{K}}$ is the K-dimensional all-ones vector. Convexity in q over the simplex ensures $\mathscr{P}$ a (global) minimiser of the Bayes risk, and strict convexity yields strict properness (Gneiting and Raftery 2007; M. D. Reid and R. C. Williamson 2010).

Binary proper losses. For example, consider a binary classification problem. Let $\eta:=P(Y=1\mid\mathbf{x})\in[0,1]$ be the true posterior and $p\in(0,1)$ the model's predicted probability for class 1. Define the conditional Bayes risk as

$$\bar{S}(p;\eta):=\mathbb{E}_{Y\sim\mathrm{Bern}(\eta)}[S(p,Y)],$$

where $\mathrm{Bern}(\eta)$ denotes the Bernoulli distribution on $\{0,1\}$ with $P(Y=1)=\eta$ and $P(Y=0)=1-\eta$. In the binary case, the simplex constraint is absorbed into a single scalar p, so the first-order condition reduces to

$$\frac{\partial}{\partial p}\bar{S}(p;\eta)=0\quad\text{at } p=\eta.$$

Then *binary Brier score* $S_{\mathrm{Brier}}(p,y)=(y-p)^2$ is strictly proper, because its conditional Bayes risk is

$$\bar{S}_{\mathrm{Brier}}(p;\eta)=\eta(1-p)^2+(1-\eta)p^2=(p-\eta)^2+\eta(1-\eta).$$

Its derivative w.r.t. predictor

$$\frac{\partial}{\partial p}\bar{S}_{\mathrm{Brier}}(p;\eta)=2(p-\eta)$$

is equal to zero only at $p=\eta$, and the second derivative is strictly positive

$$\frac{\partial^2}{\partial p^2}\bar{S}_{\mathrm{Brier}}(p;\eta)=2>0.$$

Similarly, for *binary cross-entropy* (often referred to as log loss) with the loss expression

$$S_{\log}(p,y)=-y\log p-(1-y)\log(1-p)$$

and its correspondent conditional Bayes risk

$$\bar{S}_{\log}(p;\eta)=-\eta\log p-(1-\eta)\log(1-p),$$

strict properness holds because its derivative

$$\frac{\partial}{\partial p}\bar{S}_{\log}(p;\eta) = -\frac{\eta}{p} + \frac{1-\eta}{1-p} = \frac{p-\eta}{p(1-p)}$$

is zero only at $p = \eta$. The second derivative is positive

$$\frac{\partial^2}{\partial p^2}\bar{S}_{\log}(p;\eta) = \frac{\eta}{p^2} + \frac{1-\eta}{(1-p)^2} > 0$$

on $p \in (0,1)$, giving strict properness.

Properness is theoretically appealing loss characteristic as strictly proper losses are uniquely minimised only by the Bayes-optimal predictor (i.e., the true class posterior). In practice, however, the global optimum is often unattainable due to finite samples, optimisation challenges, and model/data imperfections. Still, most commonly used training losses, including the Brier score and cross-entropy, are strictly proper, which suggests, but does not by itself guarantee, practical benefits for optimisation.

However, Guo et al. (2017) (primarily on image data, with some document classification) reported that modern deep networks trained with cross-entropy tend to be overconfident, and that simple post-hoc calibration via temperature scaling substantially improves calibration. Recent studies such as (Blasiok et al. 2023) criticises the unconditional use of proper losses in practice, arguing that additional conditions are needed for proper loss optimisation to yield calibrated predictions, formalising a "local optimality" criterion under post-hoc recalibrations, providing counterexamples in restricted model classes (e.g., logistic regression), and identifying regimes where the conditions are plausibly met (e.g., sufficiently expressive predictors, including modern large models).

In contrast, many recent approaches to address overconfidence are implemented as modifications of standard proper loss training, rather than abandoning properness altogether. Focal loss augments cross-entropy to downweight easy examples (T.-Y. Lin et al. 2017). Label smoothing replaces hard targets with softened ones, which can improve calibration and robustness (Müller, Kornblith, and G. E. Hinton 2019). The maximum mean calibration error approach (MMCE) adds a differentiable calibration penalty alongside cross-entropy during training (Kumar, Sarawagi, and Jain 2018). Finally, regularisation terms often extend cross-entropy with additive penalties (e.g., L1, L2 weight norms) to mitigate overconfidence and overfitting (Bishop and Nasrabadi 2006).

Taken together, while the theoretical advantages of proper losses are clear, their practical performance is not automatically optimal; nevertheless, training with proper losses or their modifications remains a principled and often effective approach.

2.2. Uncertainty in Machine Learning

An ever-growing number of machine learning systems are now deployed, and their predictions are trusted to drive decisions across nearly every aspect of daily life, yet those predictions remain inherently uncertain. The reliability of these predictions is limited by finite and noisy data, embedded in imperfect models, and exposed to environments that can shift unexpectedly. This uncertainty arises from two main sources (Der Kiureghian and Ditlevsen 2009; Hüllermeier and Waegeman 2021). *Aleatoric uncertainty* arises from variability in the data-generating process, including, for example, sensor inaccuracies, measurement error, label noise, or inherent randomness. It is irreducible given a fixed information set (features, sensors, and data-collection protocol): collecting more data of the same kind does not reduce this variability (Kendall and Gal 2017). Richer measurements or additional features may convert part of this variability into explained signal. *Epistemic uncertainty*, by contrast, comes from gaps in the model's knowledge, regions of the input space for which it has seen few or no examples, and therefore mirrors our incomplete understanding of the underlying process. In principle, it can be reduced by gathering more representative data or adopting more expressive models. Understanding how aleatoric and epistemic uncertainties arise, evolve, and interact is essential for managing system performance rationally and safely in real-world conditions.

In many practical settings, it is not feasible to cover every region of the input space with dense, representative data, because the long tail of rare scenarios and irreducible randomness guarantees that some areas remain data-poor. Yet deployed models must still deliver reliable predictions across this entire domain, including those sparse or unseen regions.

To mitigate this problem, several strategies could be considered. Firstly, during evaluation and model selection, we can emphasise metrics that assign extra weight to rare or high-stakes regions, so that models are rewarded for performing well where mistakes would be most costly. One pragmatic option is to keep the training procedure for a single model unchanged and, among the multiple candidates produced by the same pipeline (e.g., different random seeds, checkpoints, or hyperparameter settings), select the model that scores highest under the chosen cost-aware metric. A more proactive approach is to embed the cost structure directly into the training objective, enabling the model to learn from the outset to minimise expected cost.

Another practical challenge is obtaining reliable confidence estimates for each prediction, even when a model's discrimative power, including accuracy, is high. Such estimates are crucial in high-stakes settings, where recognising a model's own uncertainty guides downstream planning, risk management, and human oversight. Although metrics for evaluating reliability exist, the optimal way to train, fine-tune, or calibrate a model to produce reliable probabilities remains unsettled. In principle, minimizing a *proper* loss should yield reliable model's confi-

dence, yet in practice the most common proper loss, cross-entropy, often produces overconfident predictions and, therefore, requires post-hoc calibration (Guo et al. 2017). Gaining a clearer picture of how training losses, calibration transformations, and evaluation metrics interact is thus critical for constructing guidelines for practitioners for improving real-world applications performance.

2.2.1. Cost-sensitive learning

Alongside advances that drive raw discriminative performance (e.g., accuracy, squared error, log-loss), extensive work has adapted supervised learning to diverse practical domains and operational constraints. Among them, cost-sensitive and class-imbalance learning arise frequently in real applications, in which misclassification errors have unequal costs or classes have substantially different frequencies.

Consider a **cost matrix**

$$C = (c_{ij}) \in \mathbb{R}^{\mathrm{K}\times\mathrm{K}},$$

where c_{ij} is the cost of predicting class j when the true class is i. Typically, $c_{ii} = 0$ for all i, and $c_{ij} > 0$ for $i \neq j$, representing the penalty for that misclassification. In class-imbalanced settings, a common heuristic is to scale costs **inversely with class frequency** (H. He and Garcia 2009), e.g., $c_{ij} \propto 1/\pi_i$ for $i \neq j$, where $\pi_i = P(Y = i)$ or its empirical estimates $\pi_i \approx n_i/n$.

The goal of cost-sensitive learning is to minimise evaluation metrics that account for typically asymmetric misclassification costs, as specified by a cost matrix. For example, the classical error rate can be *cost-weighted* to produce the *total (misclassification) cost*, the most common metric in cost-sensitive learning. Given random variables Y and $\hat{Y}$ for the true and predicted labels, the (population) total cost is defined as the expected cost per prediction:

$$C_{\mathrm{tot}}(Y,\hat{Y}) \;:=\; \mathbb{E}\Big[c_{Y,\hat{Y}}\Big] \;=\; \sum_{i=1}^{\mathrm{K}}\sum_{j=1}^{\mathrm{K}} c_{ij}\,\Pr(Y=i,\hat{Y}=j).$$

Given a decision rule $\hat{y}(\mathbf{x})$, the **empirical total cost** over a dataset $\mathscr{D} = \big((\mathbf{x}_t, y_t)\big)_{t=1}^{n}$ is

$$\widehat{C}_{\mathrm{tot}}(\mathscr{D},\hat{y}) = \frac{1}{n}\sum_{t=1}^{n} c_{y_t,\hat{y}(\mathbf{x}_t)}.$$

When all off-diagonal costs equal 1 and $c_{ii} = 0$, $\widehat{C}_{\mathrm{tot}}$ reduces to the error rate.

Beyond class-dependent matrices, many applications use instance-dependent (example-dependent) costs, where the penalty depends on the input $\mathbf{x}$, e.g. $c_{y,\hat{y}}(\mathbf{x})$ reflecting, for example, patient severity, transaction value, or regulatory constraints (Zadrozny, Langford, and N. Abe 2003). In such cases, the learning target is still an expected total cost.

Still, it is not a settled question which training loss best minimises cost-sensitive metrics such as total cost. Because total cost is defined after a discrete decision rule, it is piecewise-constant and non-differentiable, so it cannot be optimised directly with gradient-based training of deep neural networks. In practice, a common pipeline is to train with a surrogate loss (e.g., cross-entropy or another proper loss) and inject costs at one or more training levels.

At the data level, proportional sampling schemes are widely used: oversampling (duplicates the samples from minority class) is often a strong baseline (Buda, Maki, and Mazurowski 2018), undersampling reduces the majority class (Japkowicz and Stephen 2002), and hybrid approaches such as SMOTE and its variants synthesize new minority examples via augmenting (Ramentol et al. 2012; Chawla et al. 2002). These techniques target class imbalance, but can be adapted to cost imbalance by replacing class ratio heuristics with cost ratios; they also extend to instance-dependent settings via per-example sampling probabilities (Zadrozny, Langford, and N. Abe 2003). Such data-level strategies are straightforward to apply with deep neural networks.

At the model level, one can (i) modify the training loss (e.g., class-weighted losses, cost-sensitive SVM/logistic, margin adjustments) (Elkan 2001; Hastie, Tibshirani, and Friedman 2009; Kukar, Kononenko, et al. 1998; Y. Lin, Lee, and Wahba 2002; Iranmehr, Masnadi-Shirazi, and Vasconcelos 2019), (ii) adjust thresholds or decision rules post-training according to cost scenario (Elkan 2001; Lipton, Elkan, and Naryanaswamy 2014; Ye et al. 2012), and (iii) calibrate or rescale probabilities before cost-aware decision making (Collell, Prelec, and Patil 2016). Hybrid methods combine data- and model-level ideas and are surveyed in (Johnson and Khoshgoftaar 2019). Many of these approaches port cleanly to neural networks (e.g., weighted losses, cost-aware thresholds, post-hoc calibration), providing practical alternatives when direct optimisation of total cost is infeasible.

2.2.2. Model calibration

Model calibration is an important aspect of classification performance, assessing the quality of predicted probabilities. This matters because many downstream decisions assume those probabilities are numerically meaningful and are close to true outcome frequencies; miscalibration can therefore degrade real-world performance even when accuracy is high. The most common calibration measure is the Expected Calibration Error (ECE), which averages the absolute gap between predicted confidence and observed accuracy across probability bins. In multiclass settings, several variants of ECE have been proposed, including confidence (top-label), full-class, classwise, and marginal ECE, which differ in how they aggregate calibration across classes (Silva Filho et al. 2023). Among these, confidence ECE is the most widely used in practice, and we adopt this definition in what follows. Let $Y \in \{1,\dots,\mathrm{K}\}$ denote the true class label, $\hat{p} = (\hat{p}_1,\dots,\hat{p}_\mathrm{K})$ the predicted probability vector, and $\hat{y} = \arg\max_k \hat{p}_k$ the predicted class (ties broken by

any fixed deterministic rule). Let $\hat{p}_{\max} = \max_k \hat{p}_k$ denote the model's confidence. The confidence-ECE is then defined as

$$\mathrm{ECE} = \mathbb{E}_{\hat{p}_{\max}}\Big[\,\big|\,\mathbb{P}(\hat{y} = Y \mid \hat{p}_{\max}) \;-\; \hat{p}_{\max}\,\big|\,\Big],$$

where the expectation is taken over the distribution of predicted confidence scores (Gupta and Ramdas 2022). This quantity cannot be computed in practice, as the true class probabilities are unknown. Therefore, an empirical approximation of ECE is typically computed instead.

Given the validation set $\mathscr{D}_{\mathrm{val}}$, for each instance $\mathbf{x}_j^{\mathrm{val}}$, let $\hat{y}_j = \arg\max_k \mathscr{Q}_\theta(k \mid \mathbf{x}_j^{\mathrm{val}})$ and $\hat{p}_j = \max_k \mathscr{Q}_\theta(k \mid \mathbf{x}_j^{\mathrm{val}})$ denote the predicted class and its confidence. Partition $[0,1]$ into bins $\{I_b\}_{b=1}^{B}$, typically using equal-width intervals $I_b = ((b-1)/B, b/B]$. Alternative schemes, such as equal-mass (equal-frequency) binning, are also possible but lead to different empirical ECE estimates. Then define $S_b = \{j : \hat{p}_j \in I_b\}$, and for each non-empty bin S_b set

$$\mathrm{acc}(b) = \frac{1}{|S_b|} \cdot \sum_{j \in S_b} \mathbb{I}\{y_j^{\mathrm{val}} = \hat{y}_j\}, \qquad \mathrm{conf}(b) = \frac{1}{|S_b|} \cdot \sum_{j \in S_b} \hat{p}_j.$$

The empirical ECE is then

$$\widehat{\mathrm{ECE}} = \sum_{b=1}^{B} \frac{|S_b|}{m} \cdot \big|\mathrm{acc}(b) - \mathrm{conf}(b)\big|.$$

As ECE is estimated on a finite validation set using binning, it inevitably incurs approximation error and can be sensitive to the choice of binning scheme (Tygert 2025).

Poor calibration, as indicated by a high ECE, undermines the model's reliability, which is especially important in safety- or cost-critical applications, since the model's stated confidence no longer reflects real-world risks and utilities. To address model miscalibration, various calibration techniques are applied. These methods are typically grouped into training-time approaches and post-hoc approaches applied after training.

Among calibration approaches, post-hoc calibration is the most widely used: a learned transformation is applied to a trained model's logits or probabilities, leaving the underlying predictor unchanged (Silva Filho et al. 2023). The transformation is selected (and tuned) on a held-out validation set to improve the agreement between predicted and empirical probabilities, typically by minimising a strictly proper scoring rule (Gneiting and Raftery 2007). Temperature scaling is the simplest and most popular instance: a single parameter $T > 0$ scales the logits before the softmax and often yields substantial improvements with minimal complexity (Guo et al. 2017). When additional flexibility is needed, practitioners consider matrix/vector scaling (Guo et al. 2017), Bayesian binning into quantiles (BBQ) (Naeini, Cooper, and Hauskrecht 2015), and Dirichlet calibration (Kull, Perello

Nieto, et al. 2019). For a broader overview of calibration methods and practice in deep learning, see the survey by Wang et al. (C. Wang 2023).

Alternatives to post-hoc calibration include training-time approaches, which use calibration-aware losses or penalties (e.g., MMCE, label smoothing, confidence penalties (Müller, Kornblith, and G. E. Hinton 2019; Kumar, Sarawagi, and Jain 2018; Pereyra et al. 2017)), model-based methods such as deep ensembles (T. Abe et al. 2022) or Bayesian neural networks (Kendall and Gal 2017; Gal and Ghahramani 2016), and data/optimisation strategies (e.g., mixup/augmentation (Thulasidasan et al. 2019), early stopping (Minderer et al. 2021)), that aim to produce well-calibrated probabilities without post-hoc adjustment. Originally introduced to address class imbalance in object detection, focal loss has also been shown to improve calibration, often yielding competitive ECE even without post-hoc or other adjustments (T.-Y. Lin et al. 2017; Mukhoti et al. 2020).

Although these calibration methods were shown to be effective, there is no one-size-fits-all solution; the best choice depends on the dataset, model family, and deployment constraints, and should be selected via validation against the calibration metric that matters for the application.

3. BINARY COST-SENSITIVE LEARNING UNDER COST UNCERTAINTY (PUBLICATION I)

3.1. Introduction

This chapter presents the thesis's first contribution, advancing cost-sensitive learning by explicitly modelling class cost uncertainty and examining the loss functions best suited to minimise the expected total cost. Whereas prior work either assumes fixed misclassification costs or focuses on evaluation only, we treat each class cost as a sample from a known probability distribution, collected, for example, from domain experts' estimates, and focus on selecting training losses that minimise expected total cost. Guided by this objective, we derive a family of surrogate losses that is mathematically equivalent to the expected total cost expression. We then evaluate this family in extensive experiments, assessing both classical and cost-sensitive metrics and studying performance after post-hoc temperature scaling calibration. We explore the following research questions.

- Which loss function should be used to minimise cost-sensitive evaluation metrics (e.g., expected total cost), especially in the presence of class-cost uncertainty?
- Which proper loss families perform best in cost-sensitive settings when the evaluation metric exactly matches the training loss?

3.2. Motivation

Classification problems are rarely balanced: some events occur far more frequently than others, data collection often favours particular classes, and class priors can shift at deployment. Moreover, equal class sizes rarely imply equal importance: misclassification costs are asymmetric and dictated by domain considerations. In practice, such costs are commonly specified in advance (often from expert judgement) and used to convert common balanced metrics such as accuracy into cost-sensitive such as total cost. Changing the metric of interest from class-symmetric to cost-sensitive may alter which training losses and optimisation schemes are appropriate; methods tuned for accuracy may no longer be optimal. While numerous cost-sensitive methods have been proposed (see Chapter 2.2.1), important research gaps remain, which this chapter addresses.

Most studies focus on misclassification costs fixed a priori, which may be suboptimal because the initial cost settings can be inaccurate, biased, or drifted during deployment. A more robust approach is to explicitly model uncertainty around the specified costs. Prior work that incorporates cost uncertainty often treats it only at the evaluation stage or is limited to classical models (e.g., decision trees). In contrast, we derive loss functions that minimise cost-sensitive metrics under cost uncertainty and evaluate them on deep learning architectures.

In many practical settings, uncertainty about misclassification costs is greatest precisely where costs are highest, and some high-penalty situations may be underrepresented in the data because systems are designed to avoid them. Our approach can encode such patterns by specifying distinct, potentially heavy-tailed cost distributions for different classes or error types and then training with a loss that coincides with the resulting expected misclassification cost. Nevertheless, because training still follows an empirical risk minimisation paradigm under the assumed cost distributions, performance remains constrained by data coverage and does not, by itself, correct for systematic underrepresentation of rare, high-cost scenarios.

3.3. Main findings

We derive proper losses from the expected total cost expression under cost uncertainty and outline a few probability distributions with potential practical value. We benchmarked proposed losses against commonly used cross-entropy, focal loss and label smoothing in image classification tasks, evaluating performance on both standard (e.g., accuracy, log loss) and cost-sensitive metrics (total cost). Detailed findings are presented in the subsections below.

