Pattern Recognition 178 (2026) 113389

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Privileged learning via a multi-task distilled approach

Mario Martínez-García [a,*], Jon Vadillo [b], Marco Pedersoli [c], Iñaki Inza [b], Jose A. Lozano [a,b]

[a] Basque Center for Applied Mathematics, BCAM, Bilbao, Spain
[b] Computer Science Faculty, University of the Basque Country UPV/EHU, San Sebastián, Spain
[c] Dept. of Systems Engineering, LIVIA, ETS Montreal, Canada

ARTICLE INFO

Keywords:
Learning using privileged information (LUPI)
Multi-task learning
Knowledge distillation
Convolutional neural networks
Neural networks

ABSTRACT

The learning using privileged information paradigm leverages relevant features unavailable at deployment time for model training. In this paper, we propose a multi-task privileged framework that combines two types of tasks. First, the privileged-prediction task involves using regular features (available in both training and deployment) to predict privileged information, working as an intermediate step to guide the learning process. Second, the main learning objective, the target task, uses the predicted privileged information along with the regular features to make the final target prediction. Furthermore, knowledge distillation techniques are included within the target task to enhance the knowledge transfer of privileged information. Experimental results show improvements in tabular datasets and image-related problems compared to state-of-the-art approaches. Additionally, we analyze misclassification causes and refine the proposed multi-task privileged learning to reduce errors.

1. Introduction

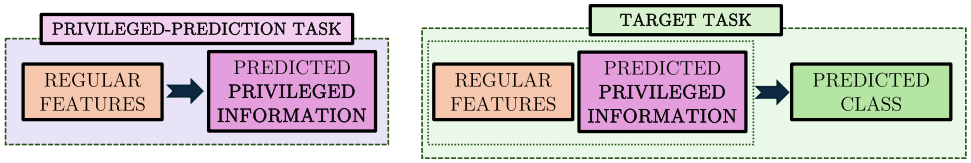
The quantity and quality of data are crucial for obtaining robust and representative machine learning models. However, while nowadays a large amount of data is available for learning models, useful information is sometimes discarded. For instance, high-quality features available for model training but not accessible at deployment time. This type of features are known as privileged information and are considered within the Learning Using Privileged Information (LUPI) paradigm [1]. An example of a privileged scenario can be found in the clinical field [2,3]. Suppose our goal is to automatically classify biopsy images (cancer vs. non-cancer) when the image is provided. Moreover, some days later, the therapist provides a report with detailed information about the biopsy image. This information, known as privileged information, is available for previous patients and can be used for training the models. However, at deployment time, the patient's diagnosis is predicted based exclusively on the biopsy image. Unlike the traditional supervised learning paradigm, which considers the set of pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ for n instances, LUPI takes a set of triplets $\{(\boldsymbol{x}_i, \boldsymbol{x}_i^*, y_i)\}_{i=1}^n$ as inputs composed of regular features $\boldsymbol{x}_i \in \mathbb{R}^l$, privileged features $\boldsymbol{x}_i^* \in \mathbb{R}^m$, and labels y_i.

The LUPI paradigm is related to the 'teacher-student' idea [1]. Suppose a student is preparing an exam with all the theoretical material, which constitutes the regular features. However, during classes, the student can take advantage of the privileged information provided by the teacher, such as the most important points or possible questions that may appear on the final exam. Thus, the teacher helps the student focus on key concepts while preparing for the exam, but their assistance will not be available in the final exam.

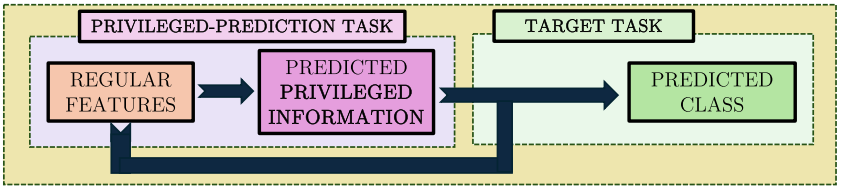
A widely used approach for addressing the LUPI paradigm is focused on predicting privileged information from regular features and leveraging these predictions for training and deployment. This approach, known as Knowledge Transfer [2–4], aims to convey knowledge from the privileged space to the regular space. Specifically, the process is divided into two stages. First, privileged information is predicted from regular features, a step referred to as the privileged-prediction task. Then, the predicted privileged features together with the regular ones are used to predict the labels, which is denoted as the target task. Although the state-of-the-art approach learns these two tasks separately (see Fig. 1a), we propose to address the privileged problem jointly as a multi-task learning [5,6] problem (see Fig. 1b). This approach aims to enhance generalization by leveraging information from complementary tasks, offering potentially better results compared to learning tasks independently. It does this by learning tasks in parallel; what is learned for each task can help other tasks to be learned better [6]. Therefore, by training on multiple tasks simultaneously, the model tends to learn more general representations rather than memorizing the peculiarities of a single task. Moreover, this reduces the risk of overfitting by preventing the model from becoming too specialized on a single task [7].

* Corresponding author.

https://doi.org/10.1016/j.patcog.2026.113389
Received 7 May 2025; Received in revised form 26 January 2026; Accepted 25 February 2026
Available online 1 March 2026



(a) Knowledge Transfer



(b) Multi-task privileged learning

Fig. 1. (a) Knowledge transfer: Privileged-prediction and target tasks are learned separately. (b) Multi-task privileged learning: Privileged-prediction and target tasks are learned jointly.

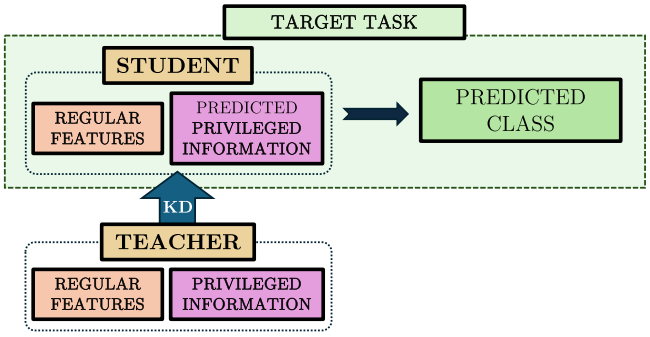


Fig. 2. Knowledge distillation scheme introduced in the target task.

This effect is further supported by the sharing of common representations, which allows the model to exploit relationships between tasks [6]. Multi-task learning benefits from related tasks, such as the privileged-prediction and the target task. Thereby, we have considered multi-task framework within the privileged setting to enhance model performance.

In addition to the advantages that multi-task learning can provide in the privileged paradigm, we find other effective approaches for handling the LUPI paradigm that should be considered. Specifically, we leverage knowledge distillation [8,9], which shares the privileged idea of the "teacher-student". Knowledge distillation's objective is to perform a proper transfer of knowledge from a teacher learned with privileged information (either independently [9] or alongside regular features [10,11]) to a student learned with only regular features. Owing to the advantages that knowledge distillation can provide to the privileged learning, we embed knowledge distillation within the multi-task privileged learning. It should be noted that both techniques are complementary and can be used within the same learning framework. While multi-task learning enables joint learning from both the privileged-prediction task and the target task, knowledge distillation allows the transfer of high-quality information from a teacher learned with the real privileged information. This integration allows privileged information to be leveraged from two complementary perspectives: the teacher model used in knowledge distillation and the privileged-prediction task of the multi-task learning framework. It is worth noting that the distillation process influences the target task but the privileged-prediction task remains unmodified (see Fig. 2). Furthermore, contrary to state-of-the-art privileged distillation approaches [8,10,11], the student not only has access to the regular features but can also rely on the privileged predictions. Thus, the multi-task privileged learning is addressed through two privileged distillation methods: the traditional Privileged Feature Distillation (PFD) [11], and Teacher Privileged Distillation (TPD) [10], focused on dealing with imperfect teachers.

In this work, the advantages of multi-task learning and knowledge distillation are combined to improve the LUPI paradigm. Multi-task learning allows a better generalization [5,6], while distillation helps to transfer specific information between the teacher and the student [8,9]. To the best of our knowledge, this combination represents a novel approach that integrates the strengths of both methods within a single model. Thus, the main contributions of this work are the following:

- A Multi-Task Privileged framework is introduced (MTP): LUPI with knowledge transfer is addressed from a multi-task perspective. Unlike traditional approaches that treat the privileged-prediction task and the target task as a sequential process, our proposed MTP framework adopts a joint learning strategy. This simultaneous optimization fosters richer interactions and a more effective exploitation of privileged information.
- An improvement of the proposed multi-task privileged (MTP) framework with knowledge distillation techniques is developed to effectively leverage the privileged information. Specifically, a multi-task privileged with PFD [11] (MTP-PFD) and TPD [10] (MTP-TPD) are presented.
- An analysis to determine the causes of misclassification in the evaluated instances is developed: *Does the error come from a weak prediction of the privileged feature, or is it due to a teacher failure?* Furthermore, a set of insights are presented to adjust the proposed multi-task privileged learning approach in an attempt to correct misclassified instances.

The rest of the paper is organized as follows. Section 2 covers the related work. Section 3 compiles our proposal, the multi-task privileged learning approaches. Section 4 presents the experimental results for some tabular and image datasets, while Section 5 provides a discussion to further explore and understand the approaches. Finally, Section 6 concludes the manuscript.

2. Related work

Three paradigms are essential to the development of this work: learning using privileged information, multi-task learning, and knowledge distillation.

2.1. Learning using privileged information

The LUPI paradigm was initially proposed by Vapnik and Vashist through Support Vector Machines (SVMs) [1]. The original idea was refined with more efficient models [12] and extended to other learning algorithms. For instance, some privileged approaches for logistic regression [13], neural networks [14,15], and tree-based learning [16] were proposed.

A relevant approach within LUPI, knowledge transfer (generally described in [17]), was introduced by Vapnik and Izmailov [2,3]. It seeks to transfer information from the privileged space to the final regular space. Based on this, numerous studies have emerged related to feature selection [4], neural networks [18], and explainable machine learning [19].

Moreover, various fields such as fairness [20], streaming [21], multi-view learning [22], and multi-label [23] have considered privileged information in their learning processes. Even in scenarios without privileged information, its extraction has been proposed through explainability models [24]. Recently, there has been growing interest in using privileged information to deal with label noise [25–27].

