24 July 2025, Preprint v2 · CC-BY 4.0

Qeios

PREPRINT

Review Article

Evaluation and Learning with Multiple Inaccurate True Targets

Yongquan Yang[1]

[1]. Institute of Sciences for AI, China

In many real-world machine learning (ML) scenarios, obtaining accurate true targets (ATTs) for evaluation and learning is difficult, expensive, or fundamentally infeasible. This article proposes a unified scientific paradigm—evaluation and learning with multiple inaccurate true targets (MIATTs)—that addresses this challenge by integrating the fundamental principles of two recently proposed frameworks: Logical Assessment Formula (LAF) and Undefinable True Target Learning (UTTL). Both LAF and UTTL operate under a relaxed but shared assumption that the true target for a given ML task is not assumed to exist as a well-defined object in the real world, which motivates us to define MIATTs as a collection of weak yet partially informative targets, each capturing a different aspect of the underlying true target. Building on this foundation, we present a comprehensive theoretical framework, which encompasses MIATT generation, model construction, metric formulation, and model optimization, for formalizing the evaluation and learning of predictive models with MIATTs. The article offers a principled and practical alternative for ML scenarios marked by true target ambiguity, providing a viable path where conventional ATT-based methods prove inadequate or inapplicable.

Corresponding author: Yongquan Yang, remy_yang@foxmail.com

1. Introduction

Advances in machine learning (ML) techniques—particularly deep learning (DL) based on deep neural networks [1]—have driven the rapid development of artificial intelligence (AI) technologies over the past decade, enabled by the availability of large-scale data and increasingly powerful computing resources. Today, large-scale models such as ChatGPT [2] represent the forefront of modern AI, fundamentally transforming knowledge acquisition, expanding the boundaries of productivity, and redefining paradigms of human–machine collaboration. A typical ML pipeline for a specific prediction task generally involves three key stages: (1) collecting task-relevant data; (2) designing or selecting an appropriate ML algorithm; and (3)

doi.org/10.32388/UMHEFG.2

optimizing the ML algorithm using the collected data and available computing resources to obtain a predictive model that can be deployed for the target task.

For a given ML task, the data typically consist of three fundamental components: instance, label, and target. An instance refers to an event or entity to be analyzed—such as a numerical vector or an image matrix—depending on the nature of the task. A label is commonly assigned to an instance and encodes semantic information relevant to the specific prediction objective. A target is usually a computationally equivalent transformation derived from a label, designed to facilitate computational procedures in ML, such as validation and learning. In practice, given available computational resources, a predictive model is typically obtained by optimizing a suitable ML algorithm using processed data consisting of paired instances and targets (rather than raw labels) [3]. Within this optimization process, validation and learning represent two essential computational procedures. Validation evaluates the model's ability to generalize by measuring the discrepancy between the model's predicted target for an instance and the true target associated with that instance, while learning focuses on producing a predictive model that maps each instance to its corresponding true target. Further explanations and interrelations among the terminologies of instance, label, target, evaluation, and learning in ML are provided in Section 2.