3.3.1. Theoretical findings

We first formulate the commonly used misclassification cost (MC) in *binary* case with fixed costs, where $c_0 := c_{01} > 0$ is the cost of a false positive and $c_1 := c_{10} > 0$ is the cost of a false negative; costs of correct decisions are equal to zero: $c_{00} = c_{11} = 0$. Let $p \in [0,1]$ denote the predicted probability of the positive class ($y = 1$) and $y \in \{0,1\}$ the true class label. Then, for a decision threshold $t \in [0,1]$ and indicator function $\mathbb{I}\{\cdot\}$ (equals 1 if the argument event holds, 0 otherwise).

$$\mathsf{MC}(c_0, c_1, p, y) = c_0 \cdot \mathbb{I}\{y = 0, p \geq t\} + c_1 \cdot \mathbb{I}\{y = 1, p < t\}.$$

Following (Hernández-Orallo, P. Flach, and Ferri 2012), it is natural to reparameterise costs into *magnitude* $b = c_1 + c_0$ and *proportion* $c = \frac{c_0}{c_1 + c_0}$, thereby separating scale from ratio and yielding cleaner and more intuitive decision rules. Under this reparameterisation, the MC can be rewritten as:

$$\mathsf{MC}(b, c, p, y) = b \cdot c \cdot \mathbb{I}\{y = 0, p \geq t\} + b \cdot (1 - c) \cdot \mathbb{I}\{y = 1, p < t\}.$$

Assuming calibrated predicted probabilities, the Bayes-optimal decision threshold (i.e., the one that minimises total cost) is equal to the cost proportion $t = c$ (Hernández-Orallo, P. Flach, and Ferri 2012). We therefore use this threshold throughout.

Introducing uncertainty in costs is reasonable in practice. A typical applied ML workflow includes training, validation, and deployment. During training, misclassification costs are often incorporated, based, for example, on domain expert

estimates, to balance class importance. At deployment, however, actual costs may change because estimates were noisy or biased, class priors shift, or operational goals evolve. For this reason, we treat class costs as uncertain at both training and validation: we model them as random variables and optimise the expected objective over plausible cost scenarios, yielding models that are more robust to cost misspecification and shift. We note that focusing solely on the expected value does not capture higher-order characteristics of the cost distribution, such as its variance, so in principle a model with slightly higher expected cost but substantially lower variability might be preferable in risk-averse settings. In this chapter, however, we take expected total cost as our primary criterion and leave explicit control of risk measures (for example, variance or tail risk) to future work.

We model class costs as random variables C_0 and C_1 (equivalently, magnitude $B = C_0 + C_1$ and proportion $C = \frac{C_0}{C_0+C_1}$), making **no** independence assumption between them. Under independence, the cost for one class would, by construction, convey no information about the cost for the other, which is rarely realistic in practice, since class costs are often tied to a common budget, policy, or risk tolerance and tend to move together. Consequently, misclassification cost $\mathsf{MC}(B,C,p,y)$ is itself a random variable. We then define the *expected misclassification cost* under cost uncertainty as

$$\mathbb{E}_{B,C}[\mathsf{MC}(B,C,p,y)] = \mathbb{E}_{B,C}\Big[B\cdot C\cdot \mathbb{I}\{y=0, p\geq C\} + B\cdot(1-C)\cdot\mathbb{I}\{y=1, p<C\}\Big].$$

We assume that the cost proportion C has an absolutely continuous distribution on $[0,1]$, so that it admits a probability density function f_C and the expected cost admits the integral and gradient representations below. We now state a theorem that expresses the expected misclassification cost as an integral with respect to the probability density function (PDF) of cost proportion C.

Theorem 3.3.1 (Restated from Theorem 1 of **Publication I**). *Let $p \in [0,1]$ denote the predicted probability of the positive class and $y \in \{0,1\}$ the true class label. Let B and C be random variables corresponding to the cost magnitude and cost proportion such that $\mathbb{E}[B] < \infty$. Let $f_C(\cdot)$ be the PDF of C, and $w(c) = \mathbb{E}[B|C=c]\cdot f_C(c)$. If thresholding is performed with $t = C$, then the expected misclassification cost on this instance is equal to:*

$$\mathbb{E}_{B,C}\left[\mathsf{MC}(B,C,p,y)\right] = \int_0^p (c-y)\cdot w(c)dc + \mathbb{I}\{y=1\}\cdot(1-\mathbb{E}[C]).$$

Similarly, we state the corresponding result under the (C_1, C_0) parametrisation (Theorem 2, **Publication I**):

$$\begin{aligned}&\mathbb{E}_{C_0,C_1}\left[\mathsf{MC}(C_0,C_1,p,y)\right]\\ &= \int_0^p (c-y)\Big(\int_0^\infty f_{C_0,C_1}(bc, b(1-c))db\Big)dc + \mathbb{I}\{y=1\}\cdot(1-\mathbb{E}_{C_0,C_1}[C]).\end{aligned}$$

The important practical outcome from these theorems is that these expected cost expressions admit convenient gradient-based optimisation: p appears only

as a limit of integration, so by Leibniz's rule the derivative reduces to a simple boundary term. For example, for proportion-magnitude parametrisation, the gradient for a given instance is equal to:

$$\frac{\partial}{\partial p}\mathbb{E}_{B,C}\left[\mathsf{MC}(B,C,p,y)\right] = (p-y)\cdot w(p).$$

Integral and weight function representations of proper losses similar to the form above have appeared earlier in the literature. For binary probability estimation, Miller et al. (Miller, Goodman, and Smyth 1991) show that losses whose conditional risk is minimised, with respect to the predictive argument, at the posterior probability (i.e., proper losses) can be represented via suitable weighting functions on the unit interval. Cid-Sueiro et al. (Cid-Sueiro et al. 1999) extend this analysis to the multiclass setting, considering different assumptions on how to construct multiclass losses and on the structure of the principal directions. Buja et al. (Buja, Stuetzle, and Y. Shen 2005) provide a systematic treatment of binary proper scoring rules in terms of weight functions on class probabilities. Specific examples of such losses based on particular choices of weight functions, including Beta functions, are given by Guerrero-Curieses et al.(Alicia Guerrero-Curieses and Jesús Cid-Sueiro and Rocío Alaiz-Rodríguez and Aníbal R. Figueiras-Vidal 2004).

Our formulation interprets the weighting function instead as a probability density over class costs that are uncertain during training and explicitly identifies the corresponding loss as the expected misclassification cost under cost uncertainty at the instance level. Developing this cost uncertainty interpretation, and using the resulting expected cost losses as objectives for training modern deep neural networks, is the main focus of this work.

The gradient shows that optimisation properties, such as continuity, differentiability, and smoothness, can be controlled by the weighting term $w(p)$. By modelling costs with a convenient parametric family for w, we obtain well-behaved, computationally tractable gradients suitable for gradient-based training. Crucially, this lets us treat the expected misclassification cost not only as the evaluation metric M but also as a training objective L using the cost expression itself as the loss when fitting machine learning models.

We suggest the *Beta* distribution as a convenient choice for $w(p)$: by varying $\alpha,\beta > 0$ we can control cost proportion *mean* location $\alpha/(\alpha+\beta)$ and, when $\alpha,\beta > 1$, its *mode* $(\alpha-1)/(\alpha+\beta-2)$. The cost proportion *uncertainty* can be modeled by varying the distribution parameters α,β to control its variance $\alpha\beta/\left[(\alpha+\beta)^2(\alpha+\beta+1)\right]$. In our interpretation, $\alpha/(\alpha+\beta)$ represents the *expected cost proportion*, i.e., the average share of total misclassification cost attributed to the positive class; the variance quantifies the *uncertainty* around this cost split. These quantities can, at least approximately, be elicited from domain experts or estimated from historical cost data, giving the hyperparameters a tangible meaning.

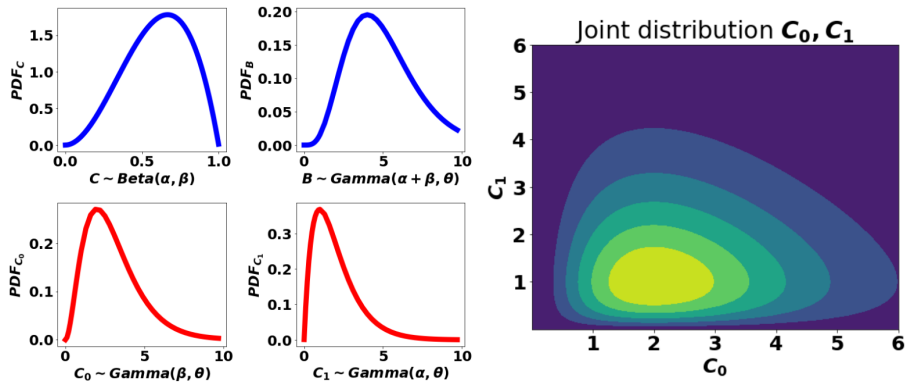

Figure 1: Class costs joint, marginal, sum and proportion distribution density functions for $\alpha = 3$, $\beta = 2$, $\gamma = 1$. The unimodality of both marginal parametrisations leads to a unimodal joint distribution, representing a practical scenario in which the most probable costs cluster near the expert-specified central region. Adapted from Publication I.

This yields a Beta family of losses with analytic, well-behaved gradients; it reduces to the *Brier* case when $\alpha = \beta = 1$ (uniform cost weighting), and approaches the *cross-entropy* in the limit $\alpha, \beta \to 0^+$, where $w(p) \propto p^{\alpha-1}(1-p)^{\beta-1} \to 1/\big(p(1-p)\big)$.

Alternatively, we suggest modelling the class costs directly as independent Gamma variables with a *same* scale parameter: $C_1 \sim \text{Gamma}(\alpha, \theta)$ and $C_0 \sim \text{Gamma}(\beta, \theta)$. Assuming a common scale parameter reflects the idea that both types of error are subject to the same underlying cost volatility (e.g., market prices, operational budgets), with α and β then primarily controlling their relative magnitudes. For a Gamma(k, θ) distribution, the mean is $k\theta$ and the variance is $k\theta^2$ (standard deviation $\sqrt{k}\,\theta$), so α, β control the class-specific cost means and variances in a convenient way. The common scale θ simplifies the following derivations: the cost magnitude $B = C_0 + C_1$ will be distributed Gamma$(\alpha+\beta, \theta)$, while the cost proportion $C = \frac{C_1}{C_0+C_1}$ will be Beta(α, β), and B and C will be independent. Hence $\mathbb{E}[B \mid C = c] = \mathbb{E}[B] = (\alpha+\beta)\theta$, and the weighting term in the gradient reduces to $w(p) = \mathbb{E}[B]\, f_C(p) = (\alpha+\beta)\theta\, p^{\alpha-1}(1-p)^{\beta-1}/\mathrm{B}(\alpha, \beta)$, where $\mathrm{B}(\cdot, \cdot)$ is the Beta function.

An example of the smooth joint cost distribution and its decomposition into (C_1, C_0) and (B, C) parametrisations for a Gamma cost model is shown in Fig. 1. The marginal and joint PDFs are smooth and unimodal, as expected for Gamma distributions with shape parameters $\alpha, \beta > 1$, with the joint density peaking near the intersection of the marginal modes, i.e., at the most plausible combination of class costs. Modelling multimodal or non-smooth cost beliefs would require alternative families, which we do not consider here.

These are only a few possible choices for cost modelling, but they are attractive in practice because they combine mathematical convenience with a flexible, inter-

pretable parametrisation: the Beta and Gamma families are standard for modelling proportions and positive-valued quantities, yield closed-form expressions for the weighting function and its gradient, and allow domain experts to specify means, modes, and variance directly through α, β, γ. In particular, they conveniently capture scenarios where the cost distribution is unimodal, with a flexible choice of the mode (most likely cost level) and a tunable degree of uncertainty around it.

3.3.2. Experimental findings

In the previous subsection, we showed how the expected misclassification cost under class cost uncertainty can be used as a training loss for ML models by deriving an expression suitable for gradient-based optimisation. We also proposed several practically convenient cost distributions, most notably a Beta family for the cost proportion parametrisation. We then benchmark this loss family on several binary image classification datasets against commonly used baseline losses: cross-entropy, focal loss, and label smoothing.

Optimal Beta loss hyperparameters. First, we used the training and validation sets of the considered datasets to gain a preliminary understanding of the Beta loss's performance across its hyperparameters. For that, we consider 99 cost uncertainty scenarios to create Beta distribution based evaluation metrics, and consider 63 different Beta hyperparameters across cost-sensitive losses. We use six binary class balanced image classification datasets, obtained by extracting **three** class pairs (cat vs horse, bird vs frog, deer vs plane) from *CIFAR-10* (Krizhevsky, Nair, and G. Hinton 2009) and **three** class pairs (Pullover vs Coat, T-shirt/top vs Shirt and Shirt vs Pullover) from *Fashion-MNIST* (Xiao, Rasul, and Vollgraf 2017) datasets. To make the task sufficiently challenging for deep networks, we injected additive Gaussian noise (mean 0, standard deviation 30) and randomly dropped a fraction of pixels, tuning the corruption level until validation accuracy fell below 90%. This makes the experiments more informative: when accuracy is very high, predicted probabilities concentrate near 0 or 1, so only extreme cost imbalances would change decisions relative to the default 0.5 threshold.

Our preliminary experiments revealed that simply training with the loss induced by the cost uncertainty distribution underlying a given evaluation metric rarely yields the best, or even competitive, performance. This observation motivated a different strategy: instead of tuning a separate loss for every cost scenario, we looked for a single set of Beta loss hyperparameters that performs reliably well across many such scenarios. To identify such a default, we benchmarked all 63 candidate Beta losses by counting how often each appeared among the top-10 losses across 99 cost metrics and six datasets (Table 1). The hyperparameters Beta(25,25) emerged as the strongest overall candidate.

Table 1: Best Beta loss parameters per dataset (three extracted class pairs from CIFAR-10 and three from Fashion-MNIST). Each column corresponds to a dataset; each row lists the best Beta loss parameters, ranked by the frequency with which they appear in the top-10 across 99 cost-sensitive metrics (among 63 candidate losses) on the validation set. Beta(25,25) was most frequently in the top-10 (among 53% of metrics). Adapted from Publication I.

Deer v Pl.	Bird v Fr.	Sh v Car	P. v Coat	Top v Sh.	Dr v P.	Total
(8, 12)	(16, 24)	(24, 16)	(15, 15)	(15, 15)	(8, 12)	**(25, 25)**: 53%
(16, 24)	(12, 18)	(12, 28)	(10, 10)	(25, 25)	(10, 10)	(20, 30): 47%
(20, 30)	(25, 25)	(20, 30)	(6, 4)	(12, 18)	(18, 12)	(18, 12): 47%
(12, 28)	(12, 28)	(18, 12)	(5, 5)	(12, 8)	(16, 24)	(20, 20): 43%
(25, 25)	(20, 30)	(30, 20)	(4, 6)	(18, 12)	(20, 20)	(24, 16): 43%

Performance against baselines on cost-sensitive metrics. We evaluated the performance of the Beta(25, 25) loss against focal, cross-entropy, and label-smoothed cross-entropy across 99 cost-sensitive metrics, each derived by assuming one of 99 Beta distributions over cost uncertainty and computing the corresponding expected misclassification cost. Fig. 2 offers a single, detailed visualisation summarising performance over 99 evaluation metrics and six datasets, highlighting when Beta(25, 25) outperforms the baselines. Each subplot corresponds to a single dataset; the X and Y axes are the Beta parameters (α, β) that define the evaluation metric. Each circle corresponds to the metric at its (α, β) location, which specifies the corresponding cost uncertainty scenario. The blue number inside each circle shows the relative % cost improvement: for the green circle, how much Beta(25,25) lowers cost vs the best baseline; for the orange, how much the best baseline lowers cost vs Beta(25,25). In most cases, Beta(25,25) performs better, especially for more cost-balanced metrics, though the gains and patterns vary by dataset.

Performance against baselines on standard metrics. We also compared the performance of the Beta(25, 25) loss against the considered baselines on commonly used evaluation metrics, including log-loss, MSE, AUC, and accuracy. We additionally chose **three** class-imbalanced datasets PneumoniaMNIST, ChestMNIST, and BreastMNIST (Yang et al. 2023) without added noise. The class imbalance ratio in these datasets is approximately 3.5:1. Prior to evaluation, we applied standard temperature scaling to all approaches, with the temperature chosen on a validation set by minimising log-loss. The mean and standard deviation of performance, aggregated over 20 random seeds across nine datasets, are presented in Table 2. We observe that, across many metrics and datasets, Beta(25, 25) often achieves noticeably better performance.

To assess statistical significance, we computed average ranks and critical differences using the Nemenyi test ($\alpha = 0.05$), as shown in Table 3. Significant superiority of Beta(25, 25) was found only relative to cross-entropy when evaluated by log-loss and accuracy; for other baselines, despite large gaps in average

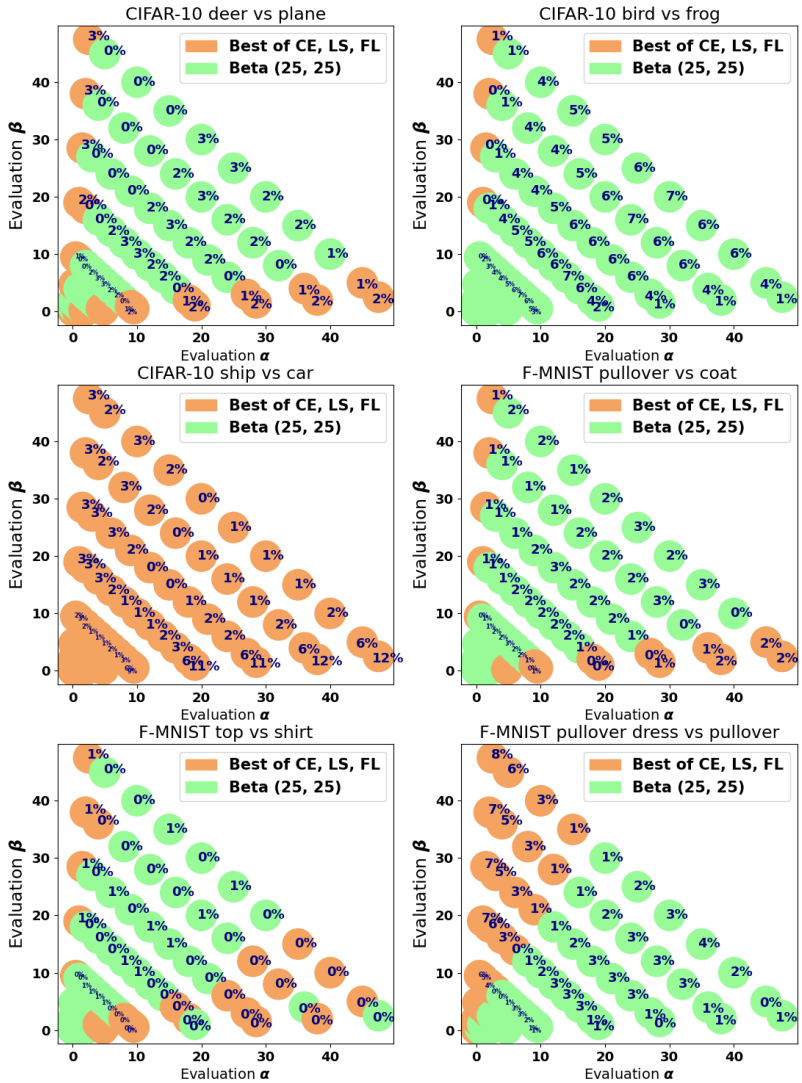

Figure 2: Comparison of Beta(25,25) with the best of cross-entropy, label smoothing, and focal across six datasets on 99 Beta cost-sensitive metrics. The plot indexes metrics by Beta parameters (α, β) on the X and Y axes; light green cells indicate Beta(25,25) outperforms the best baseline, orange indicates the converse. Blue numbers report the relative improvement (%) of the better method (positive = Beta(25,25), negative = baseline); e.g., on PneumoniaMNIST with evaluating on Beta(5,45), Beta(25,25) is 3% worse than the best baseline. Adapted from Publication I.

rank, the limited number of datasets made it difficult to detect additional significant differences.