2.2. Knowledge distillation

A key point within LUPI is its relationship with knowledge distillation, which is defined as a type of model compression and acceleration: a smaller student model effectively learns from a larger teacher model. Lopez Paz et al. [9] unified the LUPI paradigm with the knowledge distillation and named it Generalized Distillation (GD). In this case, the knowledge from an intelligent teacher learned in the privileged space is transferred to a student in the final regular space. The privileged distillation process begins by learning from the set $\{(\boldsymbol{x}_i^*, y_i)\}_{i=1}^n$ an intelligent teacher $f_t \in F_t$, where F_t is a class of teacher functions. Thus, the student $f_s \in F_s$ captures the knowledge distilled from the teacher $f_t \in F_t$ by minimizing:

$$f_s = \underset{f \in F_s}{argmin} \frac{1}{n} \sum_{i=1}^{n} \left[(1-\lambda) \cdot \ell(y_i, \sigma(f(\boldsymbol{x}_i))) + \lambda \cdot \ell(s_i, \sigma(f(\boldsymbol{x}_i))) \right] \tag{1}$$

where σ represents the softmax operator, $\ell(\cdot)$ the loss function (specifically the cross-entropy) and λ the imitation parameter. Moreover, s_i term reports the soft predictions $s_i = \sigma(f_t(\boldsymbol{x}_i^*)/T)$ where the temperature $T > 0$ controls the smoothing degree of predicted probabilities.

Xu et al. [11] introduced Privileged Feature Distillation (PFD), which operates similarly to GD (see Eq. (1)), but with a teacher model trained using both regular and privileged features $(\boldsymbol{x}_i, \boldsymbol{x}_i^*, y_i)_{i=1}^n$, thereby the soft predictions of Eq. (1) are redefined as $s_i = \sigma(f_t(\boldsymbol{x}_i^*, \boldsymbol{x}_i)/T)$. This results in a more knowledgeable teacher, providing better guidance for the student model. Later, Yang et al. [28] examined the conditions required for the robust performance of PFD through both empirical studies and theoretical analysis for linear models.

Alternatively, Teacher Privileged Distillation (TPD) [10] was developed to deal with imperfect teachers, i.e., teachers who make mistakes. Specifically, the teacher is learned as in PFD, using both regular and privileged features $(\boldsymbol{x}_i, \boldsymbol{x}_i^*, y_i)_{i=1}^n$. Thereby, following the notation of Eq. (1) the knowledge is transferred from the teacher $f_t \in F_t$ to the student $f_s \in F_s$ by optimizing:

$$\begin{aligned} f_s = \underset{f \in F_s}{argmin} \frac{1}{n} \sum_{i=1}^{n} \Big[& (1-\lambda) \cdot \ell(y_i, \sigma(f(\boldsymbol{x}_i))) \\ & + \lambda \cdot \left(\delta_{y_i \hat{y}_i^t} \cdot \ell(\sigma(f(\boldsymbol{x}_i)), s_i) - (1 - \delta_{y_i \hat{y}_i^t}) \cdot \beta \cdot \ell(\sigma(f(\boldsymbol{x}_i)), s_i) \right) \Big] \end{aligned} \tag{2}$$

where λ is the imitation parameter, β parameter controls the influence of misclassified instances and $\delta_{y_i \hat{y}_i^t}$ is the Kronecker delta function: it is 1 when the real label and teacher's prediction are equal, and 0 when they diverge. Thus, $\delta_{y_i \hat{y}_i^t}$ handles a correction of the misclassifications made by the teacher. Furthermore, it should be noted that the pure distillation ($\lambda = 1$) term is composed of a cross-entropy loss where the inputs order are swapped to appropriately penalize the interaction between the student and the teacher. Although the teacher's output can be seen as a form of label noise, important differences arise in the training process compared to standard label-noise approaches [26]. In this case, TPD explicitly incorporates the clean labels, whereas in label noise approaches the clean labels are only used for evaluation.

2.3. Multi-task learning

Multi-task learning [5,6] aims to train multiple tasks simultaneously, enhancing both efficiency and performance compared to training separate models for each task. Thus, sharing information across complementary tasks can give better generalization than learning the tasks in isolation. Note that, multi-task learning is addressed from the privileged perspective in different ways. For instance, Tang et al. [29] propose a multi-task privileged approach where the elements of the regular and privileged spaces are projected into a joint space to inform the similarity between training samples. Multi-objective optimization is also used to arbitrate multi-task learning between the target task and the privileged task [30]. Furthermore, the multi-task privileged approach has also been addressed through Generative Adversarial Networks, where a deep multi-task network acts as a generator to identify and distinguish the predicted privileged information [31]. While our approach focuses on a supervised scenario, semi-supervised alternatives based on a privileged multi-task setting are also developed. Specifically, Liu et al. [32] use the privileged information to mitigate the impact of unlabeled samples. Recently, a multi-task mutualistic transformer [33] was proposed for gaze prediction, where privileged information from one task is leveraged via a mutualistic attention mechanism to guide another task, improving feature learning and inter-task communication.

In this paper, unlike the joint space approach in [29] and the GAN-based alternative in [31], we share the objectives of the multi-objective framework proposed in [30], addressing both the privileged-prediction and target tasks of the LUPI paradigm. While our method shares similarities with Kendall et al. [34] in weighting the loss functions according to the homoscedastic uncertainty of each task, it differs in the task architecture. Our multi-task privileged framework follows a sequential architecture, known as the multi-task cascade [35], in which the target task explicitly depends on the output of the privileged-prediction task. This design allows for more effective knowledge transfer from the privileged-prediction task to the target task, improving learning efficiency and model performance compared to conventional multi-task approaches. Furthermore, when incorporating knowledge distillation, we train a model that exploits privileged information from two complementary sources –the teacher from the knowledge distillation and the privileged– prediction task of the multi-task learning framework- representing a novel integration of these approaches within a unified model.

3. Multi-task privileged learning

The LUPI paradigm has the same objective as standard supervised learning: minimizing the error in target prediction. However, when LUPI paradigm is addressed from the knowledge transfer perspective additional tasks are leveraged. First, the prediction of privileged information, i.e., the privileged-prediction task, serves as an intermediate stage to guide the learning process. Second, the target prediction task is addressed. Since multiple tasks are involved, we propose to handle knowledge transfer from a multi-task perspective, where both the privileged-prediction and target tasks are learned jointly. Furthermore, multi-task privileged learning is defined as a multi-task cascade [35] problem: the privileged-prediction task influences the target task, i.e., the predicted privileged information acts as an additional input for the target task. Note that, in the proposed framework, privileged information is assumed

to be numeric so the privileged-prediction task is formulated as a regression task. In contrast, for the target task, a classification scenario is chosen. Thus, assuming m privileged features and n instances, the Multi-task Privileged learning (MTP) is defined as follows:

$$\underset{\substack{f_r\in F_R,\, f_c\in F_C,\\ \theta_j,\tau}}{argmin} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{1}{2\theta_j^2}\ell_R(x^*_{i,j}, f_r(\boldsymbol{x}_i)_j) + \log(\theta_j)\right) + \frac{1}{\tau^2}\ell_C(y_i, \sigma(f_c(\boldsymbol{x}_i, f_r(\boldsymbol{x}_i)))) + \log(\tau) \tag{3}$$

where F_R and F_C represent classes of regression and classification functions, respectively. $\ell_R(\cdot)$ computes the squared difference between the real privileged feature value and the predicted one, and $\ell_C(\cdot)$ evaluates the cross-entropy loss for the classification task. Furthermore, θ_j and τ are used to weight the $\ell_R(\cdot)$ loss of each j privileged feature ($j = 1, \dots, m$) and $\ell_C(\cdot)$ loss, respectively. Logarithms act as regularizers to prevent the uncontrolled growth of θ_j and τ. Note that, as [34] defines, θ_j and τ capture the trade-off between tasks, which are computed based on the uncertainty inherent to the regression and classification tasks, respectively. In this way, instead of relying on computationally expensive hyperparameter tuning, the loss weights are computed during model training.

Furthermore, in order to extract more information from the privileged features, knowledge distillation techniques are included within the target task of the multi-task privileged learning. Thus, instead of using the traditional cross-entropy loss, two distillation methods are considered: PFD [11], and TPD [10]. Thereby, Eq. (3) is rewritten with PFD as follows:

$$\underset{\substack{f_r\in F_R,\, f_c\in F_C,\\ \theta_j,\tau}}{arg\,min} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{1}{2\theta_j^2}\ell_R(x^*_{i,j}, f_r(\boldsymbol{x}_i)_j) + \log(\theta_j)\right) + \frac{1}{\tau^2}\Big[(1-\lambda)\cdot\ell_C(y_i, \sigma(f_c(\boldsymbol{x}_i, f_r(\boldsymbol{x}_i)))) + \lambda\cdot\ell_C(s_i, \sigma(f_c(\boldsymbol{x}_i, f_r(\boldsymbol{x}_i))))\Big] + \log(\tau) \tag{4}$$

where the loss associated to the classification task is changed with the distillation structure of the PFD defined in Eq. (1). Thus, Eq. (4) formulates the Multi-task Privileged learning with PFD (MTP-PFD). Alternatively, when TPD is considered, Eq. (3) is modified to obtain the Multi-task Privileged learning with TPD (MTP-TPD):

$$\underset{\substack{f_r\in F_R,\, f_c\in F_C,\\ \theta_j,\tau}}{argmin} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{1}{2\theta_j^2}\ell_R(x^*_{i,j}, f_r(\boldsymbol{x}_i)_j) + \log(\theta_j)\right) + \frac{1}{\tau^2}\Bigg[(1-\lambda)\cdot\ell_C(y_i, \sigma(f_c(\boldsymbol{x}_i, f_r(\boldsymbol{x}_i)))) + \lambda\cdot\Big(\delta_{y_i\hat{y}_i^t}\cdot\ell_C(\sigma(f_c(\boldsymbol{x}_i, f_r(\boldsymbol{x}_i))), s_i) - (1-\delta_{y_i\hat{y}_i^t})\cdot\beta\cdot\ell_C(\sigma(f_c(\boldsymbol{x}_i, f_r(\boldsymbol{x}_i))), s_i)\Big)\Bigg] + \log(\tau) \tag{5}$$

where each term of TPD is defined in Eq. (2). Note that, by incorporating distillation within multi-task learning, additional insights are collected by the student. More precisely, with this approach, the student not only has access to the regular features but also to the predicted privileged information (see Fig. 2). Consequently, we hypothesize that the distillation performed within multi-task learning may outperform traditional privileged distillation, as we experimentally validate in Section 4.

4. Experimental evaluation

In this section, the three presented multi-task privileged frameworks are evaluated: multi-task privileged (MTP), multi-task privileged with PFD (MTP-PFD), and multi-task privileged with TPD (MTP-TPD). In our experiments, the regression (f_r) and classification (f_c) models are implemented using neural networks, with parameters ω_r and ω_c, respectively. It is worth noting that the gradient backpropagation is carried out jointly, so the gradients affecting the classification network also flow into the regression network due to their sequential connection, as we outline in Fig. 3. More architectural details will be provided in Sections 4.1 and 4.2.

The evaluation of the approaches is conducted using an imitation parameter value of $\lambda = 0.5$, a temperature of $T = 1$ and $\beta = 1$ (only with TPD) in the distillation models. We compare the proposed multi-task privileged frameworks with traditional Knowledge Transfer (KT), where regression and classification tasks are learned separately, as well as with standard privileged distillation methods, PFD [11] and TPD [10]. Moreover, we also consider two additional alternatives for KT with distillation in the classification task: using PFD (KT-PFD) and TPD (KT-TPD). In addition to the privileged models, we also report the performance of two baseline classification models: the teacher model trained with both privileged and regular features, and the regular model trained exclusively with regular features. The results are presented in an ablation format, allowing a straightforward comparison of the impact of privileged information across each approach: knowledge transfer, knowledge distillation, and multi-task learning. Furthermore, the evaluation of these methods is conducted on eight binary datasets and on the EuroSat image dataset [36].