In the current ML literature, approaches for the evaluation of predictive models can be categorized into three types based on the accuracy of the true targets associated with the evaluation data: (1) conventional evaluation with accurate true targets (ATTs) [4][5][6], (2) conventional evaluation with inaccurate true targets (IATTs) [7][8], and (3) logical assessment formula (LAF)-based evaluation, which can operate with multiple inaccurate true targets (MIATTs) [9][10]. Conventional evaluation methods based on ATTs or IATTs typically assume that ATTs are reliably embedded within the provided true targets associated with the evaluation instances, even when the overall set of provided true targets may contain inaccuracies—that is, when the set constitutes IATTs. In contrast, the recently proposed LAF framework relaxes this assumption by allowing that ATTs may or may not be present among the given true targets [9]. In the absence of a strict requirement for ATTs, LAF offers a more adaptable and realistic framework for evaluation, particularly suited to scenarios with uncertain or noisy supervision. Following a similar perspective, approaches for the learning of predictive models in the current ML literature can likewise be categorized into three main types based on the accuracy of the true targets associated with the learning data: (1) conventional learning with ATTs [3][11][12][13][14][15], (2) conventional learning with IATTs [16][17][18][19][20], and (3) undefinable true target learning (UTTL), which can operate with MIATTs [21][22][23][24]. Conventional learning approaches based on ATTs or IATTs generally assume that the true target exists as a well-defined object in the real world. In contrast, the recently introduced UTTL paradigm relaxes this assumption by positing that the true target does not exist as a precisely defined object in the real world [24]. Accordingly, UTTL eliminates the need for ATTs during learning, making it well-suited for scenarios in which the true target is ill-defined or inherently ambiguous. A more detailed discussion of these validation and learning approaches, along with a comparison of their underlying assumptions, is provided in Section 3.

Respectively, LAF and UTTL offer distinct frameworks for evaluation with MIATTs and learning with MIATTs, without relying on ATTs. These approaches are particularly well-suited to addressing the challenge that, in many real-world ML scenarios, acquiring ATTs is difficult, expensive, or fundamentally infeasible. However, treating LAF and UTTL in isolation may hinder their broader applicability and practical deployment. To provide a more unified and effective solution, this article proposes a cohesive scientific paradigm—evaluation and learning with MIATTs—which integrates the foundational principles of both LAF and UTTL into a single framework.

Specifically, drawing on prior studies [9][10][21][22][23][24][25], we establish the scientific foundation for proposing the unified paradigm of evaluation and learning with MIATTs. The respective principles of the LAF and UTTL have independently demonstrated the feasibility of evaluation and learning in the absence of ATTs. From these principles, we derive a shared foundational assumption: the true target for a given ML task is not assumed to exist as a well-defined object in the real world. This assumption motivates the formal definition of MIATTs, conceptualized as collections of weak yet partially informative targets that each reflect certain aspects of the underlying true target. The logical progression connecting these elements—assumption, definition, and principles—collectively establishes the scientific foundation for proposing the unified paradigm. Further details on this foundation are presented in Section 4.

Building upon this foundation, we then present a comprehensive theoretical framework for formalizing the evaluation and learning of predictive models using MIATTs. This framework systematically outlines the procedures for MIATT generation, predictive model construction, metric formulation, and model optimization. By operationalizing this paradigm, it provides principled guidance for applying MIATT-based evaluation and learning of predictive models. The full framework is detailed in Section 5.

Overall, the unified scientific paradigm of evaluation and learning with MIATTs proposed in this article provides a novel and flexible framework for addressing ML tasks in domains where ATTs are difficult, costly, or fundamentally infeasible to obtain. The key contributions of this work are as follows:

- It integrates the foundational principles of both LAF and UTTL into a unified SP of evaluation and learning with MIATTs

- It clarifies the concepts and interrelationships among core ML terminologies, including instance, label, target, evaluation, and learning
- It provides a concise review of existing evaluation and learning approaches, along with a comparative analysis of their underlying assumptions
- Drawing on prior studies, it establishes the scientific foundation for a unified paradigm of evaluation and learning with MIATTs by integrating the principles of LAF and UTTL
- It presents a comprehensive theoretical framework that systematically guides the processes of MIATT generation, predictive model construction, metric formulation, and model optimization, which formalize the evaluation and learning of predictive models with MIATTs.

The remainder of this article is organized as follows. Section 2 introduces fundamental terminologies in ML and clarifies their interrelations. Section 3 reviews related work on evaluation and learning approaches, categorizing them based on assumptions regarding the availability of true targets. Section 4 establishes the scientific foundation for the proposed unified paradigm of evaluation and learning with MIATTs, building on prior studies and theoretical reasoning. Section 5 presents a theoretical framework for formalizing the evaluation and learning of predictive models with MIATTs, which cover the generation of MIATTs, construction of predictive models, formulation of evaluation metrics, and optimization procedures. Finally, Section 6 concludes the article and discusses open challenges and potential directions for future research.