We also investigated the effect of temperature scaling on Beta losses. Specifically, we computed the relative improvement in log-loss after applying temperature scaling (versus before) across nine datasets, alongside the other baseline losses (Fig. 3). We observe that performance improvements are largely dataset-specific, whereas the Beta and cross-entropy losses, which are also the only proper losses among those considered, typically benefit the most from post-hoc calibration.

Table 2: Mean and standard deviation of standard evaluation metrics aggregated over 20 random seeds for the Beta(25, 25) and standard loss functions with post-hoc temperature scaling calibration. Adapted from Publication I.

Losses	Evaluation metrics			
PneumoniaMNIST	CE	MSE	AUC	ACC
Cross-entropy	0.859±0.17	0.044±0.017	0.927±0.021	0.828±0.02
Label smoothing	0.813±0.14	0.017±0.016	**0.937±0.013**	0.816±0.02
Focal loss	**0.682±0.08**	**0.012±0.018**	0.928±0.008	0.808±0.03
Beta(25, 25)	0.990±0.11	0.063±0.026	0.922±0.027	**0.832±0.03**
BreastMNIST	CE	MSE	AUC	ACC
Cross-entropy	0.416±0.04	0.088±0.016	**0.851±0.04**	**0.852±0.04**
Label smoothing	0.577±0.04	0.140±0.011	0.806±0.03	0.754±0.05
Focal loss	0.468±0.03	0.115±0.012	0.841±0.04	0.815±0.04
Beta(25, 25)	**0.408±0.03**	**0.065±0.006**	0.835±0.04	0.827±0.05
ChestMNIST	CE	MSE	AUC	ACC
Cross-entropy	0.254±0.001	**0.429±0.004**	**0.562±0.03**	0.932±0.0
Label smoothing	0.236±0.000	0.445±0.000	0.506±0.01	0.936±0.0
Focal loss	0.253±0.002	0.432±0.003	0.536±0.01	0.936±0.0
Beta(25, 25)	**0.234±0.003**	0.445±0.002	0.512±0.03	**0.939±0.0**
CIFAR-10 bird vs frog	CE	MSE	AUC	ACC
Cross-entropy	0.545±0.01	0.092±0.002	0.798±0.01	0.722±0.011
Label smoothing	0.526±0.011	0.088±0.002	0.815±0.01	0.735±0.01
Focal loss	0.524±0.009	0.087±0.002	0.817±0.008	0.739±0.009
Beta(25, 25)	**0.512±0.012**	**0.085±0.002**	**0.829±0.009**	**0.747±0.01**
CIFAR-10 ship vs car	CE	MSE	AUC	ACC
Cross-entropy	0.375±0.012	0.059±0.002	0.913±0.006	0.831±0.007
Label smoothing	0.362±0.009	0.057±0.002	0.919±0.005	0.84±0.006
Focal loss	0.36±0.011	0.057±0.002	0.92±0.005	0.839±0.006
Beta(25, 25)	**0.34±0.012**	**0.053±0.002**	**0.929±0.005**	**0.852±0.008**
CIFAR-10 deer vs plane	CE	MSE	AUC	ACC
Cross-entropy	0.326±0.008	0.05±0.001	0.935±0.003	0.864±0.006
Label smoothing	0.318±0.008	0.048±0.001	0.938±0.003	0.867±0.005
Focal loss	0.314±0.006	0.048±0.001	0.939±0.003	0.869±0.005
Beta(25, 25)	**0.308±0.008**	**0.046±0.001**	**0.943±0.002**	**0.877±0.005**
Shirt vs Pullover	CE	MSE	AUC	ACC
Cross-entropy	0.324±0.005	0.05±0.001	0.936±0.002	0.864±0.002
Label smoothing	**0.316±0.006**	**0.049±0.001**	**0.939±0.002**	0.863±0.006
Focal loss	0.333±0.008	0.052±0.001	0.933±0.003	0.857±0.005
Beta(25, 25)	0.32±0.005	**0.049±0.001**	0.938±0.002	**0.868±0.004**
T-shirt/top vs Shirt	CE	MSE	AUC	ACC
Cross-entropy	0.34±0.006	0.054±0.001	0.929±0.003	0.846±0.004
Label smoothing	**0.326±0.006**	**0.051±0.001**	**0.936±0.002**	**0.852±0.004**
Focal loss	0.326±0.005	0.052±0.001	0.934±0.002	0.849±0.004
Beta(25, 25)	0.335±0.006	0.052±0.001	0.932±0.003	0.851±0.005
Pullover vs Coat	CE	MSE	AUC	ACC
Cross-entropy	0.312±0.009	0.047±0.002	0.94±0.003	0.87±0.005
Label smoothing	0.303±0.003	0.046±0.0	0.944±0.002	0.875±0.004
Focal loss	0.313±0.005	0.048±0.001	0.939±0.002	0.868±0.002
Beta(25, 25)	**0.294±0.004**	**0.044±0.001**	**0.947±0.001**	**0.882±0.003**

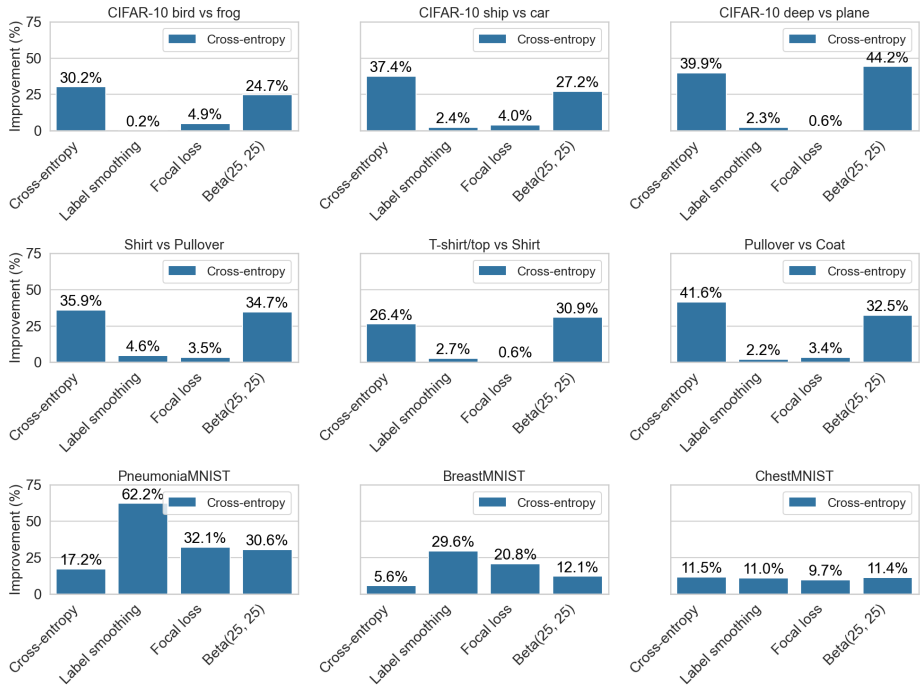

Figure 3: The relative improvement in cross-entropy, computed as relative difference before and after applying post-hoc temperature scaling, is shown for four loss functions - Beta(25, 25), label smoothing, focal loss, and cross-entropy - across nine datasets. Each of the nine subplots corresponds to one dataset: CIFAR-10 (bird vs. frog, ship vs. car, deer vs. plane classes), Fashion-MNIST (Shirt vs. Pullover, T-shirt/top vs. Shirt), Pullover vs. Coat classes, PneumoniaMNIST, ChestMNIST, and BreastMNIST. Adapted from Publication I.

Table 3: Average ranks of the CE, LS, FL, and Beta (25, 25) loss functions on standard metrics computed over nine datasets, along with critical differences. A lower rank is better and is highlighted in bold. Underlined scores denote methods that are significantly inferior (Nemenyi test with $\alpha = 0.05$) to the top-performing approach according to the critical difference. Adapted from Publication I.

	CE rank	**MSE rank**	**AUC rank**	**ACC rank**
CE	3.39	3.11	2.94	3.28
LS	2.39	2.56	2.39	2.72
FL	2.56	2.56	2.83	2.78
Beta(25, 25)	**1.67**	**1.78**	**1.83**	**1.22**
Critical difference	1.56	1.56	1.56	1.56

3.4. Summary and limitations

By assuming class cost uncertainty, our work aligns more closely with real-world applications. In practice, exact misclassification costs are rarely known during training; instead, approximate estimates, often informed by planning constraints or domain experts, are available, while precise values can only be measured at deployment. Building on these assumptions, we addressed the key question of which loss function to use in order to minimise expected total cost, a metric that directly captures the trade-offs between errors under cost uncertainty.

Starting from the expected cost formulation, we derive the Beta loss family by modelling class cost proportions with Beta distribution, which is convenient for practical applications. The Beta distribution can represent a single high-density peak (near the "most likely" cost value) or broader uncertainty, by increasing hyperparameters that control the cost variance. We show that Beta losses possess several optimisation-friendly properties: they are smooth, admit closed-form gradient expressions, and become convex for a specific range of hyperparameters.

We also show that alternative cost parameterisations are possible, such as those based directly on class cost magnitudes, and illustrate this with the example of the Gamma loss family.

We demonstrate that directly minimising Beta losses lowers the expected total cost on both the validation and test sets compared to existing cost-sensitive approaches. Also, we study whether it is optimal to align the cost parameters of the surrogate loss with the class-cost parameters of the evaluation metric.

Experiments show that training with large, nearly equal Beta loss parameters $\alpha \approx \beta >> 1$, consistently delivers lower cost-sensitive metrics than standard cross-entropy and performs comparably with focal loss and label smoothing, which are losses typically chosen for handling class imbalance. High and balanced Beta loss parameters represent a low-variance, balanced prior over misclassification costs, which surprisingly keeps total cost low across most cost distributions except extreme imbalance. Moreover, applying temperature scaling further enhances the performance of probabilistic measures of the model trained with Beta loss.

Despite its scope, this study has clear limitations: it assumes a specific form of cost uncertainty and considers only certain distributions within it (mostly Beta). In experiments, the study focuses only on image classification tasks that are class-balanced or only mildly imbalanced. Still, even within these limitations, the study highlights important aspects of selecting losses for cost-uncertainty metrics and provides a foundation for future work. In the multiclass setting, ideas from the framework of Cid-Sueiro et al. (Cid-Sueiro et al. 1999) can be reused to extend the analysis beyond the binary case under suitable structural assumptions on the cost matrix. In that context, the Dirichlet distribution, as the natural multiclass generalisation of the Beta distribution, becomes a primary candidate for modelling cost uncertainty via Dirichlet-distributed class-wise cost proportions.

4. IMPROVING CALIBRATION BY RELATING FOCAL LOSS, TEMPERATURE SCALING, AND PROPERNESS (PUBLICATION II)

4.1. Introduction

This chapter presents the thesis's second contribution, exploring calibration of separable losses (that depend only on true class predicted probability), with a primary focus on focal loss. We show that focal loss admits a decomposition into a proper loss component and a fixed calibration map component, and we examine how these parts affect calibration during training, validation, and deployment. We then extend this decomposition to broader families of separable losses and propose new training losses together with their associated proper and calibration map parts. Finally, we evaluate the calibration of focal loss and the proposed losses on image classification datasets, pairing them with newly derived families of calibration maps.

We focus on the following research questions:

- Why does the (improper) focal loss yield better calibration than the commonly used (proper) cross-entropy loss (as reported by (Mukhoti et al. 2020)), and how does focal loss relate to properness?
- Can we decompose any separable loss into a proper component and a fixed calibration map component, and can some separable losses perform comparably to cross-entropy?

4.2. Motivation

Model calibration is essential in real-world systems because reliable predicted probabilities support accurate downstream decisions. Still, many modern deep learning models suffer from miscalibration due to overconfidence, when predicted confidences are overly optimistic. While many studies propose new methods to improve calibration, it is equally important to understand the mechanisms that drive it. Post-hoc calibration, applied once optimisation is complete and the weights are fixed, fits a transformation of predicted probabilities on a validation set to better match true class probabilities; it is now a routine step in the training pipeline. Yet discovering a loss function that yields well-calibrated probabilities without post-hoc adjustments or additional regularisations would simplify deployment and deepen our understanding of training dynamics.

Focal loss often yields well-calibrated probabilities without post-hoc adjustments, while matching or even outperforming cross-entropy on discriminative metrics (e.g., accuracy). This chapter aims to explain why focal loss calibrates well and to develop and assess new loss functions with similar calibration properties.

4.3. Main findings

We derive a decomposition of the focal loss into a proper loss component and a fixed calibration map component, analyse properties of these components, propose a new family of calibration maps by combining the focal calibration component with standard temperature scaling, and investigate the practical performance of this calibration method.

Later, we extend this decomposition to a wider family within separable losses, propose new training losses and calibration maps, and show in experiments that they yield comparable performance and sometimes outperform classical baselines such as cross-entropy and focal loss on both accuracy and calibration metrics.

4.3.1. Focal loss decomposition

We consider focal loss (T.-Y. Lin et al. 2017), defined for a true class label y and a predicted probability vector p (with p_y the probability assigned to class y) by

$$L_{FL}(p,y) = -\left(1-p_y\right)^{\gamma} \log p_y,$$

with hyperparameter $\gamma \geq 0$. Throughout this subsection, we often refer to the focal loss simply as $L(\cdot,\cdot)$ for brevity (in such cases, we explicitly state in the surrounding text that L denotes the focal loss).

Binary case. We first establish a decomposition of the binary focal loss, as presented in the following proposition.

Proposition 4.3.1 (Restated from Proposition 2, Publication II)**.** *Let $L(q,y)$ be a binary focal loss parametrised by $\gamma > 0$. Then it can be decomposed into a bijective function $\hat{p} : (0,1) \to (0,1)$ (which could be seen as a fixed calibration map) and a proper loss $L^* : [0,1] \times \{0,1\} \to \mathbb{R}^+$ such that $L^*(q,y) = L(\hat{p}^{-1}(q),y)$ (equivalently, $L(q,y) = L^*(\hat{p}(q),y)$), and $\hat{p}$ is defined as:*

$$\hat{p}(q) = \frac{1}{1+\left(\frac{1-q}{q}\right)^{\gamma} \frac{(1-q)-\gamma q \cdot \log(q)}{q-\gamma(1-q)\cdot\log(1-q)}}.$$

This decomposition shows that the focal loss can be interpreted as a standard proper loss applied after a fixed, nonlinear calibration of predicted probabilities. In other words, focal loss reshapes the probability space through a bijective transformation $\hat{p}$ before applying a proper scoring rule L^*. This view connects focal loss to the general theory of composite and link-based losses introduced by Reid and Williamson (M. D. Reid and R. C. Williamson 2010) where properness is preserved under suitable link functions, and to the posterior re-mapping interpretation of Saerens et al. (2002), who expressed similar transformations for probabilistic classifiers (their Eq. (13)). Conceptually, in the binary case this decomposition clarifies that the focal loss does not discard the properness principle but rather modifies it via a calibration layer, much like a nonlinear activation preceding a proper objective in a neural network.

The main proof steps were the following.

1. Defining the transformation $\hat{p}(\cdot)$ based on ideas from composite link theory (M. D. Reid and R. C. Williamson 2010).
2. Proving that the mapping $\hat{p}(\cdot)$ is bijective.
3. Showing that the residual loss L^* is proper.

We henceforth refer to the fixed calibration map induced by this decomposition as the *focal calibration map* (parameterised by γ). We analyse this binary calibration map in logit space (combined with the sigmoid link) and observe a close resemblance to standard temperature scaling (Guo et al. 2017). Moreover, we show that, for all logits, the focal calibration map admits pointwise lower and upper bounds given by temperature scaling transformations with close temperature parameters (see the next proposition).

Proposition 4.3.2 (Restated from Proposition 3, Publication II)**.** *Let* $FC(s) = \hat{p}_\gamma(\frac{1}{1+e^{-s}})$ *be a focal calibration function applied on top of a sigmoid with a logit* s*. Then, the focal calibration could be bounded between two temperature scaling maps with* $T = \frac{1}{\gamma+1}$ *and* $T = \frac{1}{\gamma+1-\frac{\log(\gamma+1)}{2}}$ *such that*

$$\forall s < 0 \quad \frac{1}{1+e^{-\frac{s}{\gamma+1}}} > FC(s) > \frac{1}{1+e^{-\frac{s}{\gamma+1-\frac{\log(\gamma+1)}{2}}}},$$
$$\forall s \geq 0: \quad \frac{1}{1+e^{-\frac{s}{\gamma+1}}} < FC(s) < \frac{1}{1+e^{-\frac{s}{\gamma+1-\frac{\log(\gamma+1)}{2}}}}.$$

We illustrate the similarity between the focal calibration map and temperature scaling in logit space by (i) plotting, for each calibration parameter γ, the corresponding lower and upper temperature scaling bounds that enclose the focal map over the entire logit range (two sets of bounds: one proven theoretically and one obtained via numerical approximation); and (ii) plotting, alongside, the maximum gap between these bounds across logits for each γ (Fig. 4). The theoretical bounds are intentionally conservative to simplify the proof; as a result, they can be wider than necessary, while the experimental bounds are typically much tighter. The observed relation between the optimal temperature T and the focal parameter γ is approximately inverse: $\gamma \approx \frac{1}{T}$ as could be seen in Fig. 5.

Motivated by the striking empirical similarity between focal calibration and temperature scaling on the logit scale, we run numerical experiments to investigate the relationship between their parameters. We approximate the binary focal calibration curves with temperature scaling on the logit scale by choosing $T(\gamma)$ to minimise the maximum absolute deviation between the two transformations over a fine logit grid (e.g., [-20, 20] with step 0.05). This minimax fit makes their similarity especially clear, as shown in Fig. 5 (right) for $\gamma \in \{0.5, 2\}$. The left panel shows that this approximation yields an approximately linear relationship between γ and the optimal inverse temperature $\frac{1}{T(\gamma)}$.

Multiclass case. We derive focal loss decomposition into proper loss component and calibration map component in the multiclass case as presented in the

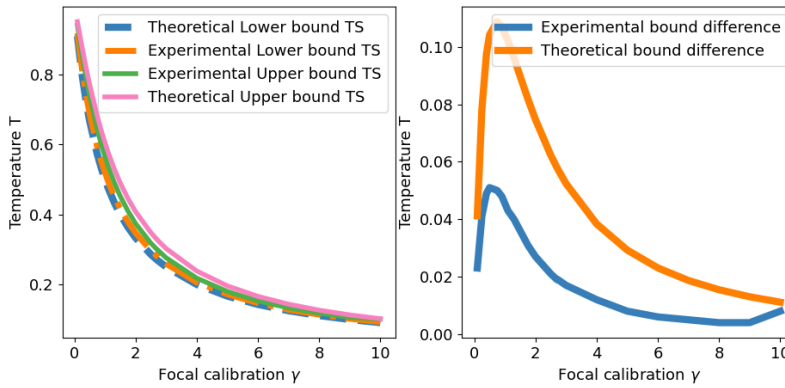

Figure 4: Left: theoretical and experimental bounds for focal calibration with temperature scaling maps. Right: theoretical and experimental widths of the bounds for different γ. The curves in the left panel overlap closely, which highlights how tight the theoretical and experimental bounds are rather than separating them for visibility. The right panel shows that the theoretical bounds, while tight, are still noticeably wider than the experimental ones, which indicates room for deriving even tighter theoretical bounds, potentially at the cost of much more complicated derivations. Adapted from Publication II.

following proposition.