We choose the error rate and the LUPI gain [3] as comparative metrics. The latter, a state-of-the-art metric in the privileged paradigm, quantifies the improvement of the privileged model by measuring the gained performance proportion between the teacher and the regular model. Thus, let us suppose the error rate of the teacher and the regular model are C and B, respectively. Therefore, the objective of a LUPI process is to achieve the lowest possible error rate, denoted as A, which is expected to be within the range $C \leq A < B$. LUPI gain is computed as *(B - A)/(B - C)*, measuring the proportion of the performance gap *(B - C)* that can be recovered by a privileged model. It is important to emphasize that the performance of the privileged model may not always conform to this expectation $C \leq A < B$. In some cases, the privileged model may achieve better generalization, resulting in superior performance compared to the teacher model ($A < C$). This scenario tends to arise more frequently when the performance gap *(B - C)* is small. Although this is a very uncommon outcome, it can occur not only in our models but also in other LUPI frameworks [10,13,16]. The code is available on GitHub.[1]

4.1. Experiments on tabular datasets

The proposed multi-task privileged learning is evaluated in 8 different binary datasets collected from the UCI [37] and OpenML [38]. These datasets (see Table 1) vary in the number of features (4–15), sample sizes (768–15545), and positive class proportions (34.9%–70.7%). It should be noted that the availability of datasets with real privileged information in the literature is limited. For this reason, most state-of-the-art LUPI studies use standard datasets where some features are chosen as privileged using dependency measures [2–4,10,13,18]. Following this approach, in our learning framework the feature with the highest correlation to the target is selected as privileged. Specifically, the absolute Spearman rank correlation between each predictor feature and the target is used [10,13]. This procedure typically results in a performance gap between the teacher and the regular models, providing a useful basis for evaluating LUPI methods.

The privileged distillation models (TPD, PFD), the teacher model, and the regular model are all implemented using a Multilayer Perceptron (MLP) with two hidden layers of dimension 20. However, two MLPs with the same architecture are employed for knowledge transfer models (KT, KT-PFD, KT-TPD) and multi-task privileged models (MTP, MTP-PFD, MTP-TPD): one is assigned to the regression task and the other is tailored for the classification task. Every model is trained for 1000 epochs

[1] https://github.com/mariomartgarcia/privileged_multitask

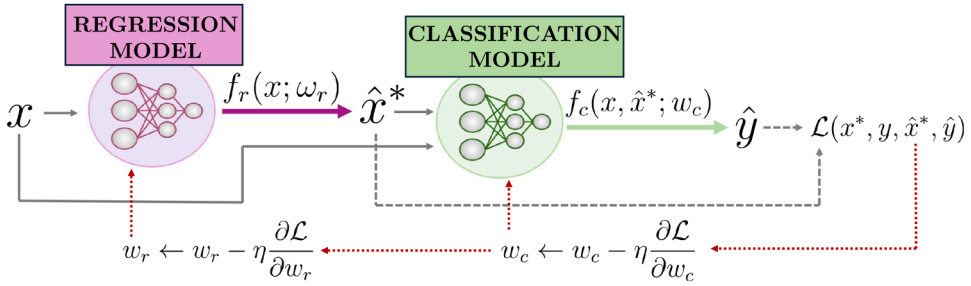


Fig. 3. Multi-task cascade diagram illustrating the privileged-learning scenario. The parameters, ω_r and ω_c, are learned to minimize a loss function $\mathcal{L}$, specified in Eq. (3) for MTP, in Eq. (4) for MTP-PFD, and in Eq. (5) for MTP-TPD. Note that, red dotted lines illustrate the path followed during gradient backpropagation for a specific learning rate η. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Name of the *Privileged Feature*, number of regular features (*#R*), number of samples (*#S*) and positive class percentage for tabular datasets.

Dataset	*Privileged Feature*	*#R*	*#S*	Positive Class
phishing [37]	empty_server_form_handler	8	1250	43.8%
obesity [37]	Weight	15	2111	40.2%
diabetes [38]	Glucose	7	768	34.9%
phoneme [38]	V4	4	5404	70.7%
mozilla [38]	start	4	15,545	67.1%
wine [37]	alcohol	10	6497	63.3%
abalone [38]	Shell weight	7	4177	50.2%
wind [38]	CLO	13	6574	53.3%

with Adam optimizer, batch size of 128, validation split of 0.2, and early stopping with a patience of 5 epochs to avoid overfitting. These training hyperparameters have been selected empirically, ensuring convergence across all datasets. Furthermore, evaluation metrics have been calculated based on the average of 50 repetitions of a 5-fold stratified cross-validation (the mean is reported along with a 95% confidence interval).

Table 2 collects the results for the 8 datasets considered in the study; as can be seen, multi-task privileged models achieve the most effective transfer of privileged information (more evaluation metrics are reported in Appendix A). Moreover, when knowledge transfer models (KT, KT-PFD, KT-TPD) are compared with multi-task approaches (MTP, MTP-PFD, MTP-TPD), the results clearly indicate that joint learning of the privileged-prediction and target tasks enhances model generalization, resulting in better performance across all datasets. Except for MTP and MTP-TPD in *wind* dataset, equal or superior performance is reported in all three multi-task privileged approaches. Additionally, distillation methods (PFD, TPD) show improvement when leveraged from the knowledge transfer perspective, with even more substantial gains when addressed from a multi-task approach. Thus, these results show that a student with access to the predicted privileged features attains a better distillation of the teacher's information.

Although the three multi-task proposals yield fairly similar results, some differences can be observed. In particular, at least one of the two distillation-based alternatives (MTP-PFD, MTP-TPD) consistently outperforms the MTP model on each evaluated dataset. Consequently, the complementary use of multi-task learning and knowledge distillation contributes significantly to achieving an effective use of the privileged information. Note that an additional comparison with LUPI models, further evaluation metrics, a sensitivity analysis of λ and T, a loss weighting ablation study, and the computational complexity are included in Appendix A.

4.2. Experiments on EuroSat image dataset

In order to show the applicability of our approaches for vision tasks, multi-task privileged learning frameworks are evaluated in an image classification scenario, the EuroSat Dataset [36]. This dataset consists of images captured by the Sentinel-2 satellite with 2500 samples per class. Each image is composed of 12 bands, all with 64x64 dimensions. In our experiment we use the common RGB bands and more specific ones such as NIR (Neared-Infrared), SWIR-C (Short-Wave Infrared-Cirrius) and SWIR (Short-Wave Infrared) with central wavelengths of 0.842 μm, 1.375 μm and 2.190 μm, respectively. Note that, infrared bands are selected as privileged information as they are comparatively less common than RGB bands and can provide relevant information to discriminate between satellite images. Thereby, three sets of experiments are conducted: the RGB-NIR, RGB-SWIRC and RGB-SWIR pairs where the regular features are the RGB bands, and the infrared band (either NIR, SWIRC or SWIR) is the privileged information. From the different available labels in the dataset, a binary classification problem is generated: *Highway vs. River* (see Fig. 4). The selected binary problem is chosen because the privileged information offers valuable insights into the learning process. Note that privileged information does not always contribute significantly (no GAP performance between the teacher and the student model is obtained) in the image classification problem for EuroSat Dataset, as classes can often be easily distinguished using only RGB images.

The network architecture used to model the EuroSat dataset shown in Fig. 5 differs from the one used with tabular domains. First, the regressor uses a reduced version of the U-Net architecture [39], chosen for its ability to effectively capture global context through its contracting path while enabling precise localization via its symmetric expanding path. This makes U-Net particularly suitable for recovering the infrared band from the RGB bands. Thereby, the contracting path of the U-Net extracts features using 3x3 convolutional layers, reducing the spatial dimensions to 32x32 and then to 16x16 using max-pooling. The expansive path gradually restores the original 64x64 size using transposed convolutions. Second, the classifier starts with a 3x3 convolutional layer to extract features from the input image, followed by a max-pooling layer and another 3x3 convolutional layer. The global average pooling layer simplifies the data, and two consecutive layers of 64 and 32 dimensions, respectively, are the final steps before extracting the predicted class. It is important to note that the predicted infrared image is added as an extra band to the RGB image. Thus, in knowledge transfer (KT, KT-PFD, KT-TPD) and multi-task (MTP, MTP-PFD, MTP-TPD) approaches, the classifier input has 4 bands: RGB and the predicted infrared. The model training is undertaken for 500 epochs with Adam optimizer, batch size of 128, and patience of 5 epochs. Since the EuroSat dataset is composed of multi-band images, which are much heavier compared to tabular data, we follow the standard evaluation in deep learning scenarios: the hold-out strategy. This enables a computationally cheaper and representative evaluation. For each class, 80% of the data is used for training and 20% for evaluation. Specifically, a 20-times repeated hold-out strategy is applied within the training set, where in each iteration, 20% of the training data is randomly selected for validation.

The results (see Table 3 and Appendix A) show that multi-task privileged approaches (MTP, MTP-PFD, MTP-TPD) achieve better perfor-

Table 2
Error rates and LUPI gain on different classification domains with tabular datasets. The results of a model learned with regular features (*Regular*), and a teacher learned with privileged and regular features (*Teacher*) are shown as baseline. Privileged distillation (*PFD*, *TPD*), Knowledge Transfer (*KT*, *KT-PFD*, *KT-TPD*) and multi-task privileged (*MTP*, *MTP-PFD*, *MTP-TPD*) models are also reported. For privileged models, the LUPI gain is included between parentheses (a minus reports a non-representative LUPI gain), and the best model is highlighted in bold.