2. Fundamental terminologies and their relations in ML

Generally speaking, the objective of ML is to construct a predictive model with data collected for a specific prediction task based on efficient computing resources. *Instance*, *label*, and *target* constitute the fundamental components of the data associated with a particular ML task. Section 2.1 provides detailed explanations of these three terms in the context of ML data. To achieve the goal of ML in practical applications, *learning* and *validation* are two essential concepts, representing the core computational procedures required to develop an appropriate predictive model for the given task. These two terminologies, pertaining to the implementation of ML solutions, are discussed in Section 2.2. Collectively, these five fundamental terms—related to both data collection and solution implementation—establish the conceptual foundation of ML. Section 2.3 illustrates the interrelationships among these five core elements.

2.1. Instance, label, and target

An instance in ML is usually regarded as an event or entity, which, for example, can be a number vector or an image matrix for a specific ML task. A label in ML is commonly assigned to an instance. When we refer to a label in ML, we can also implicitly refer to the instance associated with the label. When a number of labels and their corresponding instances are provided, there is some collected data for ML. A label assigned to an instance usually contains some semantic facts that mostly cannot be directly used in ML, since the semantic facts contained in the label can be too complex or simple and unstructured to be easily used for computation in ML [3]. To address this issue, ML practitioners must build a transformation that can transform a label into a target that can be conveniently used for computation [3]. Regarding the transformation from its associated label, a target can also be re-transformed into a label that contains some semantic facts [3]. As a result, a target is essentially a somewhat equivalent formation corresponding to a particular label, which can be conveniently used for computation in ML. Identical to the reference to a label in ML, when we refer to a target in ML, we can also implicitly refer to the instance associated with the label from which the target is transformed.

Beyond the general terminologies of label and target, some attributive words can be added to form new terminologies for a definitive label or target that has a more specific meaning. For example, we can add attributive words like ground-truth, accurate, or inaccurate to label or target to form new definitive terminologies. Some examples of definitive labels or targets and their meanings are provided in Table 1.

2.2. Evaluation and Learning

As we discussed in the previous section, a label assigned to an instance in the collected data commonly cannot be directly used for computation in ML. The two computational procedures for evaluation and learning of a predictive model for a specific ML task are usually implemented based on a number of targets and corresponding instances, which constitute data processed from the collected data with provided labels.

Based on the processed data that consists of a number of targets and corresponding instances, the evaluation procedure in ML aims to estimate the error between the predicted target of a predictive model regarding an instance and the target associated with the instance, while the learning procedure in ML aims to produce a predictive model that can map an instance to its corresponding target. Specifically, the evaluation procedure in ML is carried out by referring to some metrics to compute the error between the predicted target of a predictive model for an instance and the target associated with the instance [9][10], while the learning procedure in ML is carried out by optimizing a predefined objective function [26][27] based on a certain learning algorithm, such as deep neural networks [28][29], for a predictive model to implement the mapping from an instance to its corresponding target. Particularly, the evaluation procedure is usually an essential part of the learning procedure, as a specific strategy for the evaluation procedure can generally be used in the optimization of a predefined objective in the learning procedure [24].

Terminology	Meaning	Terminology	Meaning
Label	Semantic facts assigned to an instance	Target	A somewhat equivalent formation corresponding to a label for convenient computation in ML
Ground-truth label	Semantic facts that possess the intended information	True target	Formation corresponding to a ground-truth label
Accurate ground-truth label	Semantic facts that accurately possess the intended information	Accurate true target	Formation corresponding to an accurate ground-truth label
Inaccurate ground-truth label	Semantic facts that inaccurately possess the intended information	Inaccurate true target	Formation corresponding to an inaccurate ground-truth label

Table 1. Definitive label or target and their meanings

For a specific ML task, the two computational procedures of evaluation and learning will eventually produce an appropriate predictive model that can map an unseen instance to a predicted target, which can also be re-transformed into a predicted label for the unseen instance.