Proposition 4.3.3 (Restated from Proposition 4, Publication II)**.** *Let $L(q,y)$ be a multiclass focal loss parametrised with some $\gamma > 0$. Then, it can be deconstructed into a composition of a bijective function $\hat{p}(q)$ and a proper loss $L^*(q,y)$ such that:*

$$\hat{p}_j(q_1,...,q_{\mathrm{K}}) = \frac{\frac{1}{(1-q_j)^{\gamma}\cdot(\frac{\gamma\cdot\log(q_j)}{1-q_j}-\frac{1}{q_j})}}{\sum_{k=1}^{\mathrm{K}}\frac{1}{(1-q_k)^{\gamma}\cdot(\frac{\gamma\cdot\log(q_k)}{1-q_k}-\frac{1}{q_k})}} \quad \forall j = 1..\mathrm{K},$$

$$L^*(q,y) = L(\hat{p_1}^{-1}(q_1,...,q_{\mathrm{K}}),...,\hat{p_{\mathrm{K}}}^{-1}(q_1,...,q_{\mathrm{K}})).$$

This proposition shows that the multiclass focal loss, as in the binary case, can be understood as a standard proper loss applied after a nonlinear, invertible reparameterisation of class probabilities. A similar multiclass expression for focal transformations appears in (Charoenphakdee et al. 2021), where the authors derived a comparable reweighting of class probabilities but did not explicitly formalise the decomposition into a proper loss and a bijective calibration map. The main proof steps are essentially the same as for the binary case and are as follows.

1. Define the transformation $\hat{p}(\cdot)$ for the multiclass focal loss by extending the binary calibration map component-wise to the probability simplex, normalising it to ensure that both the input and output lie on the simplex.

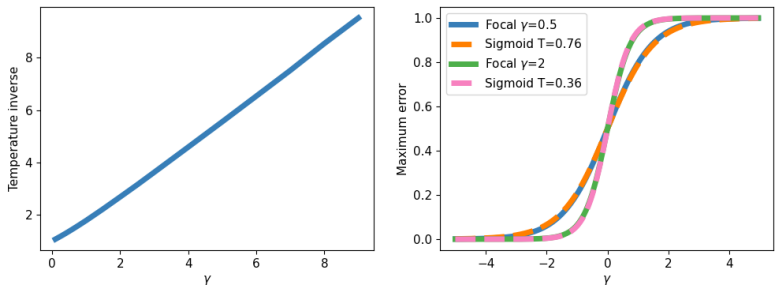

Figure 5: Left: relationship between focal calibration with parameter γ and temperature scaling (inverse temperature $\frac{1}{T}$), where $\frac{1}{T}$ is fitted to minimise the maximum absolute deviation between the two transformations over logits in range [-20, 20] (step size 0.05). Right: focal calibration curves with parameters 0.5 and 3 on the logit scale, together with their closest temperature scaling counterparts (with $T = 0.76$ and $T = 0.36$ respectively). The right-panel examples were deliberately selected to illustrate near-indistinguishability: the focal and temperature scaling mappings overlap so closely that they cannot be visually separated at this scale. Adapted from Publication II.

2. Prove bijectivity of $\hat{p}$ on the probability simplex, by analysing the component-wise derivatives and the overall Jacobian.
3. Show that the residual loss L^* is proper.

We next examine the derived focal calibration map on the probability simplex, visualising the three-class case alongside temperature scaling (Fig. 7). The focal calibration map shows the opposite of the usual overconfidence-reducing calibration behaviour: probability vectors are pushed from the simplex interior toward the vertices, similar to temperature scaling with $T < 1$, as indicated by outward-pointing arrows.

Similarly, we visualise the proper component of focal loss in the three-class case by plotting conditional risk isolines for a fixed ground truth probability vector, alongside Brier score and cross-entropy (Fig. 6). The resulting loss landscape is more complex than for the other losses and does not yield simple qualitative conclusions.

We combine the multiclass focal calibration map with temperature scaling into a new calibration method, focal temperature scaling, where the parameter pair (γ, T) is chosen on a validation set by minimising log-loss or ECE. We benchmark this method against standard temperature scaling with models trained using cross-entropy, focal loss, and training-time calibration methods (FLSD-53 (Mukhoti et al. 2020) and AdaFocal (Ghosh, Schaaf, and Gormley 2022)); evaluation uses accuracy, log-loss, and ECE, with results shown in Table 6. The results show that focal temperature scaling consistently reduces ECE further than standard temperature scaling, keeps accuracy unchanged, and achieves comparable log-loss. Moreover, it can be applied on top of training-time calibration (e.g., AdaFocal) to

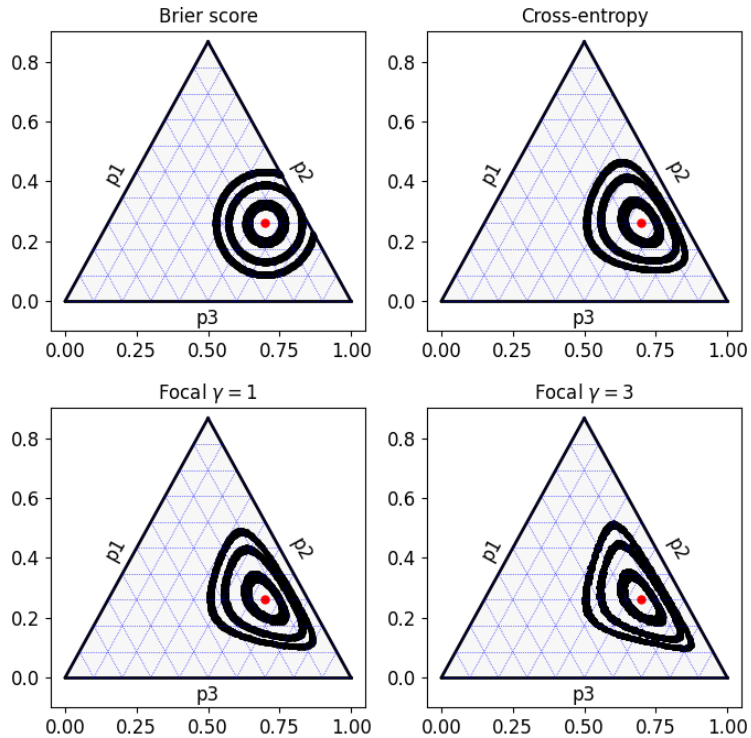

Figure 6: Brier score, cross-entropy and properized focal loss ($\gamma = 1, 3$) conditional risk isolines (defined by loss percentiles $3\%, 12\%, 20\%$) for ground truth probability $p = (0.55, 0.3, 0.15)$. Adapted from Publication II.

improve calibration even further. It is worth noting that, on some datasets with cross-entropy training, we obtain $\gamma_{\text{ev}} < 0$, which has a confidence-sharpening effect. This does not contradict the typical overconfidence of cross-entropy–trained models. In our setup, focal calibration is composed with temperature scaling, and the optimal hyperparameters often pair a mildly sharpening focal map ($\gamma_{\text{ev}} < 0$) with a *softening* temperature (typically $T > 1$), so the combined effect reduces overconfidence.

4.3.2. Extension to separable losses

Next, we generalise the focal loss decomposition to a richer loss class within separable losses (i.e., losses that depend only on the true class's predicted probability), as shown in the theorem below.

Theorem 4.3.4 (Decomposition of separable losses into calibration map and proper part)**.** *Let* $\mathrm{K} \geq 2$ *and* $\Delta^{\mathrm{K}-1} := \{p \in [0,1]^{\mathrm{K}} : \sum_{i=1}^{\mathrm{K}} p_i = 1\}$. *Let* $f : [0,1] \to \mathbb{R}$ *be convex and* C^1 *on* $(0,1]$ *with*

$$f'(x) < 0 \text{ for all } x \in (0,1], \qquad \lim_{x \downarrow 0} f'(x) = -\infty, \qquad f'(1) \in (-\infty, 0).$$

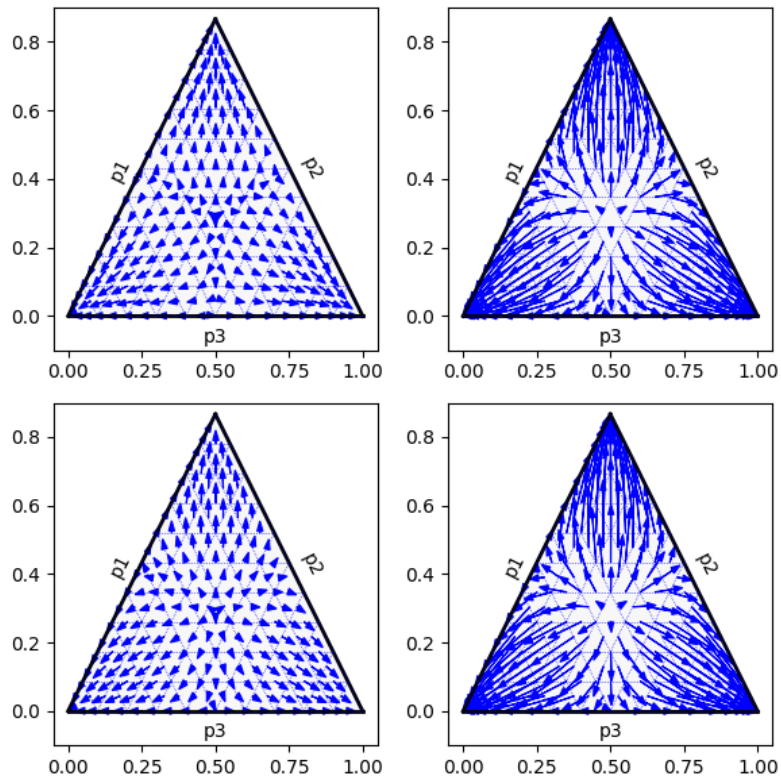

Figure 7: Focal calibration (top row) and temperature scaling (bottom row) are visualised as directional arrows over a uniform grid of three-class probability simplex points. The top-left panel uses $\gamma = 1$ and the top-right panel uses $\gamma = 3$; the bottom-left panel shows the closest temperature map (by mean squared total distance over the grid points) with $T = 0.81$, and the bottom-right panel uses $T = 0.46$. Each arrow starts at the original probability and ends at the transformed probability. Adapted from Publication II.

Consider the separable loss $L : \Delta^{\mathrm{K}-1} \times \{1, \ldots, \mathrm{K}\} \to \mathbb{R}$ *given by*

$$L(p, y) \;=\; f(p_y).$$

Define $k : (0,1] \to (0,\infty)$ *by* $k(x) := -1/f'(x)$ *and set* $k(0) := \lim_{x \downarrow 0} k(x) = 0$. *Define the calibration map* $\phi : \Delta^{K-1} \to \Delta^{K-1}$ *component-wise by*

$$\phi_i(p) \;=\; \frac{k(p_i)}{\sum_{j=1}^{\mathrm{K}} k(p_j)}, \qquad i = 1, \ldots, \mathrm{K}.$$

Then:

(a) ϕ *is a bijection. In particular,* $\phi^{-1} : \Delta^{K-1} \to \Delta^{K-1}$ *is well-defined and continuous.*

(b) *The composite loss*

$$\widetilde{L}(q,y) \;:=\; L\big(\phi^{-1}(q),y\big) \;=\; f\big((\phi^{-1}(q))_y\big), \qquad q \in \Delta^{K-1},$$

is a ***proper*** *loss. Moreover, if f is strictly convex on $(0,1)$ (equivalently k is strictly increasing), then $\widetilde{L}$ is* strictly *proper.*

(c) *(Decomposition) For every $p \in \Delta^{\mathrm{K}-1}$ and $y \in \{1,\ldots,\mathrm{K}\}$,*

$$L(p,y) \;=\; \widetilde{L}\big(\phi(p),y\big).$$

This theorem shows that any separable loss $L(p,y) = f(p_y)$ under mild assumptions can be rewritten as a proper loss applied after a fixed, bijective calibration map ϕ. In parallel with our Publication II, closely related conclusions were obtained independently by Bao and Charoenphakdee (2025), reinforcing the generality of this decomposition viewpoint. For example, for cross-entropy, the calibration map from the decomposition is the identity map, so the calibration layer vanishes and the composite reduces to the original loss; this aligns with the classical result that, up to positive affine equivalence, cross-entropy is the unique proper separable loss (Gneiting and Raftery 2007). Motivated by this, we next ask the reverse question: given a calibration map, under what conditions does it arise from some proper loss via a separable-loss decomposition, and how can this be verified?

The following lemma addresses this question. This result extends our previous work (Publication II) and will be detailed in a forthcoming study.

Lemma 4.3.5 (Representation of calibration maps and construction of a proper separable loss)**.** *Let $\Delta_\circ^{K-1} = \{p \in (0,1)^K : \sum_i p_i = 1\}$ be the open simplex and let $\phi : \Delta^{K-1} \to \Delta^{K-1}$ be a calibration map satisfying:*

(A1) ***Continuity & positivity:*** *ϕ is continuous on Δ^{K-1} and $\phi_i(p) > 0$.*

(A2) ***Permutation equivariance:*** *$\phi(\pi p) = \pi\,\phi(p)$ for every permutation π.*

(A3) ***Odds dependence:*** *for all $i \neq j$ there exists $\rho : (0,1]^2 \to (0,\infty)$ such that $\phi_i(p)/\phi_j(p) = \rho(p_i,p_j)$ for all $p \in \Delta_\circ^{K-1}$.*

(A4) ***Cocycle consistency:*** *$\rho(x,y)\rho(y,z) = \rho(x,z)$ and $\rho(x,y) = 1/\rho(y,x)$ for all $x,y,z \in (0,1]$.*

Then there exists a continuous function $k : (0,1] \to (0,\infty)$ (unique up to a positive multiplicative constant) such that

$$\boxed{\phi_i(p) \;=\; \frac{k(p_i)}{\sum_{j=1}^{K} k(p_j)}, \qquad i = 1,\ldots,K,\; p \in \Delta_\circ^{K-1}.}$$

If, in addition, ϕ extends continuously to the boundary with $\phi(e_i) = e_i$ and $\phi_i(p) = 0 \;\Leftrightarrow\; p_i = 0$, then k extends by $k(0) := 0$ and $k(1) > 0$. If ϕ is injective (rank-preserving), then k is strictly increasing.

Furthermore, if k is nondecreasing, define a separable loss $f : [0,1] \to \mathbb{R}$ by

$$f'(x) \;=\; -\frac{1}{k(x)} \quad (x \in (0,1]), \qquad f(1) := 0, \qquad f(x) := f(1) - \int_x^1 \frac{dt}{k(t)}.$$

Then f is convex with $f'(x) < 0$ on $(0,1]$ and $\lim_{x \downarrow 0} f'(x) = -\infty$. Let $L(u,y) := f(u_y)$ for $u \in \Delta^{K-1}$ and $y \in \{1,\dots,K\}$. The composite loss

$$\widetilde{L}(p,y) \;:=\; L\big(\phi^{-1}(p), y\big)$$

is strictly proper *on Δ^{K-1}.*

For example, temperature scaling satisfies A1–A4. Taking $k(x) = x^{1/T}$ with $T > 0$, we recover

$$\phi_i(p) \;=\; \frac{k(p_i)}{\sum_{j=1}^K k(p_j)} \;=\; \frac{p_i^{1/T}}{\sum_{j=1}^K p_j^{1/T}} \qquad (i = 1,\dots,K),$$

i.e., the probability-level temperature transformation.

Guided by the decomposition assumptions, we can design new focal-style objectives, allowing systematic exploration of variants that retain the key properties of focal loss and isolating whether performance is driven by the focal shape itself or by the conditions required for the decomposition. We have proposed new training losses together with corresponding calibration maps that arise from the loss decomposition into a proper term and a calibration map. Each loss is a multiplicative reweighting of cross-entropy,

$$L(p_y) = g(p_y)\,(-\log p_y),$$

with the following choices of $g(\cdot)$:

Linear-Decay Loss:	$g(p) = 1 - \beta p.$
Exponential-on-p Loss:	$g(p) = e^{-\alpha p}.$
Exponential-on-$(1-\mathbf{p})$ Loss:	$g(p) = e^{-\alpha(1-p)}.$
One-Minus-Power Loss:	$g(p) = 1 - p^{\beta}.$
Generalized Focal Loss:	$g(p) = (1-p^{\beta})^{\gamma}.$
Log-Power Loss:	$g(p) = (-\log p)^{\kappa}$ (so $L(p) = (-\log p)^{\kappa+1}$).

Here $\alpha, \beta, \gamma, \kappa \geq 0$.

We next evaluate the performance of the proposed loss functions, compared with cross-entropy and focal loss, and combined with temperature scaling, as shown in Table 4. Overall, the new losses achieve performance broadly comparable across all evaluation metrics.

Similarly, we evaluate the proposed losses paired with proposed calibration maps (not necessarily the corresponding ones), as shown in Table 5. Exploring these combinations, we observe consistent improvements in ECE over focal loss and cross-entropy. Moreover, in several settings, the proposed calibration maps, used on their own (i.e., without combining with temperature scaling), already improve calibration even relative to baselines that are paired with temperature scaling.

Loss	Param	Accuracy	Log-Loss	Log-Loss$_{T=1}$	ECE	ECE$_{T=1}$
CE	-	77.9±0.0	1.01±0.00 (2.00)	1.54±0.04	3.77±0.57 (2.20)	17.27±0.28
Focal	2.0	78.4±0.3	0.87±0.01 (1.40)	0.92±0.01	2.67±0.33 (1.45)	9.83±0.28
Linear	0.5	77.6±0.3	1.01±0.01 (2.25)	1.50±0.07	3.91±0.82 (2.35)	17.38±0.19
Expp	3.0	78.0±0.4	0.94±0.00 (1.70)	1.15±0.03	2.89±0.30 (1.85)	14.60±0.72
Exp1mp	0.5	77.6±0.3	1.03±0.01 (1.85)	1.59±0.03	4.31±0.40 (1.95)	17.57±0.23
MinusPow	0.5	77.9±0.1	0.96±0.01 (1.60)	1.51±0.05	3.11±0.41 (1.70)	16.97±0.47
LogPow	1.0	77.3±0.6	0.95±0.02 (2.00)	1.14±0.02	2.54±0.05 (2.15)	13.96±0.10

Table 4: Test set performance (mean ± standard deviation over 5 random seeds) on CIFAR-100 for the proposed losses versus cross-entropy and focal loss; within each loss family, the hyperparameter was selected by validation accuracy. Log-Loss$_{T=1}$ and ECE$_{T=1}$ denote performance before applying temperature scaling calibration. For log-loss and ECE, the optimal temperature is reported in brackets.

Loss	Param	Link	Value	Accuracy	Log-Loss	Log-Loss$_{T=1}$	ECE	ECE$_{T=1}$
LogPow	2.00	Exp1mp	1.00	75.8±0.3	0.90±0.01 (0.95)	0.90±0.01	**1.30**±0.22 (1.00)	**1.21**±0.19
Expp	7.00	Exp1mp	0.75	77.5±0.4	**0.83**±0.01 (1.00)	**0.83**±0.01	1.68±0.32 (1.05)	2.97±0.19
Focal	7.00	Focal	7.00	77.2±0.1	**0.83**±0.00 (1.00)	0.84±0.00	1.80±0.37 (1.05)	2.90±0.27
MinusPow	0.25	Exp1mp	1.50	77.8±0.1	0.85±0.00 (1.10)	0.86±0.00	2.45±0.13 (1.10)	5.28±0.19
Linear	1.00	MinusPow	7.00	77.4±0.3	0.94±0.02 (1.45)	1.20±0.02	2.53±0.35 (1.55)	13.36±0.42
Exp1mp	0.25	MinusPow	7.00	77.5±0.1	1.03±0.01 (2.10)	2.06±0.08	2.88±0.07 (2.20)	17.43±0.33
CE	-	MinusPow	7.00	**77.9**±0.0	1.02±0.01 (2.15)	2.06±0.07	3.27±0.37 (2.20)	17.31±0.27

Table 5: Test set performance (mean ± standard deviation over 5 random seeds) for the proposed losses combined with the new calibration map, compared against cross-entropy and focal loss; calibration hyperparameters were selected by validation ECE. Log-Loss$_{T=1}$ and ECE$_{T=1}$ denote performance before applying temperature scaling. For log-loss and ECE, the optimal temperature used in temperature scaling (when combined with the new calibration map family) is shown in brackets.