Method	phishing	obesity	diabetes	phoneme
Regular	0.149±0.002	0.114±0.001	0.302±0.004	0.179±0.002
Teacher	0.088±0.003	0.023±0.002	0.239±0.004	0.164±0.001
PFD	0.148±0.003 (1.6%)	0.114±0.002 (0.0%)	0.301±0.004 (1.6%)	0.178±0.001 (6.7%)
TPD	0.149±0.003 (0.0%)	0.113±0.002 (1.1%)	0.300±0.004 (3.2%)	0.181±0.001 (–)
KT	0.146±0.003 (4.9%)	0.113±0.002 (1.1%)	0.299±0.004 (4.8%)	0.173±0.001 (40.0%)
KT-PFD	0.146±0.003 (4.9%)	0.114±0.002 (0.0%)	0.300±0.004 (3.2%)	0.173±0.001 (40.0%)
KT-TPD	0.145±0.003 (6.6%)	0.111±0.002 (3.3%)	0.300±0.004 (3.2%)	0.176±0.001 (20.0%)
MTP	0.143±0.003 (9.8%)	0.106±0.002 (8.8%)	0.299±0.004 (4.8%)	0.168±0.001 (73.3%)
MTP-PFD	**0.141±0.003 (13.1%)**	0.105±0.002 (9.9%)	0.299±0.004 (4.8%)	0.170±0.002 (60.0%)
MTP-TPD	**0.141±0.002 (13.1%)**	**0.104±0.002 (11.0%)**	**0.298±0.004 (6.3%)**	**0.167±0.001 (80.0%)**
Method	mozilla	wine	abalone	wind
Regular	0.248±0.001	0.241±0.001	0.207±0.001	0.150±0.001
Teacher	0.098±0.001	0.235±0.001	0.202±0.002	0.135±0.001
PFD	0.246±0.001 (1.3%)	0.240±0.001 (16.7%)	0.204±0.001 (60.0%)	0.149±0.001 (6.7%)
TPD	0.247±0.001 (0.7%)	0.241±0.001 (0.0%)	0.205±0.001 (40.0%)	0.149±0.001 (6.7%)
KT	0.245±0.001 (2.0%)	0.239±0.001 (33.3%)	0.206±0.001 (20.0%)	0.150±0.001 (0%)
KT-PFD	0.244±0.001 (2.7%)	0.239±0.001 (33.3%)	0.204±0.001 (60.0%)	0.149±0.001 (6.7%)
KT-TPD	0.245±0.001 (2.0%)	0.240±0.001 (16.7%)	0.204±0.001 (60.0%)	0.149±0.001 (6.7%)
MTP	0.244±0.001 (2.7%)	0.237±0.001 (66.7%)	0.203±0.001 (80.0%)	0.150±0.001 (0%)
MTP-PFD	**0.243±0.001 (3.3%)**	**0.236±0.001 (83.3%)**	**0.201±0.001 (120.0%)**	**0.148±0.001 (13.3%)**
MTP-TPD	0.245±0.001 (2.0%)	**0.236±0.001 (83.3%)**	0.203±0.002 (80.0%)	0.150±0.001 (0%)

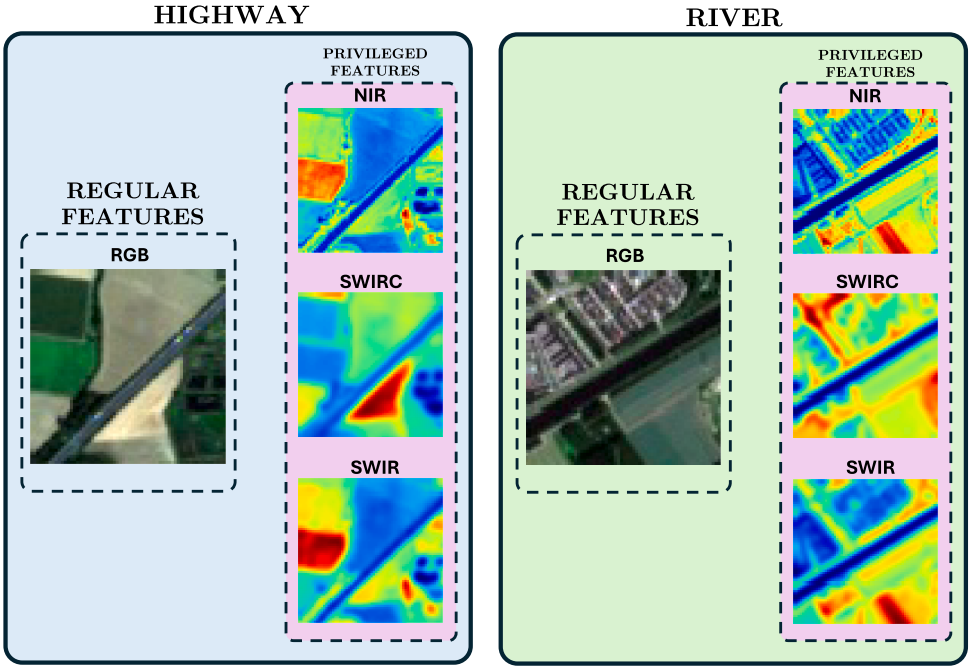


Fig. 4. Binary classification problem selected from the EuroSat dataset: Highway vs. River. The regular features are the RGB images, while the privileged feature is one of the three infrared bands: either NIR, SWIRC, or SWIR.

mance in all evaluated experiments. Thereby, the inclusion of the predicted infrared image notably improves the generalization of the models. Furthermore, the same pattern observed with tabular datasets is repeated: KT and MTP approaches outperform traditional distillation models. Among the multi-task models, the PFD-based approach (MTP-PFD) reports the highest performance, though closely followed by MTP and MTP-TPD. Although all three infrared bands provide relevant information for the classification process, the best result is attained with the SWIR band. Additional results for a EuroSat multiclass classification problem are presented in Appendix B.

5. Discussion

In this section, we present a discussion intended to deepen the understanding of the behavior of the privileged model. First, we analyze the factors that lead to its misclassifications. Subsequently, we explore

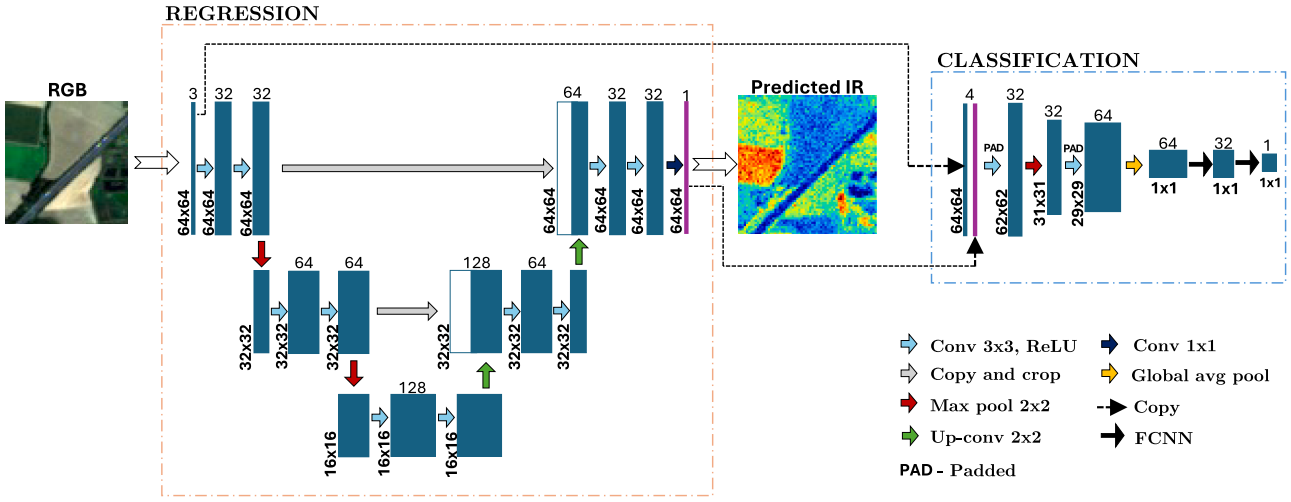


Fig. 5. Network architecture for EuroSat dataset. The regression part corresponds to a U-Net and the classification part contains convolutional and fully connected networks. Each blue box depicts a multi-channel feature map. The number of channels is denoted on top of the box. White boxes represent copied feature maps. The arrows denote the different operations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Error rates and LUPI gain for the binary problem Highway vs. River of EuroSat Dataset. The results of a model learned with regular features (*Regular*), and a teacher learned with privileged and regular features (*Teacher*) are shown. Privileged distillation (*PFD*, *TPD*), Knowledge Transfer (*KT*, *KT-PFD*, *KT-TPD*) and multi-task privileged (*MTP*, *MTP-PFD*, *MTP-TPD*) models are also reported. For privileged models, the LUPI gain is included between parentheses (a minus reports a non-representative LUPI gain), and the best model is highlighted in bold.

Method	NIR	SWIRC	SWIR
Regular	0.120±0.003	0.120±0.003	0.117±0.004
Teacher	0.039±0.001	0.050±0.001	0.044±0.002
PFD	0.115±0.003 (6.2%)	0.116±0.003 (5.7%)	0.119±0.003 (–)
TPD	0.126±0.003 (–)	0.130±0.005 (–)	0.131±0.003 (–)
KT	0.060±0.001 (74.1%)	0.065±0.001 (78.6%)	0.065±0.001 (71.2%)
KT-PFD	0.064±0.001 (69.1%)	0.073±0.002 (67.1%)	0.066±0.002 (69.9%)
KT-TPD	0.065±0.001 (67.9%)	0.069±0.002 (72.9%)	0.066±0.002 (69.9%)
MTP	0.054±0.003 (81.5%)	**0.051±0.001 (98.6%)**	0.054±0.004 (86.3%)
MTP-PFD	**0.048±0.002 (88.9%)**	**0.051±0.002 (98.6%)**	**0.046±0.002 (97.3%)**
MTP-TPD	0.050±0.003 (86.4%)	0.053±0.003 (95.7%)	0.068±0.008 (67.1%)

potential strategies to correct the instances that the privileged model misclassifies.

5.1. What causes the misclassification of the privileged model?

In order to provide insights into the performance of the multi-task privileged models, the behavior of the instances is examined. Specifically, the possible reasons for failure in the final classification are studied: *Is the error caused by a weak prediction of the privileged feature, or is it due to a failure of the teacher model?* Note that, in privileged models with distillation, it is possible to extract insights from each instance through comparison between the privileged model (P), the teacher (T), and the real labels (R). Hence, the following combinations can happen:

- P = T = R. The privileged model achieves the same prediction as the teacher and the real label. Thus, the regressor predicts the privileged feature adequately, leading to a correct classification.
- P ≠ T = R. The privileged model differs from both the teacher and the real label. Specifically, the error of the privileged model is caused by a weak prediction of the privileged feature.
- P = T ≠ R. The privileged model and the teacher make the same prediction, but it differs from the real label. An accurate prediction of the privileged feature leads to the same prediction as the teacher. However, this instance is so challenging to classify that even the teacher is unable to identify it correctly.
- P = R ≠ T. The teacher fails, but the privileged model is able to attain the correct prediction. Particularly, a double error can produce a correct classification. Hence, when the prediction of the privileged feature deviates significantly from its true value, it may lead to the prediction of a different class than the teacher, i.e., the correct class.