2.3. Relations

The relations of the illustrated five fundamental terminologies in ML are shown in Fig. 1, in which a single data point is displayed for simplicity.

The collected data for a specific ML task consist of an instance and its assigned label. As the label for the collected data usually contains some semantic facts that mostly cannot be directly used for computation in ML, the label for the instance is transformed into a target, which can also be re-transformed into the label. With the collected data and the transformation from a label to a target, we can obtain the processed data for evaluation and learning of a predictive model for the specific ML task.

On the basis of the processed data, the instance and its corresponding target transformed from the label are respectively used for computation in the evaluation and learning of a predictive model. While implementing the two computational procedures of evaluation and learning, the evaluation procedure provides a supportive foundation for the learning procedure for a predictive model for the specific ML task.

Finally, for the specific ML task, the two computational procedures of evaluation and learning of a predictive model cooperatively produce an appropriate predictive model, which can take an unseen instance as an input and output a predicted target for the unseen instance. The output predicted target of the produced predictive model regarding the input unseen instance can also be re-transformed into a predicted label that can be associated with the unseen instance.

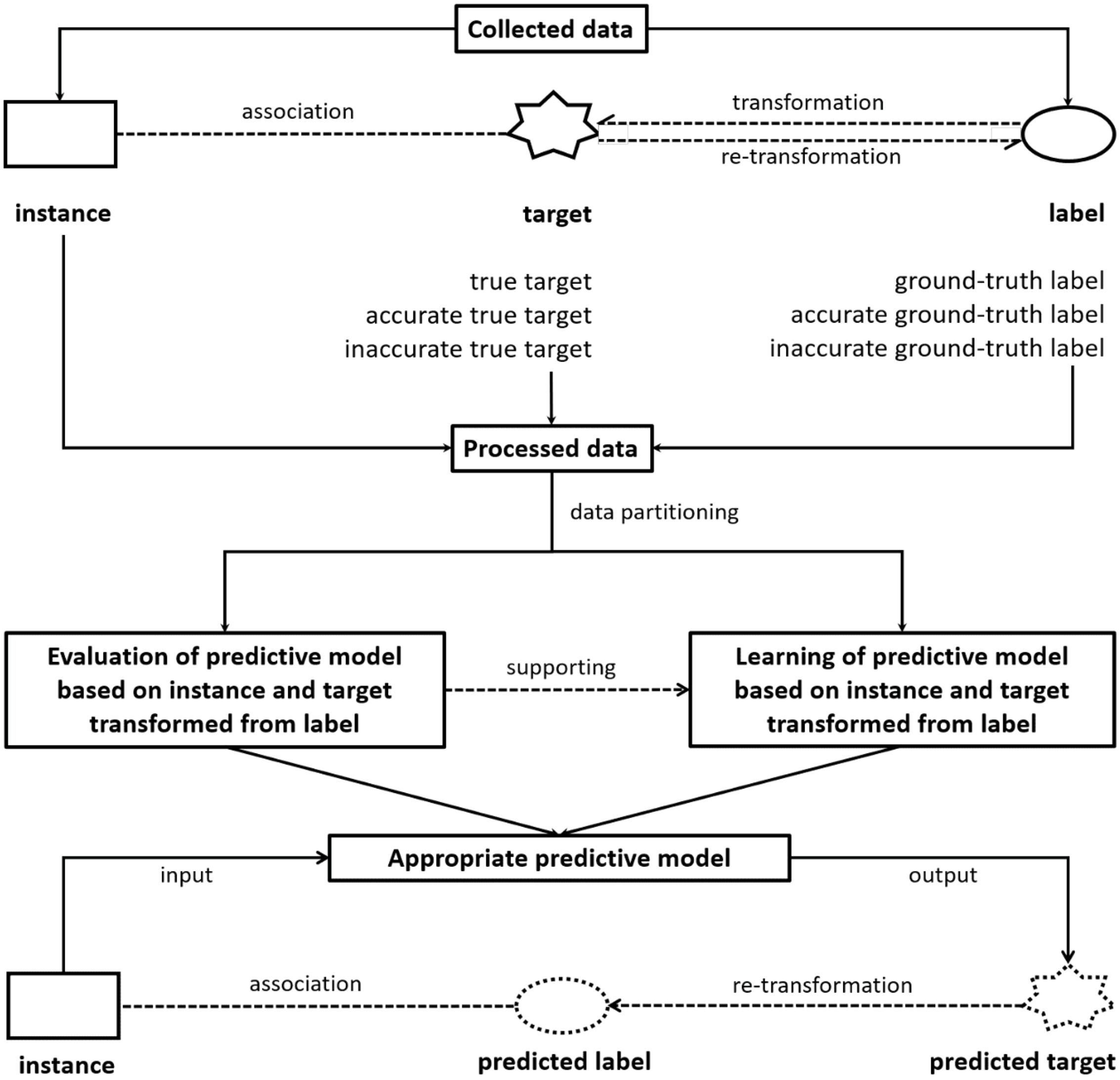

Figure 1. Outline for the relations of instance, label, target, evaluation and learning in ML.

doi.org/10.32388/UMHEFG.2

3. Related work

The evaluation and learning of predictive models are fundamental to the formulation of the new ML paradigm—evaluation and learning with MIATTs—proposed in this article. In this section, we provide a brief review of existing evaluation and learning approaches for predictive models, along with their underlying assumptions, as discussed in the current ML literature.

3.1. Evaluation of predictive models