Table 6: Test set performance for focal loss (defined by γ_{tr}), sample-dependent focal loss FLSD-53 Mukhoti et al. 2020, AdaFocal with default parameters as in Ghosh, Schaaf, and Gormley 2022 and cross-entropy trained models with temperature scaling versus focal temperature scaling (defined by γ_{ev} and temperature). The CIFAR-100, CIFAR-10 and TinyImageNet datasets were used, and the results were averaged over 5 random seeds after applying temperature scaling. The mean result is reported together with the standard deviation after the $\pm$ sign. The optimal temperature is reported in brackets; temperature choice criteria were log-loss for log-loss evaluation and ECE for ECE evaluation. The best result for each metric and dataset is highlighted in bold formatting. Adapted from Publication II.

CIFAR-100 DATASET			
APPROACH	ACCURACY	LOG-LOSS	ECE
CROSS-ENTROPY	77.6 ± 0.6	0.88 ± 0.02 (1.31)	3.01 ± 0.43 (1.45)
$+\gamma_{ev} = -0.5$	77.6 ± 0.6	0.86 ± 0.02 (1.17)	2.13 ± 0.42 (1.35)
FOCAL $\gamma_{tr} = 1$	**77.7** ± 0.3	0.83 ± 0.01 (1.05)	1.66 ± 0.23 (1.15)
$+\gamma_{ev} = -0.25$	**77.7** ± 0.3	0.82 ± 0.01 (1.00)	1.34 ± 0.17 (1.10)
FOCAL $\gamma_{tr} = 3$	77.3 ± 0.5	**0.81** ± 0.02 (0.87)	1.28 ± 0.15 (0.91)
$+\gamma_{ev} = 0.05$	77.3 ± 0.5	0.82 ± 0.02 (0.87)	1.23 ± 0.14 (0.95)
FOCAL $\gamma_{tr} = 7$	76.3 ± 0.5	0.83 ± 0.01 (0.70)	1.83 ± 0.20 (0.65)
$+\gamma_{ev} = 0.5$	76.3 ± 0.5	0.83 ± 0.01 (0.75)	**0.99** ± 0.07 (0.75)
FLSD-53	77.5 ± 0.5	0.88 ± 0.01 (1.20)	1.89 ± 0.18 (1.27)
$+\gamma_{ev} = 0.25$	77.5 ± 0.3	0.88 ± 0.02 (1.05)	1.68 ± 0.19 (1.15)
ADAFOCAL	77.6 ± 0.2	0.91 ± 0.03 (1.40)	2.96 ± 0.22 (1.52)
$+\gamma_{ev} = 0.25$	77.6 ± 0.2	0.93 ± 0.03 (1.48)	2.71 ± 0.17 (1.60)

CIFAR-10 DATASET			
APPROACH	ACCURACY	LOG-LOSS	ECE
CROSS-ENTROPY	**95.0** ± 0.1	**0.16** ± 0.00 (1.59)	1.03 ± 0.17 (1.72)
$+\gamma_{ev} = 1$	**95.0** ± 0.1	0.17 ± 0.01 (2.20)	0.71 ± 0.16 (2.36)
FOCAL $\gamma_{tr} = 1$	**95.0** ± 0.1	0.17 ± 0.01 (1.05)	1.05 ± 0.25 (1.13)
$+\gamma_{ev} = 0.5$	**95.0** ± 0.1	0.17 ± 0.01 (1.30)	0.82 ± 0.35 (1.37)
FOCAL $\gamma_{tr} = 3$	94.3 ± 0.3	0.19 ± 0.01 (0.75)	1.48 ± 0.25 (0.77)
$+\gamma_{ev} = 5$	94.3 ± 0.3	0.20 ± 0.01 (1.83)	0.93 ± 0.16 (1.87)
FOCAL $\gamma_{tr} = 7$	93.1 ± 0.1	0.23 ± 0.01 (0.49)	0.66 ± 0.07 (0.44)
$+\gamma_{ev} = 0.37$	93.1 ± 0.1	0.22 ± 0.01 (0.55)	**0.61** ± 0.10 (0.52)
FLSD-53	94.6 ± 0.1	0.18 ± 0.01 (1.40)	1.30 ± 0.13 (1.40)
$+\gamma_{ev} = 1$	94.6 ± 0.1	0.17 ± 0.01 (1.30)	1.23 ± 0.20 (1.33)
ADAFOCAL	94.9 ± 0.2	0.18 ± 0.01 (1.58)	1.80 ± 0.17 (1.63)
$+\gamma_{ev} = 5$	94.9 ± 0.2	0.20 ± 0.02 (3.70)	0.95 ± 0.10 (3.70)

TINYIMAGENET DATASET			
APPROACH	ACCURACY	LOG-LOSS	ECE
CROSS-ENTROPY	49.9 ± 0.1	2.21 ± 0.00 (1.35)	5.57 ± 0.31 (1.40)
$+\gamma_{ev} = -0.5$	49.9 ± 0.1	2.19 ± 0.00 (1.30)	3.66 ± 0.18 (1.40)
FOCAL $\gamma_{tr} = 1$	50.6 ± 0.2	2.11 ± 0.01 (1.10)	3.27 ± 0.13 (1.20)
$+\gamma_{ev} = -0.5$	50.6 ± 0.2	2.10 ± 0.01 (1.10)	1.86 ± 0.12 (1.18)
FOCAL $\gamma_{tr} = 3$	51.6 ± 0.1	2.04 ± 0.01 (0.95)	2.21 ± 0.14 (0.98)
$+\gamma_{ev} = -0.25$	51.6 ± 0.1	2.03 ± 0.01 (0.95)	1.63 ± 0.05 (0.97)
FOCAL $\gamma_{tr} = 7$	50.9 ± 0.3	**2.01** ± 0.02 (0.85)	1.01 ± 0.02 (0.85)
$+\gamma_{ev} = 0.05$	50.9 ± 0.3	**2.01** ± 0.02 (0.85)	**0.96** ± 0.00 (0.85)
FLSD-53	**52.1** ± 0.1	2.02 ± 0.01 (0.95)	2.06 ± 0.17 (0.98)
$+\gamma_{ev} = -0.25$	**52.1** ± 0.1	2.02 ± 0.01 (0.95)	1.48 ± 0.22 (0.95)
ADAFOCAL	51.6 ± 0.3	2.07 ± 0.03 (1.05)	2.97 ± 0.67 (1.10)
$+\gamma_{ev} = -0.5$	51.6 ± 0.3	2.07 ± 0.03 (1.05)	1.89 ± 0.24 (1.09)

4.4. Conclusion

We introduced a decomposition of focal loss into (i) a proper loss component and (ii) a fixed calibration map component, and extended this result to a broader class of separable losses under mild assumptions, noting that cross-entropy is the special case with the identity calibration map. We hypothesised that the proper loss component drives discriminative performance (e.g., accuracy), particularly on the training set, whereas the calibration map component mitigates miscalibration, especially on the unseen data where generalisation gaps are most pronounced. We also established the opposite link by characterising, for a given calibration map of a specific form, the corresponding separable loss that induces it. The decomposition is valuable because it bridges training losses and calibration, helping to explain the behaviour of existing objectives and to systematically design new losses and calibration methods. It also highlights the practical role of proper losses and guides the adaptation of post-hoc calibration to deployment needs (e.g., *OOD* robustness, noise resilience, context shift), enabling clearer diagnostics (which component fails) and principled transfer across architectures and datasets.

We showed, using theoretical lower and upper bounding and numerical experiments, that in the binary case, the focal calibration map is closely related to temperature scaling but never exactly coincides with it. In the multiclass case, although similarities remain, their behaviour differs noticeably near the simplex boundaries. We also examined the proper component of focal loss, visualising its shape in two- and three-dimensional space. The plots showed a broad similarity to cross-entropy while retaining a distinct geometry.

We ran experiments on several image classification datasets, including standard benchmarks and class-imbalanced medical datasets, to test whether the newly derived calibration maps, alone and in combination with temperature scaling, improve calibration relative to common baselines (such as standard temperature scaling (Guo et al. 2017), focal loss with a sample-dependent, confidence-based γ schedule (Mukhoti et al. 2020), and adaptive focal loss that adjusts γ each epoch based on validation set calibration performance (Ghosh, Schaaf, and Gormley 2022)). Although the scope is limited to image classification and a modest number of datasets, the results indicate that the proposed map families can improve calibration, especially when paired with temperature scaling. In many cases, the optimal temperature was close to 1, suggesting that temperature scaling itself was not the primary driver of the obtained gains. Because the derived calibration maps were applied post hoc and were shown to be monotone, discriminative metrics such as accuracy remained unchanged.

Moreover, we evaluated the accuracy of the newly proposed separable losses that admit a decomposition into proper and calibration components. Their performance was generally comparable to the commonly used cross-entropy and occasionally surpassed it, suggesting promise for exploring new training losses and investigating mechanisms that drive their discriminative and calibration behaviour.

5. ALIGNING THE EVALUATION OF PROBABILISTIC PREDICTIONS WITH DOWNSTREAM VALUE (PUBLICATION III)

5.1. Introduction

This chapter presents the third contribution of the thesis: aligning standard upstream evaluation metrics with domain-specific downstream objectives that ultimately drive decision-making. Unlike the previous two chapters, which focus on classification with discrete label spaces, here we study regression, with the goal of predicting targets in a continuous space $\mathscr{Y} \subseteq \mathbb{R}^d$. This change introduces additional complexity, as we now must reason about probability measures on $\mathbb{R}^d$, (e.g. via densities when they exist), not just discrete class probabilities. We propose to learn, on a validation set, a data-driven proxy mapping that predicts downstream utility directly from the upstream metric. By constraining the mapping family to positive affine transformations of the upstream metric combined with bijective reparameterisations of the predictive distribution and ground truth score, we show that the properness of the upstream measure can be preserved. Finally, we outline when such alignment is practically beneficial and demonstrate proof-of-concept results on several examples, including an inventory optimisation task.

We explore the following research questions.

- Can upstream and downstream regression evaluation be aligned via a proxy function learned on validation data?
- Under what assumptions on the transformations do both the upstream and downstream metrics remain proper?

5.2. Motivation

During design, training, and even post-deployment, the downstream evaluation metric that ultimately drives decisions and captures domain-specific nuances is often unavailable, subject to change, or fundamentally uncertain owing to stochastic external factors, insufficient data, or evolving goals.

Conversely, standard upstream evaluation metrics such as the Brier score provide a useful high-level view of model performance, but are generally insufficient when preparing high-stakes real-world applications for deployment. In many real-world applications, however, computing the downstream metric can be costly due to physical modelling, computationally intensive simulation, or other constraints. In such settings, a fast and accurate approximation of the downstream metric from standard upstream measures can be valuable. Because downstream utilities are often complex, analytic approximations are typically infeasible; instead, a data-driven approach, involving learning a mapping on validation data (unseen during

training) and constraining the alignment family to specific function classes, is more practical.

This chapter investigates how to design such an alignment, examines when properness can be preserved through the learned mapping, and presents proof-of-concept experiments demonstrating its practical performance.

5.3. Main findings

Let $L^u(\mathscr{Q}, y)$ denote the upstream scoring rule, i.e., a loss function that evaluates a predictive distribution $\mathscr{Q}$ with respect to a realised outcome $y \in \mathscr{Y} \subseteq \mathbb{R}$ and produces a scalar score (for example, a Brier score or CRPS value). Let $L^d(\mathscr{Q}, y)$ denote the corresponding downstream score or utility, which is also a real-valued function of the predictive distribution and outcome but is intended to reflect the application-specific performance measure we ultimately care about, e.g., monetary profit, decision cost, or a regulatory performance metric. We seek to express (or approximate) the downstream metric in terms of the upstream one via simple transformations of its inputs and output. To this end, we assume there exist transformations $h : \mathbb{R} \to \mathbb{R}$ and an invertible $\nu : \mathscr{Y} \to \mathscr{Y}$ such that

$$L^d(\mathscr{Q}, y) \;=\; h\Big(L^u\big(\mathscr{Q} \circ \nu^{-1},\, \nu(y)\big)\Big).$$

Here, ν reparameterises the label space, e.g., rescaling units or applying a monotone transformation, and $\mathscr{Q} \circ \nu^{-1}$ denotes the pushforward of $\mathscr{Q}$ under ν, i.e., the predictive distribution of the transformed label $\nu(Y)$ when $Y \sim \mathscr{Q}$. The outer map h then converts the upstream score computed on this transformed problem into the scale of the downstream metric, for example, turning a loss into a utility or rescaling to match a business-specific cost.

The following proposition states conditions when the properness of the upstream rule will imply the properness of the downstream utility when linked via transformations h, ν.

Proposition 5.3.1 (Restated from Proposition 1 Publication III)**.** *Let $L^u(\mathscr{Q}, y)$ be a (strictly) proper scoring rule on an outcome space $\mathscr{Y} \subseteq \mathbb{R}$. Define*

$$L^d(\mathscr{Q}, y) \;=\; h\big(L^u\big(\mathscr{Q} \circ \nu^{-1}, \nu(y)\big)\big),$$

where $\nu : \mathscr{Y} \to \mathscr{Y}$ and $h : \mathbb{R} \to \mathbb{R}$ are continuous functions. Then L^d is (strictly) proper if and only if:

1. *ν is a bijection on $\mathscr{Y}$, and*
2. *$h(s) = a\,s + b$ is an affine function with $a > 0$.*

No other choice of (h, ν) preserves properness.

In general, not all scoring rules can be perfectly aligned via the transformations h, ν. Consider a fixed predictive distribution $\mathscr{Q} = \mathrm{Unif}(0,1)$. The logarithmic score is constant w.r.t this predictive distribution because the density is $p(y) = 1$

on $(0,1)$, so $L_{\log}(\mathscr{Q},y) = -\log p(y) = 0$ for all $y \in (0,1)$. By contrast, the Continuous Ranked Probability Score (CRPS, (Gneiting and Raftery 2007)), defined as $\mathrm{CRPS}(\mathscr{Q},y) = \int_{\mathbb{R}} (P(z) - I\{y \le z\})^2 \, dz$, for the same predictive distribution $\mathscr{Q}$ varies with y:

$$\begin{aligned} L_{\mathrm{CRPS}}(\mathscr{Q},y) &= \int_0^y z^2 \, dz + \int_y^1 (1-z)^2 \, dz = \frac{y^3 + (1-y)^3}{3} \\ &= y^2 - y + \tfrac{1}{3}, \qquad 0 \le y \le 1. \end{aligned}$$

Since $L_{\log}(\mathscr{Q},\cdot) \equiv 0$ is constant, any pointwise transformation of it is also constant; consequently, the non-constant values of CRPS cannot be recovered, and perfect alignment between these scoring rules is impossible for this $\mathscr{Q}$.

However, for certain cases, the perfect alignment is attainable. The following proposition identifies some families of scoring rules L^u, L^d that can be expressed as transformations of one another under h, ν.

Proposition 5.3.2 (Restated from Lemma 2 Publication III)**.** *Let*

$$L^u(\mathscr{Q},y) = \int_{-\infty}^{\infty} w^u(x) \, k\big(\mathscr{Q}(x), \mathbb{I}\{y \le x\}\big) \, dx,$$

$$L^d(\mathscr{Q},y) = \int_{-\infty}^{\infty} w^d(x) \, k\big(\mathscr{Q}(x), \mathbb{I}\{y \le x\}\big) \, dx$$

be two strictly proper integral scoring rules sharing the same strictly proper binary loss $k : [0,1] \times \{0,1\} \to \mathbb{R}$ with strictly positive continuous weight functions w^u, w^d which both integrate to 1, i.e. have unit total mass. There exists a strictly increasing bijection $\nu : \mathscr{Y} \to \mathscr{Y}$ and an affine function $h : \mathbb{R} \to \mathbb{R}$ such that

$$L^d(\mathscr{Q},y) = h\big(L^u(\mathscr{Q} \circ \nu^{-1}, \nu(y))\big) \quad \forall \mathscr{Q}, y.$$

Moreover, there are usually multiple ways to construct an alignment, if it is attainable. Even in the trivial self-alignment case $s^d = s^u$ for CRPS, the pair (h,ν) is not uniquely identified: any positive affine transformation of ν can be absorbed by a positive scaling of h, yielding infinitely many equivalent pairs. Indeed, let

$$\widehat{\nu}(z) = az + b \quad (a > 0), \qquad \mathscr{Q}^{\widehat{\nu}}(z) := \mathscr{Q}(\widehat{\nu}^{-1}(z)), \qquad \widehat{h}(z) := h\Big(\frac{z}{a}\Big).$$

Since $\mathrm{CRPS}(\mathscr{Q}^{\widehat{\nu}}, \widehat{\nu}(y)) = a\,\mathrm{CRPS}(\mathscr{Q},y)$ (Hersbach 2000a), we have

$$\widehat{h}\big(\mathrm{CRPS}(\mathscr{Q}^{\widehat{\nu}}, \widehat{\nu}(y))\big) = h\Big(\frac{a\,\mathrm{CRPS}(\mathscr{Q},y)}{a}\Big) = h\big(\mathrm{CRPS}(\mathscr{Q},y)\big),$$

so the composite is unchanged. In particular, for self-alignment with $h = \mathrm{id}$ one may take $\widehat{h}(z) = z/a$ and obtain the same result.

For practical applications, where computing the downstream utility can be expensive due to external stochasticity, physical modelling, or heavy numerical simulation, we propose a data-driven procedure to align upstream scores with downstream utility via optimisation task. From now on we treat the losses purely as realised scalar *scores*: we write the downstream score as s^d (formerly L^d) and the upstream score as s^u (formerly L^u); for a validation instance i these are s_i^d and s_i^u, reflecting that, in the alignment step, we operate only on fixed observed scores from a held-out set, with no remaining dependence on $\mathbf{x}, y, \theta$. Using the pairs (s_i^u, s_i^d) on a held-out validation set, we train a small neural network to approximate the downstream score $\hat{s}_i^d$ using the architecture in Fig. 8 under transformations (h, v). We minimise MSE as the alignment loss $\ell(\hat{s}, s) = (\hat{s} - s)^2$ and monitor MAE as an evaluation metric of interest on the held-out set.

$$(\hat{h}, \hat{v}) \in \underset{h \in \mathscr{H},\, v \in \mathscr{V}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell\big(\hat{s}_i^d, s_i^d\big).$$

When a perfect alignment exists within $\mathscr{H} \times \mathscr{V}$, the minimum of this objective is zero; otherwise, the solution yields the best achievable alignment in the chosen families. The training, evaluation, and inference workflows are detailed in Algorithms 1 and 2 (adapted from Publication III).