Furthermore, in order to analyze why errors are committed, two novel metrics are proposed. These metrics are studied in problems with a single privileged feature. First, the Deviation from Real (*DR*) metric is defined as the difference between the predicted value of the privileged feature ($\hat{x}_i^*$) and its real value (x_i^*) for each instance:

$$DR = \hat{x}_i^* - x_i^* \tag{6}$$

This metric can be either positive or negative and indicates how far, and in which direction, the predicted privileged value is from its real value. Second, the Distance to Change (*DC*) metric is defined as the difference between the real privileged value and the closest privileged value where a class change occurs in the teacher model. Particularly, we compute the class prediction by sweeping over all possible values of the privileged information, while keeping the regular features fixed. Thus, we define p_i^* as the nearest privileged value where the class change occurs in comparison to the real privileged value:

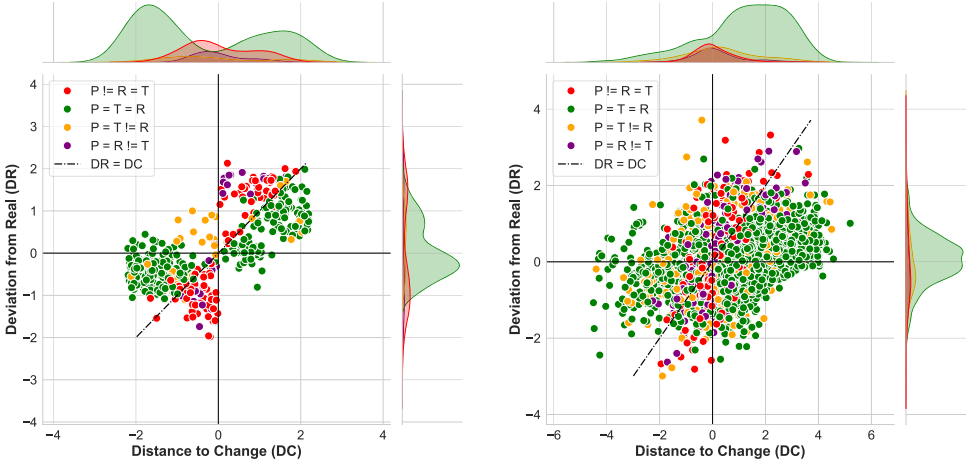
$$p_i^* = \underset{|p|}{argmin} \quad \sigma(f_{ct}(\boldsymbol{x}_i, x_i^*)) \; ! = \sigma(f_{ct}(\boldsymbol{x}_i, x_i^* + p)) \tag{7}$$

where σ represents the softmax operator and f_{ct} the teacher model. Note that this metric is computed using the teacher model, as it represents the ideal model we seek to attain. Hence, the *DC* for each instance is introduced as follows:

$$DC = p_i^* - x_i^* \tag{8}$$

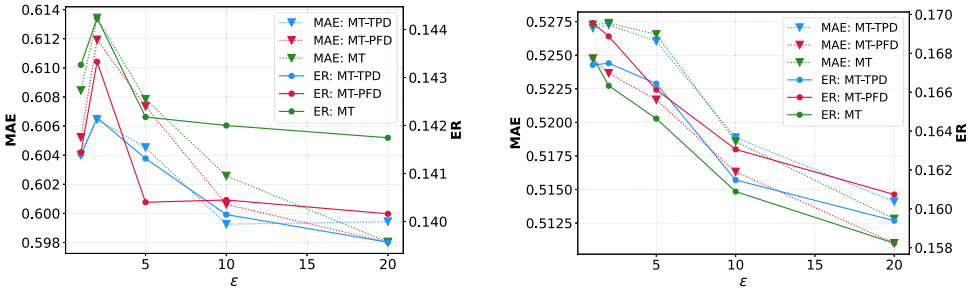
Based on these metrics, each instance can be plotted on a 2D-graph, where the axes correspond to *DR* and *DC*. Specifically, the instances evaluated with MTP-TPD are shown for *phishing* (see Fig. 6a) and *phoneme* (see Fig. 6b) datasets (remaining plots are reported in Appendix C). Furthermore, the relationship of both metrics with respect to the teacher's prediction and the real labels is indicated with different colors.

The results correspond to all the instances evaluated in each of the 5 testing-folds of a cross-validation. As depicted in the graphs, the majority of red instances (P ≠ T = R) occurs when $|DR| > |DC|$ and both metrics have the same sign. Thus, the dashed line ($DR = DC$) separates almost all the misclassified cases, i.e., red instances (P ≠ T = R), from the



(a) *phishing* dataset (b) *phoneme* dataset

Fig. 6. DR vs DC for each instance evaluated with MTP-TPD.



(a) *phishing* dataset (b) *phoneme* dataset

Fig. 7. The sensitivity to ε hyperparameter is studied for the regression task with the Mean Absolute Error (MAE) and for the classification task with the Error Rate (ER).

correctly classified ones, i.e., green instances (P = T = R). Therefore, a predicted privileged value sufficiently distant from the real privileged value causes a class change in the privileged model, leading it towards an incorrect class. It should be noted that the separation between red and green instances is not exact due to two reasons: the discrepancies between the privileged model and the teacher model, and the fact that *DC* is computed based on the teacher model. Another reason for failure occurs when the privileged model successfully replicates the teacher's performance, but the teacher fails. This type of instance is colored in yellow (P = T ≠ R).

The ideal privileged model is achieved when all instances are either green (P = T = R) or purple (P = R ≠ T), with no yellow (P = T ≠ R) or red (P ≠ T = R) instances. It should be noted that two scenarios are identified based on the complexity of the classification problem. On the one hand, when the classification problem is easy to solve, i.e., the problem presents a nearly-perfect teacher with a low error rate. Thereby, yellow (P = T ≠ R) and purple (P = R ≠ T) instances are not frequent. On the contrary, a challenging classification problem involves an imperfect teacher that tends to make more mistakes. Hence, the proportion of yellow (P = T ≠ R) and purple (P = R ≠ T) instances increases. For example, *phishing* dataset (see Fig. 6a) reports a teacher with a lower error rate (see Table 2) than *phoneme* dataset (see Fig. 6b). Consequently, *phishing* shows a lower proportion of yellow (P = T ≠ R) and purple (P = R ≠ T) instances than *phoneme* dataset.

5.2. How to correct misclassified instances of the privileged model?

As discussed in the previous Section 5.1, two types of errors for the privileged model have been identified: yellow (P = T ≠ R) and red (P ≠ T = R) instances. This leads to the question: *Can we correct any of these misclassified instances?* Yellow instances indicate an error shared by both the teacher and privileged models. They are particularly challenging to recover, as even the teacher, armed with real privileged information, fails to predict the correct class. However, red instances highlight the errors made by our privileged model that are correctly classified by the teacher. Moreover, almost all red instances share a common characteristic: they occur because the prediction of the privileged information is not accurate enough ($|DR| > |DC|$), causing the privileged model to predict something different from the teacher, which makes a correct prediction. Hence, it is important to emphasize the relevance of making accurate predictions of the privileged features to further improve the final classification performance. Thus, a hyperparameter ε is considered to control the importance of the regression task, so the Multi-task Privileged learning (MTP) (Eq. 3) is modified as:

$$\underset{\substack{f_r\in F_R, f_c\in F_C,\\ \theta_j,\tau}}{argmin} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{\varepsilon}{2\theta_j^2}\ell_R(x^*_{i,j}, f_r(\boldsymbol{x}_i)_j) + \log(\theta_j)\right) + \frac{1}{\tau^2}\ell_C(y_i, \sigma(f_c(\boldsymbol{x}_i, f_r(\boldsymbol{x}_i)))) + \log(\tau) \tag{9}$$

Note that hyperparameter ε can also be considered with multi-task privileged approaches with distillation (Eqs. 4 and 5). In order to determine how ε affects performance, we study the sensitivity of the models to ε hyperparameter. Thus, for $\varepsilon = \{1, 2, 5, 10, 20\}$ the performance of the regression and the classification is measured with the Mean Absolute Error (MAE) and the Error Rate (ER), respectively. For each ε, 50 repetitions of a 5-fold stratified cross-validation are performed. Fig. 7a

and b depict the results for *phishing* and *phoneme*, datasets, respectively (remaining figures are shown in Appendix D).

An increase in the ϵ value leads to a decrease in MAE and ER when dealing with a nearly-perfect teacher: this is shown in *phishing*, *mozilla*, and *wind*. Furthermore, in *phoneme* dataset, where the teacher is more prone to errors, the same pattern is observed. However, for *obesity*, classification worsens as ϵ increases despite slight improvements in regression. When the teacher makes more mistakes, the increase of ϵ leads to minimal variations in performance for *wine*, and a declining trend in performance for *diabetes* and *abalone*.

Therefore, the evaluated datasets provide insights into the use of the ϵ hyperparameter. Based on the results, accurately resembling the privileged information is not always the right approach. Predicting the privileged features from the regular features is often not straightforward, and putting more weight on the regression causes the learning process to shift its focus away from the main objective: improving classification performance. Thereby, what is invested in regression is usually lost in the classification. This usually occurs when the classification problem is challenging, i.e., in datasets with imperfect teachers, where increasing ϵ seldom contributes to measurable performance improvements. However, in datasets with nearly-perfect teachers, increasing the value of ϵ is generally beneficial for learning: by assigning more weight to the loss responsible for predicting privileged information, a better prediction is achieved, leading to improvements in classification.

6. Conclusions

In this paper, we propose a novel multi-task privileged framework capable of effectively exploiting privileged information, either on its own or with the support of knowledge distillation as a complementary approach. Our experiments show that multi-task privileged models deliver the best results on both tabular datasets and image-related problems. They outperform traditional knowledge transfer approaches in the privileged paradigm, demonstrating that jointly learning the prediction of the privileged information and the target prediction is better than learning the tasks separately. Consequently, adopting a multi-task perspective enhances generalization within the privileged paradigm.

Furthermore, when knowledge distillation is approached from the multi-task framework, the student not only has access to regular features as inputs but also to the predicted privileged features. This extra information facilitates the transfer of knowledge between the teacher and the student, what leads to better results than standard distillation approaches.

Conversely, the reasons for the success or failure of each evaluated instance are studied in detail. Hence, we can determine whether the prediction of the privileged feature is far enough to cause a class change, or if the misclassification is due to an incorrect classification of the teacher that is transferred to the privileged model.

Drawing from this analysis, we identify two scenarios based on the complexity of the classification problem. When the classification problem is challenging, i.e., with imperfect teachers, placing greater focus on the regression task does not always lead to improvements in classification. Hence, what is invested in regression is usually lost in classification. Nevertheless, when the classification problem is easy, i.e., with nearly-perfect teachers, emphasizing the regression task usually enhances the final classification performance.

There are several interesting avenues for future work. First, extending the framework to explicitly consider teacher errors as a form of label noise. Second, exploring its applicability to other domains, such as image segmentation or natural language processing. Third, addressing problems in multi-modal scenarios, where challenges such as handling heterogeneous representations or missing modality learning may play a central role. Finally, extending these methods to incorporate a student-friendly intermediate teacher [40] in the distillation process.

In conclusion, a versatile framework is introduced that can be applied to a wide range of problems and provides improved results compared to state-of-the-art approaches.