With respect to the accuracy of true targets associated with evaluation data, we broadly categorize the evaluation approaches for predictive models in the current ML literature into three types: conventional evaluation using accurate true targets (ATTs), conventional evaluation using inaccurate true targets (IATTs), and logical assessment formula (LAF) for evaluation based on multiple inaccurate true targets (MIATTs).

3.1.1. Conventional evaluation with ATTs

To evaluate predictive models with ATTs, standard metrics such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are defined based on the comparison between predictions and AGTLs. Derived from these, metrics like precision, recall, F-measure, accuracy, and intersection over union are commonly used for classification [6] and image segmentation tasks [4][5].

To reduce the labor-intensive effort required for labeling a large number of ATTs, several evaluation approaches utilizing limited AGTLs have been proposed. For example, Jung et al. [8] proposed using pre-trained classifiers, a small set of true labels, and multiple pseudo-ground-truth label sets. Predictions from the classifiers are ranked against both true and pseudo-label sets, and the pseudo set most correlated with the true set is selected for evaluation without expert annotations. For another example, Deng et al. [7] introduced a regression-based approach: after training a classifier $f(\theta)$on labeled data (x, y), its performance on a test set (x_t, y_t) is assessed via ATT-based metrics. Then, using data augmentation to generate multiple variants of test sets, their performance statistics are extracted and used to train a regression model that predicts the classifier's performance based on feature statistics.

3.1.2. Conventional evaluation with IATTs

To enable evaluation with IATTs, Warfield et al. [30] proposed the STAPLE method, which uses the expectation-maximization (EM) algorithm [31] and related approaches [32][33] to estimate the probability that each target in the IATTs represents a true target. These probabilistic targets—derived from human raters or predictive models—can then be used to evaluate unseen test results without requiring ATTs [34][35]. Joyce et al. [36] introduced approximate ground truth refinement (AGTR), which clusters data points assumed to share the same true label. AGTR relies on expert knowledge to group similar instances and requires an estimated error rate, enabling the derivation of performance bounds for model evaluation.