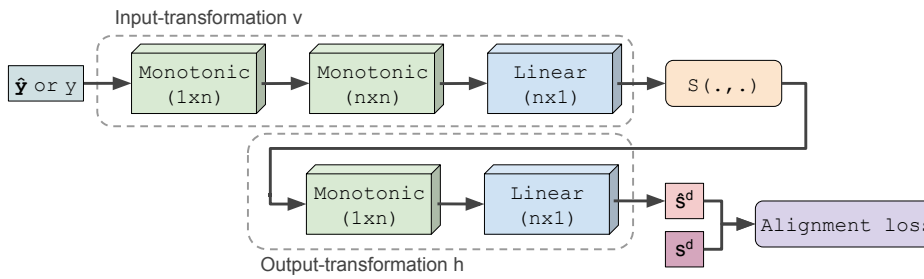

Figure 8: Alignment model architecture consists of monotonic and linear layers. Scoring rule S is passed to the network as an operator within the neural network. The preceding layers of the scoring rule perform transformation v to the scoring rule's input, and the proceeding layers perform tranformation h to the scoring rule's output. Adapted from Publication III.

For this contribution, experimental implementation and execution were carried out by my co-author (Shahroudi, **Komisarenko**, and Kull 2025). The principal aspects of the setup and findings are summarised here for completeness.

Synthetic experiments, where the ground truth downstream utility was generated with threshold-weighted CRPS (Allen 2024) with a known weighting function, demonstrate that the alignment network accurately recovers the ground-truth weighting and achieves near-perfect alignment (alignment error ≈ 0) between the transformed upstream and downstream scores.

Algorithm 1 Alignment Model Training

Input: Validation set $\mathscr{D}_{\text{val}} = \{(\hat{\mathbf{y}}_i, y_i, s_i^d)\}_{i=1}^{n_{val}}$, model f_θ, alignment loss ℓ, optimizer (e.g. SGD)

1: Split $\mathscr{D}_{\text{val}} \to \mathsf{D}_{\text{train}}, \mathsf{D}_{\text{val}}$

2: $\theta^* \leftarrow \arg\min_\theta \mathbb{E}_{(\hat{\mathbf{y}},y,s)\in \mathsf{D}_{\text{train}}}\left[\ell\big(f_\theta(\hat{\mathbf{y}}_i, y), s\big)\right]$

3: (Optionally: monitor $\mathbb{E}_{\mathsf{D}_{\text{val}}}[\ell(f_\theta(\hat{\mathbf{y}}_i, y), s)]$ for early-stopping or hyperparameter tuning)

4: **Output:** f_{θ^*} (trained model)

Algorithm 2 Alignment Model Inference and Evaluation

Input: Test set $\mathscr{D}_{\text{test}} = \{(\hat{\mathbf{y}}_i, y_i, s_i^d)\}_{i=1}^{n_{test}}$, Trained model f_{θ^*}

1: $\hat{s}_i^d \leftarrow f_{\theta^*}(\hat{\mathbf{y}}_i, y_i) \quad \forall i$

2: $\text{MAE} \leftarrow \dfrac{1}{n_{test}} \displaystyle\sum_{i=1}^{n_{test}} |\hat{s}_i^d - s_i^d|$

3: **Output:** Predicted scores $\{\hat{s}_i^d\}_{i=1}^n$, MAE

We considered a practical inventory optimisation problem of a seafood distribution using a Kaggle dataset (Cashion 2024). The base probabilistic demand model was an Exponential Smoothing forecaster (Darts library (Herzen et al. 2022)) that outputs a sample-based predictive distribution for monthly tuna demand. As the upstream evaluation metric, we used CRPS computed on these predictive samples. The downstream score was the *expected monthly profit* in a Newsvendor-style decision: choose an order quantity a_t given a forecast distribution D_t; profit is

$$\pi_t(a_t, d_t) = p_t \min\{d_t, a_t\} - c_t a_t - h_t(a_t - d_t)_+.$$

Here t indexes months; $d_t \sim D_t$ is realised demand from predictive distribution D_t; p_t is the unit selling price (revenue per unit sold), c_t is the unit *procurement cost* (purchase cost per unit ordered), and h_t is the unit *holding cost* applied to leftover inventory; $(x)_+ = \max\{x, 0\}$. Accordingly, $p_t \min\{d_t, a_t\}$ is sales revenue, $c_t a_t$ is total procurement cost, and $h_t(a_t - d_t)_+$ is the total holding cost for unsold units. The downstream score is $\mathbb{E}_{d_t \sim D_t}[\pi_t(a_t, d_t)]$ under the predictive distribution. The alignment for this problem is important because evaluating the *true* downstream profit requires solving an optimisation for each candidate model/hyperparameter (costly and slow), whereas the aligned score is a fast proxy that should preserve downstream ranking for selection and monitoring. We learned the alignment mapping on a held-out validation slice (backtesting 120 months) to transform upstream CRPS into a downstream-aligned score. With our alignment approach, the base predictor need not be retrained, and the aligner can be periodically refreshed when new downstream outcomes arrive.

The learned alignment substantially improves agreement with downstream per-

formance. On the test set, the aligned evaluation increases Kendall's rank correlation τ by $\approx 58\%$ on average relative to the non-aligned baseline; the alignment curves track downstream scores closely, and the learned weight function emphasises decision-relevant regions (Publication III).

5.4. Conclusion

We introduced the alignment problem between upstream metrics and downstream utilities and highlighted practical scenarios where such alignment matters. We proposed a data-driven procedure that learns a mapping from upstream scores to downstream utility using a small feed-forward neural network with specified activation functions. We further showed that the strict properness of the upstream scoring rule is preserved under positive affine transformations of the upstream metric combined with bijective reparameterisations of the predictive distribution and labels. Preserving properness in the downstream proxy keeps the true data-generating distribution as the unique optimum, aligning incentives for calibrated, truthful predictions. In practice, finite data, noise, and modelling constraints prevent reaching the exact optimum, but, similarly to classical supervised learning, using losses/metrics that are proper or close to proper typically reported to yield better, more stable performance.

Experiments demonstrated the possibility of accurate alignment on a few examples, including an inventory optimisation problem. However, the limited experimental scale, still, appropriate for a proof-of-concept, was a key constraint and remains a primary direction for further study. The practical feasibility of applying alignment depends on complex downstream utility computations (e.g., stochastic processes and costly physical or numerical simulations); evaluating performance in such settings is therefore a priority for future work.

6. CONCLUSION

This thesis examines the selection of training losses and evaluation metrics for specific application-oriented settings, including cost sensitivity, calibration, and downstream evaluation, through four distinct but interconnected studies. Its central objective is to develop practical guidelines for aligning loss choice with task-specific evaluation metrics, informed by proper scoring rule theory. We conclude by discussing the broader implications of our findings, noting current limitations, and outlining emerging trends and directions for future research.

RQ1*: Which loss function should be used to minimise cost-sensitive evaluation metrics (e.g., expected total cost), especially in the presence of class-cost uncertainty.* In the **Publication I**, we derive families of proper losses that are mathematically equivalent to the expected total cost under class-cost uncertainty. We propose a Beta family of losses, which conveniently models diverse cost scenarios and admits simple, analytic, smooth gradients with respect to predicted probabilities, which supports efficient gradient-based optimisation. In extensive experiments, training deep neural networks with Beta losses using balanced hyperparameters ($\alpha \approx \beta$ in a range scale 5-25) consistently improved cost-sensitive evaluation performance across diverse cost scenarios compared with classical cross-entropy baseline. For context, prior work reports that, in common deep learning settings, unweighted cross-entropy often matches or exceeds naïve class-weighted variants and probability/threshold rescaling, especially when combined with class-balanced (class-proportional) resampling (Buda, Maki, and Mazurowski 2018). Standard post-hoc calibration (e.g., temperature scaling) further improved performance, particularly when the temperature was selected on a validation set to optimise the cost-sensitive metric of interest.

RQ2*: Which proper loss families perform best in cost-sensitive settings when the evaluation metric exactly matches the training loss.* As shown in our experiments, the Beta family of losses outperformed classical cross-entropy. Although both cross-entropy and the Beta family are proper losses, we also compared them with improper cross-entropy variants such as focal loss and label smoothing. In that setting, the Beta loss advantage was less pronounced, suggesting that hybrid objectives, including adding focal- or label-smoothing–style modulation to a representative Beta loss, may yield additional gains.

Experiments also showed that seemingly minor hyperparameter and protocol choices (often under-reported) can substantially affect measured performance. Key factors include the rule for selecting the 'best' epoch (e.g., the last epoch after a fixed number of epochs; early stopping based on a validation metric such as accuracy, log loss, or a cost-sensitive metric; or taking the kth best epoch), the criterion used for temperature scaling (i.e., how the optimal temperature is chosen), the learning rate and other optimisation settings, and the often substantial performance variation across epochs of the same model.

Finally, although the experiments were substantial, we did not cover datasets

from other modalities (e.g., text, audio, or video), nor did we examine the full spectrum of model classes - from simple, few-layer NN and classical ML models to high-compute LLM architectures. Moreover, evaluating the Beta loss under extreme class imbalance or high levels of label noise would be valuable for assessing its performance in such severe conditions.

Still, we believe that Beta losses offer a promising direction for minimising cost-sensitive metrics, and further modifications on top of Beta losses may yield additional gains.

RQ3*: Why does the (improper) focal loss yield better calibration than the commonly used (proper) cross-entropy loss as reported by (Mukhoti et al. 2020), and how does focal loss relate to properness.* In **Publication II**, we decompose focal loss into a proper loss component and a calibration map component. This shows that training with focal loss is equivalent to optimising a proper loss under a fixed calibration of the predicted probabilities. We further show that the calibration map acts similarly to temperature scaling, with an nearly explicit correspondence in the binary case. However, for the standard positive range of the focal parameter ($\gamma > 0$), the derived calibration map behaves like temperature scaling with $0 < T < 1$: it sharpens the predicted probabilities, increasing confidence. In other words, it incentivises the sharpened probabilities to be calibrated, and hence, when sharpening is switched off, then the probabilities are less confident than they otherwise would be.

In our experiments, this transformation helped produce more reliable probabilities and reduced the validation set generalisation gap (Guo et al. 2017), which may help explain focal loss's reported well-calibrated validation set performance (Charoenphakdee et al. 2021).

Our experiments, conducted on a small set of classical image classification tasks with ResNet-based architectures, showed that applying the derived focal calibration map post hoc, particularly when paired with temperature scaling to balance under- and overfitting, improves calibration over standard temperature scaling. This suggests the potential for training loss-specific families of calibration maps and motivates wider empirical testing in diverse real-world settings.

RQ4*: Can we decompose any separable loss into a proper component and a fixed calibration map component, and can some separable losses perform comparably to cross-entropy.* We extend the focal loss decomposition (into a proper component and a fixed calibration map component) to the broader class of *separable* losses under mild assumptions (continuous differentiability, convexity, non-positive first derivative, and appropriate boundary behaviour). We then study the inverse question: given a calibration map, does there exist a corresponding separable loss (and associated proper base loss) that induces it? Under mild conditions (continuity and positivity, permutation equivariance, odds-dependence, and a cocycle-consistency (see Chapter 4)), the calibration map is essentially unique (up to a positive affine transformation) and determines both the proper base loss and the separable loss. We propose several new separable losses with

their corresponding calibration maps (obtained via this decomposition) and, on four small/medium scale datasets, observe that these losses can occasionally outperform cross-entropy and focal loss baselines in both predictive accuracy and calibration, especially when the post-hoc calibration family and its hyperparameters are selected on a validation set. In practice, the best-performing trained model and calibration map often come from different function families. Conceptually, these results complement convergence analyses by illustrating how alternative separable losses can perform on par with, or better than, the classical proper separable loss, cross-entropy.

RQ5*: Can upstream and downstream regression evaluation be aligned via a proxy function learned on validation data.* In the **third contribution**, we propose a simple methodology for aligning upstream and downstream regression evaluation via a data-driven proxy learned on a held-out validation set. A small, few-layer neural network is trained to approximate the downstream metric and then fixed for evaluation, keeping the proxy independent of hand-crafted cost or weight structures. While promising, the approach would benefit from broader validation across datasets and architectures. It is particularly useful when downstream utility is costly to compute, depends on variables not known a priori, or requires simulation. Limitations include optimisation on a finite validation set (with attendant risks of overfitting or hyperparameter sensitivity), and potential sensitivity to distribution shift.

RQ6*: Under what assumptions on the transformations do both the upstream and downstream metrics remain proper.* We showed that when the upstream metric is a strictly proper regression scoring rule, the downstream metric remains strictly proper if the transformation consists of an inner bijective reparameterisation applied to both the predictive CDF and the observed outcome, together with a positive affine outer (score-level) transformation. The key principle lies in preserving the Bayes-risk minimiser: the location of the minimum remains unchanged under these transformations. These constraints are fairly strict, and maintaining properness may be challenging in practical applications.

Trends and Opportunities for Future Research

The implications of this work contribute to the fields of proper scoring rules and loss-metric alignment, i.e., selecting losses that target the evaluation metric of interest. Over the period of this thesis, the research landscape shifted remarkably with the rise of foundation models, especially LLMs and emerging multimodal systems, and their rapid adoption across application domains. Notably, recent studies report divergent calibration behaviours for LLMs compared to medium-sized vision models (Guo et al. 2017): pre-trained LLMs are often roughly well-calibrated or even under-confident, but become poorly calibrated after preference alignment/RLHF, with post-hoc methods required to restore calibration (Xie et al. 2024).

In parallel, theory and evaluation practice have been re-examined: (Blasiok et al. 2023) caution that optimising a proper loss need not yield calibrated predictors in practice, and multiple works highlight pitfalls of ECE (binning bias, estimator inconsistency) and advocate for more careful calibration assessment (Minderer et al. 2021; Tygert 2025).

The research frontier is also moving in time series and tabular learning, with foundation models such as TabPFN for tabular data and Chronos for time series being introduced, followed by newer studies (Ansari et al. 2024; Hollmann et al. 2022).

Overall, the field is dynamic: there is a growing demand to quantify calibration, reliability and cost for large models, and, at the same time, to provide practical, domain-aware guidance for practitioners who may not use massive models but still require guarantees tied to their task-specific metrics.

In summary, the three publications in this thesis advance loss–metric alignment, particularly for cost-sensitive and calibration settings. We focus on aligning the training objective with minimisation of the metric of interest, at both theoretical and applied levels.

Looking ahead, a natural research plan is to better understand how training with proper losses works in practice. This includes limited data regimes, architectures that are not expressive enough or overly expressive (e.g., LLMs), and how performance carries over from the training set to a held-out set, or under distribution shift and out-of-distribution data. Many losses have been proposed (proper, modified-proper, and non-proper), yet it is often hard to compare them fairly or to explain why one wins in a given setting, as many studies that propose new approaches provide limited experimental evidence to support the effectiveness of their method in practice. This motivates an open benchmarking effort focused on predictive confidence quality across a variety of datasets and architectures. Ideally, the benchmark would be partly crowdsourced, with submitted methods carefully annotated and run under shared protocols. A companion benchmark dedicated to post-hoc calibration methods would also be helpful. Together, these studies could close part of the gap between clean theory and messy practice: even in this thesis, choices that seem theoretically well-matched, including pairing a cost distribution of interest with the loss induced by that same distribution, or using a post-hoc focal temperature map that mirrors a trained focal loss, do not always perform best in practice. Existing loss benchmarks are helpful, but they rarely centre calibration and probability estimation; their design ideas could be reused while shifting the focus. The ideal outcome would be a simple, decision-tree style guide for practitioners: given dataset size, feature characteristics, class imbalance or target distribution, and downstream needs, which training loss and evaluation metric should one use? On the theory side, two directions seem especially promising. First, study the interaction between training losses and post-hoc calibration maps more deeply: when do they reinforce each other, when do they conflict, and can we design them jointly? Second, explore being properly improper idea:

deliberate, principled departures from strict properness, as we have in focal loss, may improve both discrimination and calibration in practice. Extending current decompositions (proper loss + fixed calibration map) beyond separable losses is another concrete target. Finally, for cost-sensitive learning, more work is needed on rare, high-penalty events. Our experiments mostly considered imbalance up to 1:10; going well beyond that, and deciding what to measure and how to optimise under extreme skew, is both challenging and practically important.

Collectively, the publications in this thesis highlight the critical importance of training loss-metric alignment, especially for cost-sensitive and calibration metrics, and its application across diverse domains. We hope this thesis provides valuable insights and practical guidelines, and ultimately supports broader adoption of metric-informed training loss selection.

BIBLIOGRAPHY

1. **Komisarenko**, **Viacheslav** and Kull, Meelis (2025). "Cost-sensitive classification with cost uncertainty: do we need surrogate losses?" In: *Machine Learning* 114.132, pp. 1–36. DOI: 10.1007/s10994-024-06634-8.
2. **Komisarenko**, **Viacheslav** and Kull, Meelis (2024). "Improving Calibration by Relating Focal Loss, Temperature Scaling, and Properness". In: *European Conference on Artificial Intelligence*. IOS Press, pp. 1535–1542. DOI: 10.3233/FAIA240658.
3. Shahroudi, Novin, **Komisarenko**, **Viacheslav**, and Kull, Meelis (2025). "Aligning the Evaluation of Probabilistic Predictions with Downstream Value". In: *European Conference on Artificial Intelligence*. IOS Press, pp. 1969–1976. DOI: 10.3233/FAIA251032.
4. Burdun, Iuliia, Bechtold, Michel, Aurela, Mika, De Lannoy, Gabrielle, Desai, Ankur R, Humphreys, Elyn, Kareksela, Santtu, **Komisarenko**, **Viacheslav**, Liimatainen, Maarit, Marttila, Hannu, et al. (2023). "Hidden becomes clear: Optical remote sensing of vegetation reveals water table dynamics in northern peatlands". In: *Remote Sensing of Environment* 296. DOI: 10.1016/j.rse.2023.113736.
5. **Komisarenko**, **Viacheslav**, Voormansik, Kaupo, Elshawi, Radwa, and Sakr, Sherif (2022). "Exploiting time series of Sentinel-1 and Sentinel-2 to detect grassland mowing events using deep learning with reject region". In: *Scientific Reports* 12.1. DOI: 10.1038/s41598-022-04932-6.
6. Ingel, Anti, Shahroudi, Novin, Kängsepp, Markus, Tättar, Andre, **Komisarenko**, **Viacheslav**, and Kull, Meelis (2020). "Correlated daily time series and forecasting in the M4 competition". In: *International Journal of Forecasting* 36.1, pp. 121–128. DOI: 10.1016/j.ijforecast.2019.02.018.
7. Burdun, Iuliia, Bechtold, Michel, Sagris, Valentina, **Komisarenko**, **Viacheslav**, De Lannoy, Gabrielle, and Mander, Ülo (2020). "A Comparison of Three Trapezoid Models Using Optical and Thermal Satellite Imagery for Water Table Depth Monitoring in Estonian Bogs". In: *Remote Sensing* 12.12. DOI: 10.5194/egusphere-egu21-4698.
8. Saerens, Marco, Latinne, Patrice, and Decaestecker, Christine (2002). "Any reasonable cost function can be used for a posteriori probability approximation". In: *IEEE transactions on neural networks* 13.5, pp. 1204–1210.
9. Gubarev, Vyacheslav F, Boyun, Vitaliy P, Melnichuk, Sergey V, Salnikov, Nikolay N, Simakov, Vladimir A, Godunok, Leonid A, **Komisarenko**, **Vyacheslav** I, Dobrovolskyy, Victor Yu, Derkach, Sergey V, and Matviyenko, Sergey A (2016). "Using Vision Systems for Determining the Parameters of Relative Motion of Spacecrafts". In: *Journal of Automation and Information Sciences* 48.11. DOI: 10.1615/JAutomatInfScien.v48.i11.30.