CRediT authorship contribution statement

Mario Martínez-García : Writing – review & editing, Writing - original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Jon Vadillo:** Writing – review & editing, Methodology, Investigation, Conceptualization; **Marco Pedersoli:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization; **Iñaki Inza:** Writing - review & editing, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization; **Jose A. Lozano:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Data availability

I have shared the link to my data/code at the attached file step as code.txt

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research is supported by the Basque Government through the BERC 2022–2025 program, Elkartek, IT1504-22 and BMTF project, and by the Ministry of Science and Innovation: BCAM Severo Ochoa accreditation CEX2021-001142-S/MICIU/AEI/10.13039/501100011033, PID 2022–137442 NB-I00 and Severo Ochoa grant PRE2021-099279 funded by MICIU/AEI/10.13039/501,100,011,033 and by ESF+.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.patcog.2026.113389.

Appendix A. Additional information for tabular datasets.

A.1. Extended comparison with non-distillation LUPI methods.

This appendix reports a comparison with additional non-distillation LUPI classification methods (see Table A.4): SVM+ [1], KT-SVM [2], LR+ [13], LRIT+ [13] and RVFL+ [15]. Note that the proposed multi-task privileged distillation methods outperform all the presented state-of-the-art LUPI methods.

Table A.4
Error rate comparison with non-distillation LUPI classification methods.

Method	phishing	obesity	diabetes	phoneme
SVM+	0.301±0.113	0.358±0.092	0.318±0.022	0.280±0.007
KT-SVM	0.156±0.020	0.169±0.015	0.309±0.028	0.255±0.007
LRIT+	0.155±0.021	0.204±0.015	0.312±0.027	0.306±0.010
LR+	0.155±0.021	0.151±0.016	0.311±0.026	0.297±0.010
RVFL+	0.145±0.020	0.110±0.014	0.313±0.028	0.197±0.010
MTP	0.143±0.003	0.106±0.002	0.299±0.004	0.168±0.001
MTP-PFD	**0.141±0.003**	0.105±0.002	0.299±0.004	0.170±0.002
MTP-TPD	**0.141±0.002**	**0.104±0.002**	**0.298±0.004**	**0.167±0.001**
Method	mozilla	wine	abalone	wind
SVM+	0.280±0.005	0.323±0.012	0.437±0.089	0.308±0.094
KT-SVM	0.296±0.006	0.268±0.001	0.216±0.013	0.161±0.008
LRIT+	0.268±0.006	0.273±0.011	0.216±0.012	0.161±0.008
LR+	0.274±0.006	0.272±0.011	0.216±0.012	0.161±0.008
RVFL+	0.510±0.043	0.240±0.010	0.208±0.013	0.152±0.009
MTP	0.244±0.001	0.237±0.001	0.203±0.001	0.150±0.001
MTP-PFD	**0.243±0.001**	**0.236±0.001**	**0.201±0.001**	**0.148±0.001**
MTP-TPD	0.245±0.001	**0.236±0.001**	0.203±0.002	0.150±0.001

A.2. AUC and F1 metrics

This appendix reports additional evaluation metrics to complement the presented results. Specifically, the Area Under the Curve (AUC) and F1 scores are provided for the tabular datasets (see Table A.5) as well as for the EuroSat image dataset (see Table A.6). These tables offer a more detailed comparison of the methods discussed in the manuscript, confirming that the observed trends for AUC and F1 are consistent with the error rates reported. This further demonstrates the strong performance of our approaches across different scenarios.

Table A.5
F1 and AUC evaluation metrics with tabular datasets. The results of a model learned with regular features (*Regular*), and a teacher learned with privileged and regular features (*Teacher*) are shown as baseline. Privileged distillation (*PFD*, *TPD*), Knowledge Transfer (*KT*, *KT-PFD*, *KT-TPD*) and multi-task privileged (*MTP*, *MTP-PFD*, *MTP-TPD*) models are also reported. Best results are highlighted in bold.

Method	Metric	phishing	obesity	diabetes	phoneme	mozilla	wine	abalone	wind
Regular	AUC	0.931±0.014	0.946±0.011	0.756±0.036	0.896±0.010	0.801±0.008	0.822±0.011	0.881±0.011	0.930±0.007
	F1	0.837±0.023	0.856±0.023	0.513±0.058	0.872±0.010	0.825±0.005	0.814±0.009	0.790±0.015	0.859±0.009
Teacher	AUC	0.969±0.009	0.997±0.002	0.827±0.031	0.909±0.010	0.950±0.004	0.828±0.011	0.882±0.010	0.943±0.006
	F1	0.901±0.019	0.976±0.010	0.635±0.047	0.883±0.010	0.927±0.005	0.818±0.008	0.795±0.015	0.874±0.009
PFD	AUC	0.931±0.014	0.947±0.011	0.758±0.037	0.897±0.010	0.802±0.008	0.825±0.011	0.882±0.011	0.931±0.007
	F1	0.840±0.022	0.859±0.021	0.509±0.055	0.872±0.010	0.825±0.005	0.816±0.009	0.792±0.015	**0.861±0.009**
TPD	AUC	0.930±0.014	0.943±0.013	**0.759±0.036**	0.893±0.010	0.800±0.008	0.824±0.011	0.882±0.011	0.931±0.006
	F1	0.838±0.023	0.860±0.021	0.512±0.056	0.868±0.010	0.825±0.006	0.815±0.009	0.791±0.015	0.860±0.009
KT	AUC	0.931±0.014	0.948±0.010	0.754±0.036	0.900±0.009	0.803±0.008	0.823±0.011	0.881±0.011	0.931±0.007
	F1	**0.871±0.001**	0.852±0.001	**0.521±0.001**	**0.879±0.001**	0.825±0.001	0.805±0.001	**0.797±0.001**	0.853±0.001
KT-PFD	AUC	0.932±0.014	0.949±0.010	0.756±0.037	0.901±0.009	0.804±0.008	0.825±0.011	0.882±0.011	0.931±0.007
	F1	0.840±0.022	0.855±0.021	0.510±0.061	0.875±0.009	**0.826±0.006**	0.816±0.009	0.792±0.015	0.860±0.009
KT-TPD	AUC	0.931±0.014	0.947±0.011	**0.759±0.037**	0.898±0.009	0.803±0.008	0.823±0.011	0.882±0.011	0.931±0.007
	F1	0.839±0.022	0.858±0.020	0.514±0.059	0.873±0.009	0.825±0.005	0.816±0.009	0.792±0.015	0.860±0.009
MTP	AUC	0.934±0.014	**0.952±0.010**	0.756±0.036	0.903±0.009	0.804±0.008	**0.829±0.011**	0.883±0.011	0.930±0.006
	F1	0.845±0.021	0.869±0.019	0.517±0.058	0.878±0.009	**0.826±0.005**	0.817±0.009	0.793±0.015	0.859±0.009
MTP-PFD	AUC	**0.935±0.013**	**0.952±0.011**	0.756±0.037	0.902±0.009	**0.805±0.007**	**0.829±0.011**	**0.884±0.011**	**0.932±0.007**
	F1	0.844±0.022	**0.870±0.021**	0.511±0.061	0.876±0.009	**0.826±0.006**	**0.818±0.009**	0.794±0.015	**0.861±0.009**
MTP-TPD	AUC	0.934±0.014	0.950±0.010	0.758±0.035	**0.904±0.009**	0.803±0.008	**0.829±0.011**	**0.884±0.011**	0.930±0.007
	F1	0.845±0.021	0.869±0.019	**0.521±0.058**	0.878±0.009	0.825±0.006	**0.818±0.009**	0.793±0.015	0.859±0.009

Table A.6
F1 and AUC evaluation metrics for the binary problem Highway vs. River of EuroSat Dataset. The results of a model learned with regular features (*Regular*), and a teacher learned with privileged and regular features (*Teacher*) are shown. Privileged distillation (*PFD*, *TPD*), Knowledge Transfer (*KT*, *KT-PFD*, *KT-TPD*) and multi-task privileged (*MTP*, *MTP-PFD*, *MTP-TPD*) models are also reported. Best results are highlighted in bold.

Method	Metric	NIR	SWIRC	SWIR
Regular	AUC	0.952±0.013	0.953±0.012	0.958±0.010
	F1	0.871±0.026	0.876±0.028	0.883±0.019
Teacher	AUC	0.990±0.001	0.991±0.003	0.988±0.004
	F1	0.961±0.004	0.955±0.006	0.948±0.010
PFD	AUC	0.962±0.014	0.960±0.009	0.960±0.007
	F1	0.893±0.024	0.891±0.014	0.886±0.017
TPD	AUC	0.948±0.011	0.949±0.009	0.949±0.013
	F1	0.872±0.019	0.868±0.019	0.874±0.031
KT	AUC	0.981±0.005	0.982±0.003	0.980±0.002
	F1	0.937±0.010	0.938±0.006	0.933±0.007
KT-PFD	AUC	0.982±0.004	0.981±0.002	0.978±0.005
	F1	0.937±0.008	0.936±0.005	0.930±0.011
KT-TPD	AUC	0.980±0.005	0.979±0.005	0.975±0.008
	F1	0.937±0.008	0.935±0.008	0.926±0.014
MTP	AUC	0.986±0.008	0.987±0.008	0.984±0.006
	F1	0.948±0.017	0.945±0.020	0.948±0.017
MTP-PFD	AUC	**0.989±0.003**	**0.989±0.005**	**0.988±0.003**
	F1	0.953±0.011	**0.953±0.011**	**0.952±0.014**
MTP-TPD	AUC	**0.989±0.002**	0.988±0.004	0.982±0.007
	F1	**0.956±0.006**	**0.953±0.012**	0.944±0.016

A.3. Loss weighting ablation analysis in multi-task learning.

This appendix provides an analysis of the impact of different weightings between regression and classification losses in proposed multi-task models. The straightforward multi-task approach combines multiple objective losses. It consists of performing a simple weighted linear sum of the losses for each individual task: $L_{\text{total}} = W_R \ell_R + W_C \ell_C$, where W_R and W_C are the weights and ℓ_R and ℓ_C are the corresponding loss functions, associated with the regression and classification tasks, respectively. Specifically, we consider three configurations to be compared with the proposed uncertainty-based multi-task approach: equal weights ($W_R = 1, W_C = 1$), increased weight for the classification task ($W_R = 1, W_C = 5$), and increased weight for the regression task ($W_R = 5, W_C = 1$).

Table A.7
Error rates for multi-task weighted approaches with different fixed weights and the proposed methods based on dynamic (uncertainty-based) weighting. Best model is highlighted in bold.

Method	Configuration	phishing
MTP	($W_R = 1, W_C = 1$)	0.150±0.021
MTP-PFD	($W_R = 1, W_C = 1$)	0.148±0.022
MTP-TPD	($W_R = 1, W_C = 1$)	0.149±0.021
MTP	($W_R = 1, W_C = 5$)	0.147±0.022
MTP-PFD	($W_R = 1, W_C = 5$)	0.147±0.022
MTP-TPD	($W_R = 1, W_C = 5$)	0.147±0.021
MTP	($W_R = 5, W_C = 1$)	0.150±0.022
MTP-PFD	($W_R = 5, W_C = 1$)	0.148±0.021
MTP-TPD	($W_R = 5, W_C = 1$)	0.150±0.021
MTP	Dynamic (Kendall et al. [34])	0.143±0.003
MTP-PFD	Dynamic (Kendall et al. [34])	**0.141±0.003**
MTP-TPD	Dynamic (Kendall et al. [34])	**0.141±0.002**

The results reported for the phishing dataset in Table A.7 show that the proposed approaches (MTP, MTP-PFD, MTP-TPD) achieve the best performance. It should be noted that this result does not imply that better weight configurations do not exist. However, determining the optimal configuration for each dataset requires a higher computational cost compared to the proposed privileged multi-task approaches.