3.1.3. LAF for evaluation with MIATTs

Conventional evaluation methods based on ATTs or IATTs typically assume that ATTs are reliably embedded within the provided true targets associated with the evaluation instances. In contrast, the recently proposed LAF, which can operate with MIATTs, relaxes this assumption by allowing that ATTs may or may not be present among the given true targets [9]. Rather than strictly requiring the presence of ATTs, this relaxed assumption for LAF simplifies the preparation of MIATTs for evaluation in real-world applications.

3.2. Learning of predictive models

In line with the categorization of evaluation approaches in Section 3.1, learning approaches for predictive models in the current ML literature can similarly be classified into three main types based on the accuracy of true targets associated with the learning data: conventional learning using ATTs, conventional learning with IATTs and undefinable true target learning (UTTL) for learning with MIATTs.

3.2.1. Conventional learning with ATTs

Learning predictive models with ATTs includes both supervised learning (SL) [14][15] and semi-supervised learning (SSL) [11][12][13]. SL is typically performed using large-scale labeled datasets with ATTs. Recent studies further refine SL into subtypes based on the nature of the transformation from labels to targets in the learning data [3]. In contrast, SSL leverages a small amount of labeled data with ATTs alongside a large amount of unlabeled data. Due to the incompleteness of true target information in the learning set, SSL is often considered a form of weakly supervised learning (WSL) [37].

3.2.2. Conventional learning with IATTs

Learning predictive models with IATTs encompasses both multi-instance learning (MIL) [18][19][20] and label noise learning (LNL) [16][17]. MIL is conducted on datasets where multiple instances are grouped under a single true target, with the assumption that at least one instance in each group is correctly labeled, while the others may be mislabeled. LNL, on the other hand, involves learning on data where individual instances are directly assigned potentially inaccurate labels. Due to the inherent inexactness and noise in the true target information, both MIL and LNL are also commonly regarded as forms of WSL.

3.2.3. UTTL for learning with MIATTs

Conventional learning approaches based on ATTs or IATTs typically assume that the true target exists as a well-defined object in the real world and that ATTs can, in principle, be provided for the learning data. In contrast, the recently proposed UTTL, which operates with MIATTs, relaxes this assumption by positing that the true target does not exist as a well-defined object in the real world [24]. Accordingly, UTTL does not require ATTs for learning and instead relies solely on MIATTs.

3. Comparative analysis of assumptions

A comparative summary of the assumptions underlying different evaluation and learning approaches is presented in Table 2.

Assumption	Evaluation Approach	Assumption	Learning Approach
ATTs are reliably embedded within the provided true targets	Conventional evaluation with ATTs or IATTs	The true target exists as a well-defined object in the real world	Conventional learning with ATTs and IATTs
ATTs may or may not be present among the given true targets	LAF for evaluation with MIATTs	The true target does not exist as a well-defined object in the real world	UTTL for learning with MIATTs

Table 2. Comparative summary of assumptions underlying different evaluation and learning approaches

In this article, we propose a unified scientific paradigm that is grounded in the shared assumption underlying both LAF and UTTL.

4. Scientific foundation for evaluation and learning with MIATTs

This section establishes the scientific foundation for proposing the unified scientific paradigm of evaluation and learning with MIATTs. Drawing on prior research, Section 4.1 summarizes the fundamental principles of

LAF and UTTL. These principles lead to the formulation of their shared assumption in Section 4.2, which in turn motivates the formal definition of MIATTs in Section 4.3. Finally, Section 4.4 outlines the logical progression that connects these elements, establishing the scientific foundation for the proposed unified paradigm.

4.1. Fundamental principles of LAF and UTTL

Prior research has validated the core principles that support the practical viability of LAF and UTTL across both theoretical analyses and real-world applications [9][10][21][22][23][24][25].