10. Hernández-Orallo, José, Flach, Peter, and Ferri, Cèsar (2012). "A unified view of performance metrics: translating threshold choice into expected classification loss". In: *Journal of Machine Learning Research* 13.Oct, pp. 2813–2869.
11. Cashion, Tim (2024). *Tsukiji Tuna Prices - Time Series*. https://www.kaggle.com/datasets/tcashion/tokyo-wholesale-tuna-prices. Accessed: 2025-02-22.
12. Miller, John W, Goodman, Rod, and Smyth, Padhraic (1991). "Objective functions for probability estimation." In: *International Joint Conference on Neural Networks.* Vol. 1, pp. 881–886.
13. Cid-Sueiro, Jesús, Arribas, Juan Ignacio, Urbán-Munoz, Sebastián, and Figueiras-Vidal, Aníbal R (1999). "Cost functions to estimate a posteriori probabilities in multiclass problems". In: *IEEE Transactions on Neural Networks* 10.3, pp. 645–656.
14. Bao, Han and Charoenphakdee, Nontawat (2025). "Calm Composite Losses: Being Improper Yet Proper Composite". In: *The 28th International Conference on Artificial Intelligence and Statistics*.
15. Ouyang, Long, Wu, Jeffrey, Jiang, Xu, Almeida, Diogo, Wainwright, Carroll, Mishkin, Pamela, Zhang, Chong, Agarwal, Sandhini, Slama, Katarina, Ray, Alex, et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35, pp. 27730–27744.
16. Herzen, Julien, Lässig, Francesco, Piazzetta, Samuele Giuliano, Neuer, Thomas, Tafti, Léo, Raille, Guillaume, Van Pottelbergh, Tomas, Pasieka, Marek, Skrodzki, Andrzej, Huguenin, Nicolas, et al. (2022). "Darts: User-friendly modern machine learning for time series". In: *Journal of Machine Learning Research* 23.124, pp. 1–6.
17. Alicia Guerrero-Curieses and Jesús Cid-Sueiro and Rocío Alaiz-Rodríguez and Aníbal R. Figueiras-Vidal (2004). "Local Estimation of Posterior Class Probabilities to Minimize Classification Errors". In: *IEEE Transactions on Neural Networks* 15.2, pp. 309–317. DOI: 10.1109/TNN.2004.824266.
18. Pearson, Karl (1896). "VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia". In: *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 187, pp. 253–318.
19. Galton, Francis (1886). "Regression towards mediocrity in hereditary stature". In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15, pp. 246–263.
20. Allen, Sam (2024). "Weighted scoringRules: Emphasizing Particular Outcomes When Evaluating Probabilistic Forecasts". In: *Journal of Statistical Software* 110.8, pp. 1–26. DOI: 10.18637/jss.v110.i08. URL: https://www.jstatsoft.org/index.php/jss/article/view/v110i08.

21. Gupta, Chirag and Ramdas, Aaditya (2022). “Top-label calibration and multiclass-to-binary reductions”. In: *International Conference on Learning Representations (ICLR)*.
22. Müller, Rafael, Kornblith, Simon, and Hinton, Geoffrey E (2019). “When does label smoothing help?” In: *Advances in neural information processing systems* 32.
23. Thulasidasan, Sunil, Chennupati, Gopinath, Bilmes, Jeff A, Bhattacharya, Tanmoy, and Michalak, Sarah (2019). “On mixup training: Improved calibration and predictive uncertainty for deep neural networks”. In: *Advances in neural information processing systems* 32.
24. Kendall, Alex and Gal, Yarin (2017). “What uncertainties do we need in bayesian deep learning for computer vision?” In: *Advances in neural information processing systems* 30.
25. Gal, Yarin and Ghahramani, Zoubin (2016). “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR, pp. 1050–1059.
26. Abe, Taiga, Buchanan, Estefany Kelly, Pleiss, Geoff, Zemel, Richard, and Cunningham, John P (2022). “Deep ensembles work, but are they necessary?” In: *Advances in Neural Information Processing Systems* 35, pp. 33646–33660.
27. Bishop, Christopher M and Nasrabadi, Nasser M (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.
28. Kumar, Aviral, Sarawagi, Sunita, and Jain, Ujjwal (2018). “Trainable calibration measures for neural networks from kernel mean embeddings”. In: *International Conference on Machine Learning*. PMLR, pp. 2805–2814.
29. Niculescu-Mizil, Alexandru and Caruana, Rich (2005). “Predicting good probabilities with supervised learning”. In: *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632.
30. He, Haibo and Garcia, Edwardo A (2009). “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284.
31. Der Kiureghian, Armen and Ditlevsen, Ove (2009). “Aleatory or epistemic? Does it matter?” In: *Structural safety* 31.2, pp. 105–112.
32. Rosenblatt, Frank (1958). “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6, p. 386.
33. Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J (1986). “Learning representations by back-propagating errors”. In: *nature* 323.6088, pp. 533–536.
34. Cortes, Corinna and Vapnik, Vladimir (1995). “Support-vector networks”. In: *Machine learning* 20.3, pp. 273–297.

35. Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E (2012). “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25.
36. Gneiting, Tilmann and Raftery, Adrian E (2007). “Strictly proper scoring rules, prediction, and estimation”. In: *Journal of the American statistical Association* 102.477, pp. 359–378.
37. Glenn, W Brier et al. (1950). “Verification of forecasts expressed in terms of probability”. In: *Monthly weather review* 78.1, pp. 1–3.
38. Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan (2014). “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1, pp. 1929–1958.
39. Minderer, Matthias, Djolonga, Josip, Romijnders, Rob, Hubis, Frances, Zhai, Xiaohua, Houlsby, Neil, Tran, Dustin, and Lucic, Mario (2021). “Revisiting the calibration of modern neural networks”. In: *Advances in neural information processing systems* 34, pp. 15682–15694.
40. Domingos, Pedro (1999). “Metacost: A general method for making classifiers cost-sensitive”. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 155–164.
41. Ansari, Abdul Fatir, Stella, Lorenzo, Turkmen, Caner, Zhang, Xiyuan, Mercado, Pedro, Shen, Huibin, Shchur, Oleksandr, Rangapuram, Syama Sundar, Arango, Sebastian Pineda, Kapoor, Shubham, et al. (2024). “Chronos: Learning the language of time series”. In: *arXiv preprint arXiv:2403.07815*.
42. Hollmann, Noah, Müller, Samuel, Eggensperger, Katharina, and Hutter, Frank (2022). “Tabpfn: A transformer that solves small tabular classification problems in a second”. In: *arXiv preprint arXiv:2207.01848*.
43. Tygert, Mark (Apr. 2025). *Calibration and Bias in Algorithms, Data, and Models: A Tutorial on Metrics and Plots for Measuring Calibration, Bias, Fairness, Reliability, and Robustness*. Tutorial, International Conference on Machine Learning (ICML), Vancouver, Canada. Zenodo, Version v2; slide deck and recording available. DOI: 10.5281/zenodo.15253140. URL: https://doi.org/10.5281/zenodo.15253140.
44. Xie, Johnathan, Chen, Annie, Lee, Yoonho, Mitchell, Eric, and Finn, Chelsea (2024). “Calibrating Language Models with Adaptive Temperature Scaling”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18128–18138.
45. Blasiok, Jaroslaw, Gopalan, Parikshit, Hu, Lunjia, and Nakkiran, Preetum (2023). “When does optimizing a proper loss yield calibration?” In: *Advances in Neural Information Processing Systems* 36, pp. 72071–72095.

46. Lienen, Julian and Hüllermeier, Eyke (2021). “From label smoothing to label relaxation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 10, pp. 8583–8591.
47. Hersbach, Hans (2000a). “Decomposition of the continuous ranked probability score for ensemble prediction systems”. In: *Weather and Forecasting* 15.5, pp. 559–570.
48. Brown, Tom, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, et al. (2020). “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33, pp. 1877–1901.
49. Hersbach, Hans (2000b). “Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems”. In: *Weather and Forecasting* 15.5, pp. 559–570.
50. Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron (2016). *Deep Learning*. MIT Press. URL: https://www.deeplearningbook.org.
51. Hilbert, Martin and López, Priscila (2011). “The World’s Technological Capacity to Store, Communicate, and Compute Information”. In: *Science* 332.6025, pp. 60–65. DOI: 10.1126/science.1200970.
52. Ashish, Vaswani (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30, p. I.
53. Radford, Alec, Narasimhan, Karthik, Salimans, Tim, Sutskever, Ilya, et al. (2018). “Improving language understanding by generative pre-training”. In.
54. Hüllermeier, Eyke and Waegeman, Willem (2021). “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods”. In: *Machine learning* 110.3, pp. 457–506.
55. Wang, Deng-Bao, Feng, Lei, and Zhang, Min-Ling (2021). “Rethinking calibration of deep neural networks: Do not be afraid of overconfidence”. In: *Advances in Neural Information Processing Systems* 34, pp. 11809–11820.
56. Xiao, Han, Rasul, Kashif, and Vollgraf, Roland (2017). “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms”. In: *arXiv preprint arXiv:1708.07747*.
57. Havaei, Mohammad, Davy, Axel, Warde-Farley, David, Biard, Antoine, Courville, Aaron, Bengio, Yoshua, Pal, Chris, Jodoin, Pierre-Marc, and Larochelle, Hugo (2017). “Brain tumor segmentation with deep neural networks”. In: *Medical image analysis* 35, pp. 18–31.
58. Bartlett, Peter L, Jordan, Michael I, and McAuliffe, Jon D (2006). “Convexity, classification, and risk bounds”. In: *Journal of the American Statistical Association* 101.473, pp. 138–156.
59. Japkowicz, Nathalie and Stephen, Shaju (2002). “The class imbalance problem: A systematic study”. In: *Intelligent data analysis* 6.5, pp. 429–449.

60. Rybizki, Lydia (2014). *Learning cost sensitive binary classification rules accounting for uncertain and unequal misclassification costs*. Tech. rep. IWQW Discussion Papers.
61. Johnson, Justin M and Khoshgoftaar, Taghi M (2019). "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6.1, pp. 1–54.
62. Elkan, Charles (2001). "The foundations of cost-sensitive learning". In: *International joint conference on artificial intelligence*. Lawrence Erlbaum Associates Ltd, pp. 973–978.
63. Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
64. Schervish, Mark J (1989). "A general method for comparing probability assessors". In: *The annals of statistics* 17.4, pp. 1856–1879.
65. Drummond, Chris and Holte, Robert C (2006). "Cost curves: An improved method for visualizing classifier performance". In: *Machine Learning* 65, pp. 95–130.
66. Kukar, Matjaz, Kononenko, Igor, et al. (1998). "Cost-sensitive learning with neural networks." In: *ECAI*. Vol. 15. Citeseer, pp. 88–94.
67. Shuford Jr, Emir H, Albert, Arthur, and Edward Massengill, H (1966). "Admissible probability measurement procedures". In: *Psychometrika* 31.2, pp. 125–145.
68. Zhao, Shengjia, Kim, Michael P, Sahoo, Roshni, Ma, Tengyu, and Ermon, Stefano (2021). "Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibration". In: *Thirty-Fifth Conference on Neural Information Processing Systems*.
69. Schapire, Robert E (2013). "Explaining adaboost". In: *Empirical inference*. Springer, pp. 37–52.
70. Collell, Guillem, Prelec, Drazen, and Patil, Kaustubh (2016). "Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data". In: *arXiv preprint arXiv:1606.08698*.
71. Bröcker, Jochen (2009). "Reliability, sufficiency, and the decomposition of proper scores". In: *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography* 135.643, pp. 1512–1519.
72. Flach, Peter A (2015). *Cost-sensitive classification meets proper scoring rules*.
73. Pereyra, Gabriel, Tucker, George, Chorowski, Jan, Kaiser, Łukasz, and Hinton, Geoffrey (2017). "Regularizing neural networks by penalizing confident output distributions". In: *arXiv preprint arXiv:1701.06548*.
74. Demšar, Janez (2006). "Statistical comparisons of classifiers over multiple data sets". In: *The Journal of Machine learning research* 7, pp. 1–30.

75. Yang, Jiancheng, Shi, Rui, Wei, Donglai, Liu, Zequan, Zhao, Lin, Ke, Bilian, Pfister, Hanspeter, and Ni, Bingbing (2023). "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification". In: *Scientific Data* 10.1, p. 41.
76. Ghosh, Arindam, Schaaf, Thomas, and Gormley, Matthew (2022). "AdaFocal: Calibration-aware Adaptive Focal Loss". In: *Advances in Neural Information Processing Systems* 35, pp. 1583–1595.
77. Simkanin, Lucia (2020). "Multi-emotion Recognition and Dialogue Manager for VR-based Self-attachment Therapy". PhD thesis. Imperial College London.
78. Scott, Clayton et al. (2012). "Calibrated asymmetric surrogate losses". In: *Electronic Journal of Statistics* 6, pp. 958–992.
79. Kingma, Diederik P and Ba, Jimmy (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.
80. Song, Hao, Diethe, Tom, Kull, Meelis, and Flach, Peter (2019). "Distribution calibration for regression". In: *International Conference on Machine Learning*. PMLR, pp. 5897–5906.
81. Zadrozny, Bianca and Elkan, Charles (2002). "Transforming classifier scores into accurate multiclass probability estimates". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699.
82. Platt, John et al. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3, pp. 61–74.
83. Kull, Meelis, Silva Filho, Telmo, and Flach, Peter (2017). "Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers". In: *Artificial Intelligence and Statistics*. PMLR, pp. 623–631.
84. Masnadi-Shirazi, Hamed and Vasconcelos, Nuno (2015). "A view of margin losses as regularizers of probability estimates". In: *The Journal of Machine Learning Research* 16.1, pp. 2751–2795.
85. Hand, David J (2009). "Measuring classifier performance: a coherent alternative to the area under the ROC curve". In: *Machine learning* 77.1, pp. 103–123.
86. — (2010). "Evaluating diagnostic tests: the area under the ROC curve and the balance of errors". In: *Statistics in medicine* 29.14, pp. 1502–1510.
87. Hand, David J and Anagnostopoulos, Christoforos (2014). "A better Beta for the H measure of classification performance". In: *Pattern Recognition Letters* 40, pp. 41–46.
88. Zou, Hui, Zhu, Ji, and Hastie, Trevor (2008). "New multicategory boosting algorithms based on multicategory fisher-consistent losses". In: *The Annals of Applied Statistics* 2.4, p. 1290.

89. Williamson, Robert, Vernet, Elodie, Reid, Mark, et al. (2016). “Composite multiclass losses”. In.
90. Boyd, Stephen, Boyd, Stephen P, and Vandenberghe, Lieven (2004). *Convex optimization*. Cambridge university press.
91. Buda, Mateusz, Maki, Atsuto, and Mazurowski, Maciej A (2018). “A systematic study of the class imbalance problem in convolutional neural networks”. In: *Neural Networks* 106, pp. 249–259.
92. Lipton, Zachary C, Elkan, Charles, and Naryanaswamy, Balakrishnan (2014). “Optimal thresholding of classifiers to maximize F1 measure”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 225–239.
93. Ye, Nan, Chai, Kian Ming, Lee, Wee Sun, and Chieu, Hai Leong (2012). “Optimizing F-measures: a tale of two approaches”. In: *Proceedings of the 29th International Conference on Machine Learning*. Omnipress, pp. 289–296.
94. Zadrozny, Bianca, Langford, John, and Abe, Naoki (2003). “Cost-sensitive learning by cost-proportionate example weighting”. In: *Third IEEE international conference on data mining*. IEEE, pp. 435–442.
95. Lin, Yi, Lee, Yoonkyung, and Wahba, Grace (2002). “Support vector machines for classification in nonstandard situations”. In: *Machine learning* 46.1, pp. 191–202.
96. Cui, Yin, Jia, Menglin, Lin, Tsung-Yi, Song, Yang, and Belongie, Serge (2019). “Class-balanced loss based on effective number of samples”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277.
97. Huang, Chen, Li, Yining, Loy, Chen Change, and Tang, Xiaoou (2016). “Learning deep representation for imbalanced classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384.
98. Li, Buyu, Liu, Yu, and Wang, Xiaogang (2019). “Gradient harmonized single-stage detector”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33, pp. 8577–8584.
99. Yu, Shuang, Li, Xiongfei, Zhang, Xiaoli, and Wang, Hancheng (2019). “The OCS-SVM: An objective-cost-sensitive SVM with sample-based misclassification cost invariance”. In: *IEEE Access* 7, pp. 118931–118942.
100. Pendharkar, P (2009). “Misclassification cost minimizing fitness functions for genetic algorithm-based artificial neural network classifiers”. In: *Journal of the Operational Research Society* 60.8, pp. 1123–1134.
101. Pendharkar, Parag and Nanda, Sudhir (2006). “A misclassification cost-minimizing evolutionary–neural classification approach”. In: *Naval Research Logistics (NRL)* 53.5, pp. 432–447.

102. Cao, Kaidi, Wei, Colin, Gaidon, Adrien, Arechiga, Nikos, and Ma, Tengyu (2019). “Learning imbalanced datasets with label-distribution-aware margin loss”. In: *Advances in neural information processing systems* 32.
103. Ramentol, Enislay, Caballero, Yailé, Bello, Rafael, and Herrera, Francisco (2012). “SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory”. In: *Knowledge and information systems* 33.2, pp. 245–265.
104. Chawla, Nitesh V, Bowyer, Kevin W, Hall, Lawrence O, and Kegelmeyer, W Philip (2002). “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16, pp. 321–357.
105. Zhang, Zhilu and Sabuncu, Mert (2018). “Generalized cross entropy loss for training deep neural networks with noisy labels”. In: *Advances in neural information processing systems* 31.
106. Ding, Nan and Vishwanathan, SVN (2010). “t-Logistic regression”. In: *Advances in Neural Information Processing Systems* 23.
107. Ding, Nan (2013). “Statistical machine learning in the t-exponential family of distributions”. PhD thesis. Purdue University.
108. Amid, Ehsan, Warmuth, Manfred K, and Srinivasan, Sriram (2019). “Two-temperature logistic regression based on the tsallis divergence”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2388–2396.
109. Godambe, Vidyadhar P (1960). “An optimum property of regular maximum likelihood estimation”. In: *The Annals of Mathematical Statistics* 31.4, pp. 1208–1211.
110. Wright, Raymond E (1995). “Logistic regression.” In.
111. Balakrishnama, Suresh and Ganapathiraju, Aravind (1998). “Linear discriminant analysis-a brief tutorial”. In: *Institute for Signal and information Processing* 18.1998, pp. 1–8.
112. Iranmehr, Arya, Masnadi-Shirazi, Hamed, and Vasconcelos, Nuno (2019). “Cost-sensitive support vector machines”. In: *Neurocomputing* 343, pp. 50–64.
113. Amid, Ehsan, Warmuth, Manfred KK, Anil, Rohan, and Koren, Tomer (2019). “Robust Bi-Tempered Logistic Loss Based on Bregman Divergences”. In: *Advances in Neural Information Processing Systems* 32.
114. Buja, Andreas, Stuetzle, Werner, and Shen, Yi (2005). “Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications”. In: *Working draft, November* 3.
115. Charoenphakdee, Nontawat, Vongkulbhisal, Jayakorn, Chairatanakul, Nuttapong, and Sugiyama, Masashi (2021). “On Focal Loss for Class-Posterior Probability Estimation: A Theoretical Perspective”. In: *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5202–5211.

116. Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li (2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248–255.
117. Guo, Chuan, Pleiss, Geoff, Sun, Yu, and Weinberger, Kilian Q (2017). "On Calibration of Modern Neural Networks ". In: *International Conference on Machine Learning*. PMLR, pp. 1321–1330.
118. He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian (2016). "Deep Residual Learning for Image Recognition ". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
119. Komisarenko, Viacheslav (2024). *Code Repository for Improving Calibration by Relating Focal Loss, Temperature Scaling, and Properness*. Accessed: 20.08.2024. URL: `https://github.com/slavikkom/focal_temperature_scaling`.
120. Krizhevsky, Alex, Nair, Vinod, and Hinton, Geoffrey (2009). "CIFAR-10 and CIFAR-100 datasets". In: *URl: https://www.cs.toronto.edu/kriz/cifar.html* 6.
121. Kull, Meelis, Perello Nieto, Miquel, Kängsepp, Markus, Silva Filho, Telmo, Song, Hao, and Flach, Peter (2019). "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration". In: *Advances in Neural Information Processing Systems* 32.
122. Lin, Tsung-Yi, Goyal, Priya, Girshick, Ross, He, Kaiming, and Dollár, Piotr (2017). "Focal Loss for Dense Object Detection". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988.
123. Mukhoti, Jishnu, Kulharia, Viveka, Sanyal, Amartya, Golodetz, Stuart, Torr, Philip, and Dokania, Puneet (2020). "Calibrating Deep Neural Networks using Focal Loss ". In: *Advances in Neural Information Processing Systems* 33, pp. 15288–15299.
124. Naeini, Mahdi Pakdaman, Cooper, Gregory, and Hauskrecht, Milos (2015). "Obtaining Well Calibrated Probabilities Using Bayesian Binning ". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1.
125. Peters, Ben, Niculae, Vlad, and Martins, André FT (2019). "Sparse Sequence-to-Sequence Models". In: *arXiv preprint arXiv:1905.05702*.
126. Reid, Mark D and Williamson, Robert C (2010). "Composite Binary Losses". In: *The Journal of Machine Learning Research* 11, pp. 2387–2422.
127. Silva Filho, Telmo, Song, Hao, Perello-Nieto, Miquel, Santos-Rodriguez, Raul, Kull, Meelis, and Flach, Peter (2023). "Classifier calibration: a survey on how to assess and improve predicted class probabilities". In: *Machine Learning*, pp. 1–50.
128. Wang, Cheng (2023). "Calibration in Deep Learning: A Survey of the State-of-the-Art ". In: *arXiv preprint arXiv:2308.01222*.

129. Wang, Cheng, Balazs, Jorge, Szarvas, György, Ernst, Patrick, Poddar, Lahari, and Danchenko, Pavel (2022). “Calibrating Imbalanced Classifiers with Focal Loss: An Empirical Study”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 145–153.
130. Komisarenko, Viacheslav and Kull, Meelis (2024). “Improving Calibration by Relating Focal Loss, Temperature Scaling, and Properness”. In: *arXiv preprint arXiv:2408.11598*. URL: `http://arxiv.org/abs/2408.11598`.

ACKNOWLEDGEMENTS

My first thought on starting a PhD at the University of Tartu dates back to the Ph.D. introduction event in 2018. I remember that day, from meeting with Meelis Kull earlier that day at the old Computer Science building to discuss a simple theoretical problem related to the expected misclassification cost formulation, to the event itself in the building opposite the main university building. I remember prof. Marlon Dumas saying to the audience that the main recipe for success for the PhD journey is a synergy between the student and the supervisor. And after more than seven years since that event, in my own experience, I can only agree with it.

That is why my first thank-you goes to my supervisor, prof. Meelis Kull. Thank you for leading by example, showing how a real scientist should look and which qualities one should have and continue to improve over years. Thank you for being a kind person and for many pleasant memories during formal and informal events alike. These include basketball, fishing, sauna, board games and long hours near the whiteboard when paper deadlines were approaching.

I also thank my opponents, Prof. Cid-Sueiro and Assoc. Prof. Filippone, for your careful reading and insightful feedback on my dissertation. You managed to ask questions that have not occurred to me for years, and to hint towards existing works that I have not managed to find before. Special thanks go to the internal reviewer, Assoc. Prof. Laur, for very detailed but fair comments and suggestions for improving the manuscript.

I express my gratitude to the administration of the Institute of Computer Science for always being approachable and for making all administrative processes really smooth; this makes PhD life easier.

Thank you my university colleagues from ML and neighboring groups (e.g., NAIL). This includes (but is not limited to): Novin Shahroudi, Mari-Liis Allikivi, Marharyta Domnich, Markus Kängsepp, Joonas Järve, Mihkel Lepson, as well as newcomers Marek Leibl, Heili Aavola, Halil Ibrahim Aysel, Laura Altin. Thank you for all team events, including simple dinners, board games, basketball, and drawing; there were many! I hope that all of you who are not PhDs yet will become one soon.

Special thanks go to Novin Shahroudi, the only colleague who started PhD studies in the same year as I did and the only colleague so far whose collaboration with me resulted in publishing a research paper.

Now I will smoothly transition from colleagues to friends in the gratitude section, as friends supported me, my well being, and created warm memories during these years, and we have had great times together and I hope we will have many more.

To make this transition as smooth as possible, I will start with Dr. Iuliia Burdun and Dr. Oleksandr Karasov, who went from being friends to also being collaborators, exactly in this order. I will continue with the cluster of friends I met during

my visit to Ghent University, thank you, Dr Marko Palangetić, Dr Slađana Babić, Dr Camila Olarte Parra, Dr Paweł Morzywołek, Dr Oliver Urs Lenz, Dr Martina Amongero for board games and many pleasant evenings and other moments together. Then, I would like to mention my friends among the Master's studies alumni of the University of Tartu: Artem Domnich, (almost Dr) Marharyta Domnich, Kateryna Peikova, Alina Vorontseva, Dr Vladyslav Fediukov; and almost-graduated students: Viktor Mysko and Anton Potapchuk.

I would further like to express my gratitude to my friends who live in Tallinn, including Mykhailo Yaroshenko, Kateryna Adermann, Dr Andres Kuusk, Taavi Adermann and others.

To conclude, I now turn to the people whose support has been the most fundamental throughout this journey: my family. Thank you to my parents and all my relatives from my home village Hlevakha and Vasylkiv, as well as my wife's family from Kyiv. Thank you, my dear wife Dr Olha Kaminska, for always being by my side and supporting me.

And finally, my gratitude goes to the Armed Forces of Ukraine, and all people of Ukraine and beyond who help to defend my country.

In case I forgot or misspelled anyone, I apologise; you are always welcome to come to me, and I will make a personal correction in the thesis book with a pen and my signature.

SISUKOKKUVÕTE

Sügavõppe treeningkao sobitamine hindamismõõdikutega

Viimaste aastate edusammud masinõppes on kiirendanud masinõppesüsteemide kasutuselevõttu paljudes valdkondades ja tulemused ületavad sageli nii algoritmilisi baastasemeid kui mõnel juhul ka inimsooritust. Selle arengu on võimalikuks teinud paljud vastastikmõjus olevad tegurid, sealhulgas suurte ja üha kvaliteetsemate andmestike kättesaadavus, rikkalikke kõrgemõõtmelisi esitusi võimaldavad tehisnärvivõrkude arhitektuurid ning optimeerimisprotseduurid, mis on praktikas piisavalt laialt kasutatavad. Optimeerimisvalikute hulgas on keskne roll kaofunktsioonil, mida võib vaadelda kui mõõdikut, mis määrab trahvi suuruse ennustava mudeli iga eksimuse korral. Kaofunktsioon defineerib optimeerimismaastiku ja selle abil määratakse, kuidas mudeli kaalusid treenimise käigus muudetakse.

Klassifitseerimisel on kaofunktsiooni vaikevalikuna kasutatud näiteks ristentroopiat. Kui aga suuri mudeleid treenitakse ristentroopiaga, võib mudeli ennustustes ilmneda liigne enesekindlus ja see halvendab mudeli usaldusväärsust nendele ennustustele tuginevates rakendustes. Kuna masinõppesüsteemid on üha enam otsustusprotsessidesse põimitud, on konkreetses rakendusvaldkonnas olulise hindamismõõdiku alusel sooritust parandava kaofunktsiooni valimisest saanud oluline praktiline probleem.

Lisaraskusi tekitab asjaolu, et sooritust hinnatakse lõpuks hindamismõõdikutega, mis võivad olla katkevuspunktidega, mittediferentseeruvad või väga ülesandespetsiifilised. Sellised mõõdikud sageli ei sobi otseselt treenimiseks, mistõttu kasutatakse asenduskaofunktsioone, nagu näiteks ristentroopia. See võib aga tekitada ebakõla treeningul optimeeritava eesmärgi ja lõpprakenduses olulise eesmärgi vahel. Klassikalistest mõõdikutest, nagu täpsus, paljudes tänapäevastes rakendustes ei piisa, mistõttu toetuvad praktikud üha enam hinnatundlikele mõõdikutele, mis kaaluvad vigu nende tähtsuse järgi, ja kalibreerituse mõõdikutele, mis mõõdavad tõenäosuslike ennustuste kvaliteeti. Need mõõdikud kajastavad usaldusväärsuse aspekte, mida täpsus üksi väljendada ei suuda, ja on seetõttu muutunud olulisteks tööriistadeks soorituse hindamisel praktilistes rakendustes.

See väitekiri koosneb kolmest omavahel seotud uurimusest, mille eesmärk on parandada treenimisel kasutatavate kaofunktsioonide ja praktikas oluliste hindamismõõdikute kooskõla. Esimene uurimus käsitleb kaofunktsiooni valikut hinnatundlikus klassifitseerimises, kus hindamismõõdik kaalub iga veatüüpi klassispetsiifiliste kuludega, mille andmed on tavaliselt saadud valdkonna ekspertidelt. Praktikas on need kulud harva täpselt teada ja esinevad pigem ebakindlate hinnangutena, mis võivad aja jooksul ja kasutusolude muutudes teiseneda. Selle ebakindluse modelleerimiseks on uurimuses käsitletud klassispetsiifilisi kulusid etteantud jaotusega juhuslike suurustena ja tuletatud kaofunktsioonide perekonnad, mis on matemaatiliselt samaväärsed vale klassifitseerimise eeldatava kuluga antud mää-

ramatuse korral. Tuvastatud on praktikas mugavad jaotuste perekonnad, näiteks beetajaotused kuluproportsioonide ja gammajaotused toorkulude jaoks, ning näidatud, kuidas nende

parameetrid määravad keskmist kulu, asümmeetriat ja ebakindlust. Katsed mitme andmestiku ja kulustsenaariumiga näitavad, et mõned tuletatud kaofunktsioonid saavutavad vastavatel hinnatundlikel mõõdikutel järjepidevalt tugeva sobivuse, pakkudes põhimõttelist alternatiivi *ad hoc* kaalumisskeemidele.

Teine uurimus keskendub mudeli kalibreeritusele, mis iseloomustab seda, kui hästi vastavad ennustatud tõenäosused tegelikele tinglikele klassitõenäosustele. Mudeli kalibreeritus on hädavajalik mudeli praktilisel rakendamisel, kus ennustatud tõenäosuste põhjal tehakse praktilisi otsuseid. Eriti tähtis on kalibreeritus ohutuse või kulude seisukohalt kriitilistes olukordades. Kuigi ristentroopia on rangelt korralik kaofunktsioon ja teoreetiliselt soodustab kalibreeritud tõenäosusi, on sellega treenitud sügavad tehisnärvivõrgud sageli valesti kalibreeritud. Seevastu fokaalse kaofunktsiooni puhul on korduvalt täheldatud, et see annab paremini kalibreeritud mudeleid isegi ilma hilisema kohandamiseta. Doktoritöös on seda nähtust uuritud ja näidatud, et fokaalset kaofunktsiooni saab väljendada kahe komponendi kompositsioonina: korralik kaofunktsioon, mida minimeerivad tegelikud klassitõenäosused, ja fikseeritud kalibreerimisfunktsioon, mis sarnaneb temperatuuri skaleerimisega. See dekompositsioon selgitab, miks fokaalne kaofunktsioon on praktikas sageli hästi kalibreeritud, kuna see seob korraliku kaofunktsiooni treenimise ajal rakendatava sisseehitatud kalibreerimisteisendusega. Seda dekompositsiooni on laiendatud laiale lahutuvusomadusega kaofunktsioonide klassile ning on tuletatud valemid nendega seotud korralike komponentide ja kalibreerimisfunktsioonide jaoks. Need tulemused võimaldavad disainida uusi kaofunktsioone, millel on soovitud kalibreerimis- ja eristusomadused. Mitu töös tuletatud uut kaofunktsiooni koos nende juurde kuuluvate kalibreerimisfunktsioonidega saavutavad täpsuse ja kalibreerituse poolest tulemusi, mis suudavad konkureerida standardsete baastasemetega või on neist paremad.

Kolmas uurimus käsitleb lõhet treenimisel kasutatavate hindamismõõdikute ja rakendusepõhiste kasufunktsioonide vahel, kus kasufunktsioonid kvantifitseerivad mudeli ennustustel põhinevate otsuste domeenispetsiifilist väärtust. Rakendusepõhised kasufunktsioonid võivad sõltuda kontekstist tulenevatest teguritest, mis pole treenimise ajal kättesaadavad, ja nende hindamine võib olla kulukas, nõudes mõnikord simulatsiooni või füüsilisi eksperimente. Kui rakendusepõhist kasu ei saa suures mahus arvutada, muutub mudelite valimine keeruliseks. Selle probleemi leevendamiseks on töös välja pakutud meetod, milles valideerimisandmetel treenitud väike närvivõrk aitab treenimisel kasutatavat hindamismõõdikut paremini sobitada rakendusepõhise kasufunktsiooniga. Töös on analüüsitud tingimusi, mille korral sellised teisendused säilitavad korralikkuse, ja demonstreeritud meetodi rakenduvust mitmes ülesandes, sealhulgas laoseisu optimeerimise probleemi korral. Selle meetodiga saab rakenduse jaoks hästi sobiva mudeli valida ka olukorras, kus rakendusepõhise kasufunktsiooni otsene hindamine on ebapraktiline.

Kokkuvõttes näitavad selle väitekirja tulemused treenimisel kasutatava kaofunktsiooni ja hindamismõõdiku sobitamise olulisust. Töös on uuritud, kuidas erinevad kaofunktsiooni valikud kujundavad treenitava mudeli ennustuslikku käitumist, ja loodud meetodeid mudelite soorituse parandamiseks rakendusvaldkonnas olulistel mõõdikutel.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Full name:	Viacheslav Komisarenko
Date of birth:	17.12.1995
Nationality:	Ukraine
E-mail:	viacheslav.komisarenko@gmail.com

Education

2019 – present	Ph.D. in Computer Science, University of Tartu
2021 – 2022	Exchange studies in Computer Science, Ghent University
2017 – 2019	MSc in Computer Science, University of Tartu
2013 – 2017	BSc in Computer Science, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Employment

2021 – 2026	Junior Research Fellow in Machine Learning, University of Tartu
2024 – 2026	Data Scientist, Sixfold
2019 – 2020	Data Scientist, Bondora
2018 – 2019	Deep Learning Engineer, KappaZeta

Supervision

2025, MSc Thesis	Karl H. Tamkivi, Viacheslav Komisarenko[1]; Tetiana Shtym; *"Enhancing Mowing Event Detection by Mitigating Semi-Transparent Cloud Anomalies in Optical Satellite Image Time Series"*
2020, MSc Thesis	Kerstin Äkke, Viacheslav Komisarenko[1]; Anti Gruno; *"Snow cover detection in Estonia from SAR images using machine learning"*

Teaching

Spring 2022, 2023	Machine Learning II (teaching assistant)
Fall 2019	Machine Learning (teaching assistant)

[1] Supervisor in charge

ELULOOKIRJELDUS

Isikuandmed

Täisnimi:	Viacheslav Komisarenko
Sünniaeg:	17.12.1995
Kodakondsus:	Ukraina
E-mail:	viacheslav.komisarenko@gmail.com

Haridus

2019 – tänaseni	Ph.D. arvutiteaduses, Tartu Ülikool
2021 – 2022	Vahetusõpingud arvutiteaduses, Genti Ülikool
2017 – 2019	MSc arvutiteaduses, Tartu Ülikool
2013 – 2017	BSc arvutiteaduses, Ukraina Riiklik Tehnikaülikool "Igor Sikorsky Kiievi Polütehniline Instituut"

Teenistuskäik

2021 – 2026	Masinõppe nooremteadur, Tartu Ülikool
2024 – 2026	Andmeteadlane, Sixfold
2019 – 2020	Andmeteadlane, Bondora
2018 – 2019	Süvaõppe insener, KappaZeta

Juhendamine

2025, MSc lõputöö	Karl H. Tamkivi, Viacheslav Komisarenko[1]; Tetiana Shtym; *"Niitmissündmuste tuvastusmudeli täiustamine optiliste satelliidipiltide aegridades olevate poolläbipaistvate pilveanomaaliate tuvastamise kaudu"*
2020, MSc lõputöö	Kerstin Äkke, Viacheslav Komisarenko[1]; Anti Gruno; *"Lume tuvastamine Eestis tehisavardari piltidelt masinõppe meetoditega"*

Õppetöö

Kevad 2022, 2023	Masinõpe II (õppeassistent)
Sügis 2019	Masinõpe (õppeassistent)

[1] Vastutav juhendaja

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.
51. **Ahto Salumets.** Bioinformatics analysis of various aspects in immunology. Tartu 2024, 198 p.
52. **Mohammed Abdulhameed Shaif Ali.** Deep Learning Methods for Cell Microscopy Image Analysis. Tartu 2024, 143 p.
53. **Pille Pullonen-Raudvere**. Foundations of Efficient and Secure Algorithm Development for Secure Multiparty Computation. Tartu 2024, 265 p.
54. **Marili Rõõm.** Multiple approaches to learners' success and factors affecting it in computer programming MOOCs. Tartu 2024, 170 p.
55. **Shivananda Rangappa Poojara.** Design and Orchestration of Scalable, Event-Driven Serverless Data Pipelines for Internet of Things (IoT) Applications. Tartu 2024, 172 p.
56. **Hassan Abdulgaleel Hassan Salim Eldeeb.** Empowering Machine Learning Pipelines with Automated Feature Engineering. Tartu 2024, 121 p.
57. **Muhammad Uzair.** Soft decision making for agri-food 4.0. Tartu 2024, 158 p.
58. **Kirill Milintsevich.** Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity. Tartu 2024, 130 p.
59. **Maksym Del.** Multilingual and Multi-Domain Representational Patterns Across Trpansformer-Based Models. Tartu 2024, 131 p.
60. **Kristo Raun.** Adaptive Out-of-order Handling in Streaming Conformance Checking. Tartu 2024, 118 p.
61. **Toivo Vajakas.** Towards integration of mobile network data into analyzing human mobility. Tartu 2024, 103 p.
62. **Katsiaryna Lashkevich.** Data-Driven Analysis and Optimization of Waiting Times in Business Processes. Tartu 2024, 169 p.
63. **Alejandra Duque-Torres.** Classifying, Constraining and Ranking Metamorphic Relations. Tartu 2025, 159 p.

64. **Mariia Bakhtina.** A Method for Information Security and Privacy Management in Smart Solutions. Tartu 2025, 199 p.
65. **Andre Tättar.** Multilingual Machine Translation for Under-Resourced Languages. Tartu 2025, 170 p.
66. **Mahmoud Shoush.** Prescriptive Process Monitoring Under Uncertainty and Resource Constraints. Tartu 2025, 178 p.
67. **Alireza Akhavi Zadegan.** A Multimodal approach for refining Mapping and Localization by Integrating Generative AI and Pedestrian-Centric Data. Tartu 2025, 147 p.
68. **Eerik Muuli.** Automating the assessment and feedback processes in IT teaching – improving creation and maintenance from the teaching staff perspective. Tartu 2025, 196 p.
69. **Kateryna Kubrak.** Towards User-Centered Prescriptive Process Monitoring Systems. Tartu 2025, 151 p.
70. **Zhigang Yin.** Computing and Sensing in a Smart Ring. Tartu 2025, 251 p.
71. **Abdul-Rasheed Olatunji Ottun.** Practical Trustworthy Artificial Intelligence with Human Oversight. Tartu 2025, 239 p.
72. **Sander Mikelsaar.** Analysis and Optimization of Iteratively Decodable Codes. Tartu 2025, 146 p.
73. **Marharyta Domnich.** Advancing Human-Centric Counterfactual Explanations in Explainable AI. Tartu 2025, 210 p.