A.4. Sensitivity analysis to λ and T.

In Fig. A.8, a sensitivity analysis of MTP-PFD with respect to λ and T hyperparameters is reported. The figure illustrates that the selected hyperparameters for the main experiments ($\lambda = 0.5$ and $T = 1$) lead to a promising performance.

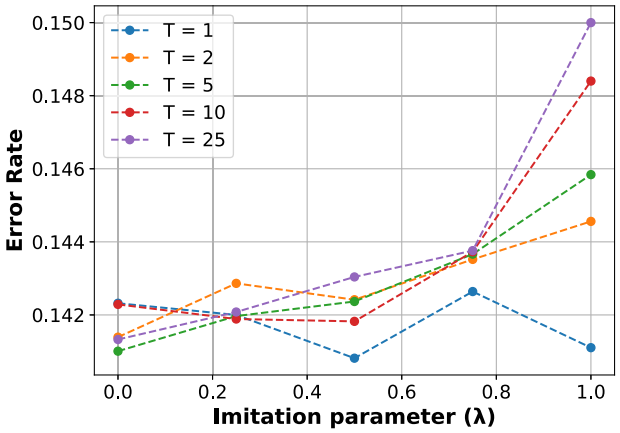


Fig. A.8. Sensitivity analysis of MTP-PFD with respect to λ and T hyperparameters in phishing dataset.

A.5. Computational complexity.

The computational complexity of our method is mainly determined by the architecture of the underlying regression and classification models employed. Furthermore, for MLPs the computational complexity is dominated by the operations between hidden layers [41]. Let d denote the input dimension, L the number of hidden layers, and h the number of neurons per layer. The computational cost of the regression network is $\mathcal{O}(dh + (L-1)h^2 + 2)$, while the classification network, has a cost of $\mathcal{O}((d+1)h + (L-1)h^2 + 2)$. The overall computational cost corresponds to the sum of both components, yielding $\mathcal{O}(2(L-1)h^2 + (2d+1)h + 4)$. As the dominant term is quadratic in the number of hidden neurons, the resulting asymptotic complexity of the proposed model is $\mathcal{O}(Lh^2)$.

Appendix B. EuroSat multiclass classification problem.

A multiclass problem is presented for the EuroSat dataset with 6 different classes: Pasture with 2000 samples; Highway, River, and PermanentCrop with 2500 samples; and Forest and AnnualCrop with 3000 samples. Note that the privileged feature is the infrared band that provides the best results in the binary case, i.e., SWIR.

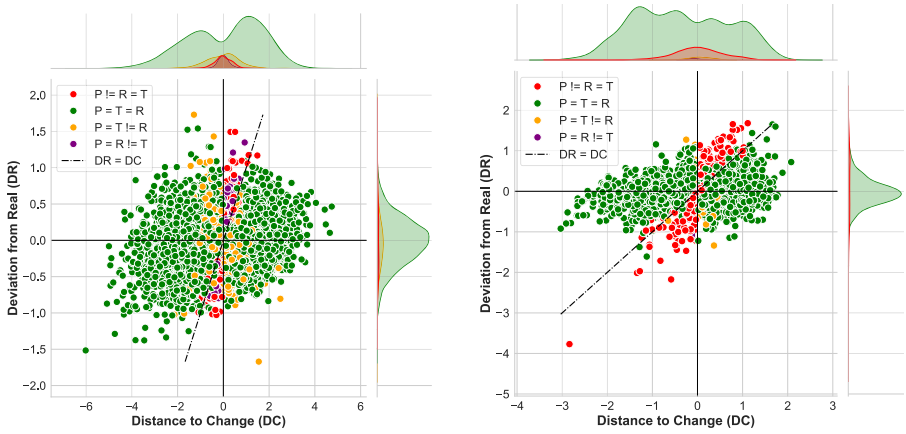
Table B.8
Overall error rate and per-class error rates for the multiclass EuroSat dataset with SWIR band as privileged information. The results of a model learned with regular features (*Regular*), and a teacher learned with privileged and regular features (*Teacher*) are shown. Privileged distillation (*PFD*, *TPD*), Knowledge Transfer (*KT*, *KT-PFD*, *KT-TPD*) and multi-task privileged (*MTP*, *MTP-PFD*, *MTP-TPD*) models are also reported. For privileged models, the LUPI gain is included between parentheses (a minus reports a non-representative LUPI gain), and best results are highlighted in bold.

Model	Overall	Highway	River	Pasture	Forest	PermanentCrop	AnnualCrop
Regular	0.189±0.024	0.373±0.102	0.229±0.057	0.186±0.058	0.039±0.028	0.182±0.047	0.160±0.059
Teacher	0.129±0.020	0.201±0.081	0.055±0.018	0.225±0.052	0.039±0.030	0.158±0.054	0.132±0.057
PFD	0.203±0.026 (-)	0.353±0.086	0.261±0.083	0.213±0.075	0.039±0.024	0.209±0.054	0.183±0.052
TPD	0.224±0.032 (-)	0.372±0.070	0.264±0.084	0.254±0.070	0.061±0.042	0.277±0.086	0.164±0.054
KT	0.159±0.017 (50.0%)	0.237±0.060	**0.107±0.015**	0.201±0.052	0.044±0.023	0.234±0.072	0.161±0.065
KT-PFD	0.161±0.014 (46.7%)	0.259±0.077	0.110±0.018	0.208±0.048	0.054±0.020	0.209±0.051	0.156±0.047
KT-TPD	0.170±0.016 (31.7%)	0.283±0.104	0.109±0.015	0.189±0.041	0.054±0.022	0.236±0.060	0.174±0.046
MTP	**0.125±0.030** (106.7%)	**0.156±0.062**	**0.107±0.015**	0.153±0.060	**0.030±0.018**	0.154±0.055	**0.113±0.041**
MTP-PFD	0.138±0.025 (85.0%)	0.197±0.085	0.146±0.049	0.191±0.053	**0.030±0.020**	**0.144±0.058**	0.147±0.042
MTP-TPD	0.285±0.162 (-)	0.552±0.406	0.530±0.408	**0.141±0.049**	0.078±0.051	0.251±0.129	0.191±0.072

The results (see Table B.8) follow the trend observed in the binary case, with MTP and MTP-PFD models achieving the best performance, although MTP-TPD shows poor behavior. Note that TPD is specifically designed for binary datasets, which may limit its generalization ability in multiclass settings. TPD method approaches the teacher when it classifies correctly, but move away from it in cases of misclassifications. In binary problems, moving away from an incorrect teacher redirects the learning in the true direction. However, in a multiclass problem, the teacher's error is avoided without steering the model toward a specific class. Further research is needed to address this limitation.

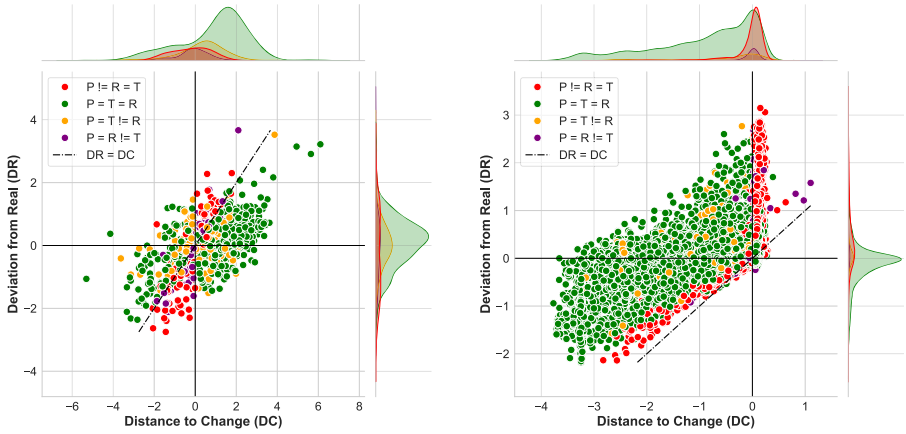
Appendix C. DR vs DC scatter plots.

The instances evaluated with MTP-TPD are plotted on the 2D graph according to their respective DR and DC values. Furthermore, different colors are used to indicate how each instance relates to the teacher's prediction and the true labels across both metrics. The results for *wind* (see Fig. C.9a), *obesity* (see Fig. C.9b), *diabetes* (see Fig. C.10a), *mozilla* (see Fig. C.10b), *wine* (see Fig. C.11a), and *abalone* (see Fig. C.11b) datasets are shown.



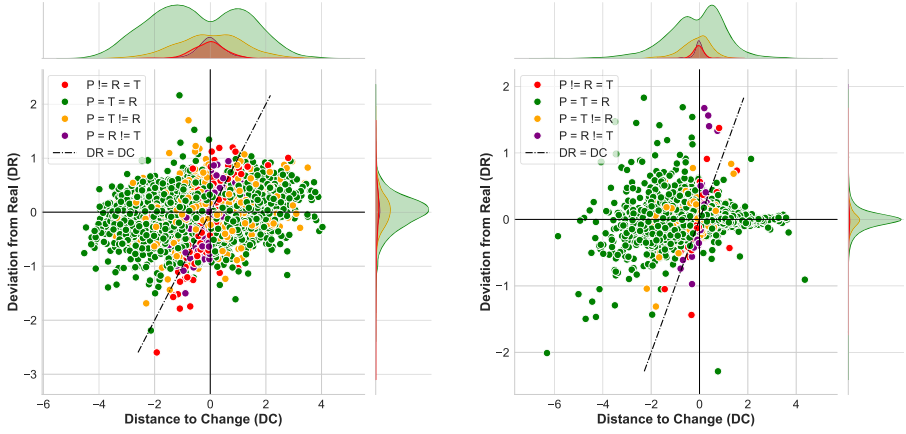
(a) *wind* dataset (b) *obesity* dataset

Fig. C.9. DR vs DC for each instance evaluated with MTP-TPD.



(a) *diabetes* dataset (b) *mozilla* dataset

Fig. C.10. DR vs DC for each instance evaluated with MTP-TPD.

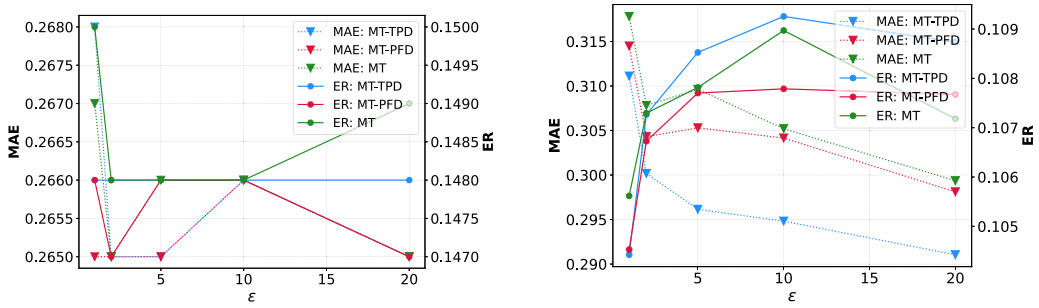


(a) *wine* dataset (b) *abalone* dataset

Fig. C.11. DR vs DC for each instance evaluated with MTP-TPD.

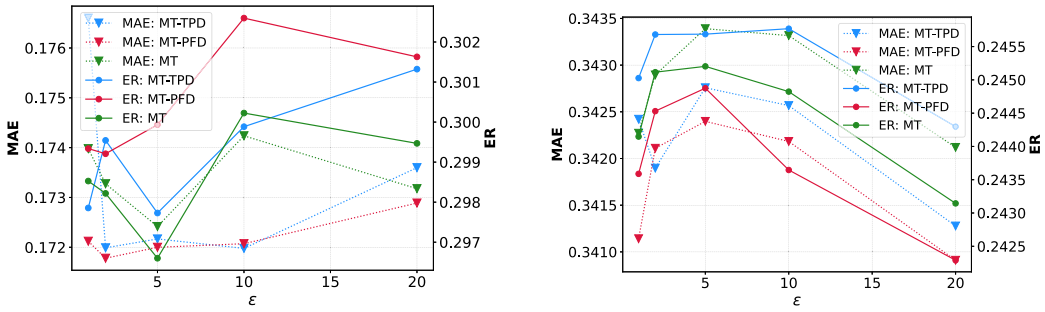
Appendix D. Sensitivity to ϵ hyperparameter

The sensitivity to the ϵ hyperparameter is evaluated by testing values $\epsilon = \{1, 2, 5, 10, 20\}$, with regression assessed via Mean Absolute Error (MAE) and classification via Error Rate (ER). The results for *wind* (see Fig. D.12a), *obesity* (see Fig. D.12b), *diabetes* (see Fig. D.13a), *mozilla* (see Fig. D.13b), *wine* (see Fig. D.14a), and *abalone* (see Fig. D.14b) datasets are shown.



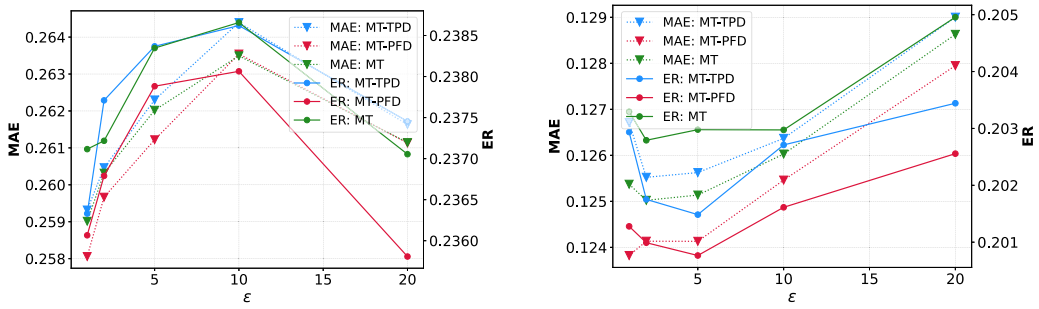
(a) *wind* dataset (b) *obesity* dataset

Fig. D.12. The sensitivity to ϵ hyperparameter is studied for the regression task with the mean absolute error (MAE) and for the classification task with the error rate (ER).



(a) *diabetes* dataset (b) *mozilla* dataset

Fig. D.13. The sensitivity to ϵ hyperparameter is studied for the regression task with the Mean Absolute error (MAE) and for the classification task with the error rate (ER).



(a) *wine* dataset (b) *abalone* dataset

Fig. D.14. The sensitivity to ϵ hyperparameter is studied for the regression task with the mean absolute error (MAE) and for the classification task with the error rate (ER).

References

[1] V. Vapnik, A. Vashist, A new learning paradigm: learning using privileged information, Neural Netw. 22 (5) (2009) 544–557.
[2] V. Vapnik, R. Izmailov, Learning using privileged information: similarity control and knowledge transfer, J. Mach. Learn. Res. (JMLR) 16 (61) (2015) 2023–2049.
[3] V. Vapnik, R. Izmailov, Learning with intelligent teacher, in: Conformal and Probabilistic Predictions with Applications (COPA), 2016, pp. 3–19.
[4] R. Izmailov, B. Lindqvist, P. Lin, Feature selection in learning using privileged information, in: IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 957–963.
[5] J. Baxter, A model of inductive bias learning, in: Journal of Artificial Intelligence Research (JAIR), 2000, pp. 149–198.
[6] R. Caruana, Multitask learning, Mach. Learn. 28 (1997) 41–75.
[7] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. 34 (12) (2022) 5586–5609.
[8] G.E. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: Neural Information Processing Systems (NeurIPS), 2015.
[9] D. Lopez-Paz, L. Bottou, B. Schölkopf, V. Vapnik, Unifying distillation and privileged information, in: International Conference on Learning Representations (ICLR), 2016.
[10] M. Martínez-García, I. Inza, J.A. Lozano, Teacher privileged distillation: how to deal with imperfect teachers? Knowl. Based Syst. 316 (2025) 113338.
[11] C. Xu, Q. Li, J. Ge, J. Gao, X. Yang, C. Pei, F. Sun, J. Wu, H. Sun, W. Ou, Privileged features distillation at taobao recommendations, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020, pp. 2590–2598.
[12] D. Pechyony, V. Vapnik, Fast Optimization Algorithms for Solving SVM+, Chapman and Hall/CRC, 2011.
[13] M. Martínez-García, S. García-Gutierrez, L. Barreñada, I. Inza, J.A. Lozano, Extending the learning using privileged information paradigm to logistic regression, Neurocomputing 615 (2025) 128869.
[14] M. Chevalier, N. Thome, G. Hénaff, M. Cord, Classifying low-resolution images by integrating privileged information in deep CNNs, Pattern Recognit. Lett. 116 (2018) 29–35.
[15] P.-B. Zhang, Z.-X. Yang, A new learning paradigm for random vector functional-link network: RVFL+, Neural Netw. 122 (2020) 94–105.
[16] R. Pasunuri, P. Odom, Learning with privileged information: decision-trees and boosting, in: International Joint Conference on Artificial Intelligence (IJCAI), 2016.
[17] R. Brachman, H. Levesque, Knowledge Representation and Reasoning, Elsevier, 2004.
[18] V. Vapnik, R. Izmailov, Knowledge transfer in SVM and neural networks, Ann. Math. Artif. Intell. 81 (1) (2017) 3–19.
[19] N. Gauraha, H. Bostrom, Investigating the contribution of privileged information in knowledge transfer LUPI by explainable machine learning, in: Conformal and Probabilistic Predictions with Applications (COPA), 2023, pp. 470–484.
[20] N. Quadrianto, V. Sharmanska, Recycling privileged learning and distribution matching for fairness, in: Neural Information Processing Systems (NeurIPS), 2017.
[21] P. Zhao, L. Xie, J. Wang, Y. Zhang, Q. Tian, Progressive privileged knowledge distillation for online action detection, Pattern Recognit. 129 (2022) 108741.
[22] C. Bian, W. Lu, W. Feng, S. Wang, Learning with privileged stereo knowledge for monocular absolute 3D human pose estimation, Pattern Recognit. Lett. 189 (2025) 143–149.
[23] S. Wang, S. Chen, T. Chen, X. Shi, Learning with privileged information for multi-label classification, Pattern Recognit. 81 (2018) 60–70.
[24] S. Fu, T. Dong, Z. Wang, Y. Tian, Weakly privileged learning with knowledge extraction, Pattern Recognit. 153 (2024) 110517.
[25] M. Collier, R. Jenatton, E. Kokiopoulou, J. Berent, Transfer and marginalize: explaining away label noise with privileged information, in: International Conference on Machine Learning (ICML), 2022, pp. 4219–4237.
[26] G. Ortiz-Jimenez, M. Collier, A. Nawalgaria, A. D'Amour, J. Berent, R. Jenatton, E. Kokiopoulou, When does privileged information explain away label noise?, in: International Conference on Machine Learning (ICML), 2023, pp. 26646–26669.
[27] K. Wang, G. Ortiz-Jimenez, R. Jenatton, M. Collier, E. Kokiopoulou, P. Frossard, Pi-DUAL: using privileged information to distinguish clean from noisy labels, in: International Conference on Machine Learning (ICML), 2024, pp. 51214–51236.
[28] S. Yang, S. Sanghavi, H. Rahmanian, J. Bakus, S.V.N. Vishwanathan, Toward understanding privileged features distillation in learning-to-rank, in: Neural Information Processing Systems (NeurIPS), 2022.
[29] F. Tang, C. Xiao, F. Wang, J. Zhou, L.-w.H. Lehman, Retaining privileged information for multi-task learning, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 1369–1377.
[30] Y. Song, Z. Lou, S. You, E. Yang, F. Wang, C. Qian, C. Zhang, X. Wang, Learning with privileged tasks, in: International Conference on Computer Vision (ICCV), 2021.
[31] Y. Shu, Q. Li, L. Liu, G. Xu, Privileged multi-task learning for attribute-aware aesthetic assessment, Pattern Recognit. 132 (2022) 108921.
[32] B. Liu, B. Li, Y. Xiao, Z. Wang, B. Zhou, S. He, C. Ye, F. Cao, Semi-supervised manifold regularized multi-task learning with privileged information, Inf. Sci. 711 (2025) 122112.
[33] W. Chen, Y. Chai, X.-J. Wu, H. Zhu, Q. Yu, Z.-M. Du, F. Han, W. Gao, C. Zheng, H. Fan, Privileged information-guided multitask mutualistic transformer for gaze prediction, IEEE Trans. Multimedia (2025) 1–16.
[34] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7482–7491.
[35] J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, in: Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3150–3158.
[36] P. Helber, B. Bischke, A. Dengel, D. Borth, EuroSAT: a novel dataset and deep learning benchmark for land use and land cover classification, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (2019) 2217–2226.
[37] D. Dua, C. Graff, UCI Machine Learning Repository, 2017.
[38] J. Vanschoren, J.N. van Rijn, B. Bischl, L. Torgo, OpenML: networked science in machine learning, SIGKDD Explorations 15 (2) (2013) 49–60.
[39] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.
[40] D. Chen, Y. Li, J. Liu, J. Zhou, Y. Gao, SATE: efficient knowledge distillation with implicit student-aware teacher ensembles, Pattern Recognit. 172 (2026) 112355.
[41] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.