4.1.1. Fundamental principle of LAF

Under the assumption that ATTs may or may not be present among the given true targets (see Table 2), the foundational principle of LAF for evaluation with MIATTs was derived through deductive reasoning, demonstrating that, based on MIATTs, LAF can approximate conventional ATT-based evaluation reasonably well in complex tasks, while potentially exhibiting greater deviations in simpler ones [9][25]. Building on this principle, a series of logical assessment metrics was developed specifically for image segmentation, and the application of these LAF-based metrics in real-world scenarios has demonstrated their effectiveness [10]. Thus, we have the following Principle 1 for the practicability of LAF.

Principle 1 (Practicability of LAF). *Based on MIATTs, LAF can approximate conventional ATT-based evaluation reasonably well in complex tasks, while potentially exhibiting greater deviations in simpler ones.*

4.1.2. Fundamental principle of UTTL

Under the assumption that the true target does not exist as a well-defined object in the real world (see Table 2), a class of learning methods has been developed based on abductive reasoning, demonstrating that, based on MIATTs, UTTL can be effectively implemented within a multi-target learning framework [24][25]. These methods have been successfully applied to various medical image segmentation tasks, demonstrating notable effectiveness in practice [21][22][23]. Thus, we have the following Principle 2 for the practicability of UTTL.

Principle 2 (Practicability of UTTL). *Based on MIATTs, UTTL can be effectively implemented within a multi-target learning framework.*

4.2. Shared assumption of LAF and UTTL

From Principle 1 and Principle 2, it can be observed that the practicality of both LAF and UTTL relies on the usage of MIATTs, despite differences in their original formulations. This observation suggests the existence of a shared foundation within the respective assumptions for LAF and UTTL. In fact, the assumption underlying LAF implicitly encompasses that of UTTL, as the assumption for LAF allows for the possibility that the true target either exists or does not exist in the real world. Consequently, the shared assumption between LAF and UTTL can be distilled as follows: the true target is not necessary to exist as a well-defined object in the real world. Based on this understanding, we formulate the following Assumption 1, which represents the common foundation of both LAF and UTTL.

Assumption 1 (Shared Assumption of LAF and UTTL): *The true target for a given ML task is not assumed to exist as a well-defined object in the real world.*

4.3. Definition of MIATTs

Assumption 1 motivates the need to formally define MIATTs. By definition, a MIATT set contains two or more individual inaccurate true targets, all associated with the same underlying true target. While each individual inaccurate target in the MIATT set may be insufficient on its own to represent the true target comprehensively, it is assumed to capture a specific and meaningful aspect of it.

Consequently, the set of semantic facts (SFs) encoded in each individual inaccurate target within the MIATTs is considered a subset of the complete set of SFs that would fully characterize the underlying true target. Furthermore, the union of SFs across the entire MIATT set either belongs to or may approximate the complete SF set of the true target.

To characterize these properties formally, we introduce the following Definition 1 for MIATTs.

Definition 1 (Multiple Inaccurate True Targets, MIATTs). *Let t^* denote the (possibly undefinable) underlying true target for a given ML task, and let $SF(t^*)$ be the set of all semantic facts that precisely characterize t^*. A MIATT set $MIATTs$ associated with t^* is a finite collection*

$$MIATTs = \{{t_n}^* | n \in \{1, \cdots, N\}\}, \quad N \geq 2$$

where each ${t_n}^*$ is an inaccurate true target satisfying:

1) ***Partial representation:***

$SF({t_n}^*) \subset SF(t^*)$*,*

i.e, each ${t_n}^*$ encodes only a subset of the true target's semantic facts.

2) ***Collective coverage:***

$\bigcup_{n=1}^{N} SF\left(t_n{}^*\right) \subseteq SF(t^*)$,

with the possibility that $\bigcup_{n=1}^{N} SF\left(t_n{}^*\right) = SF(t^*)$.

In other words, no single $t_n{}^*$ *fully specifies* t^*, *but together the* $t_n{}^*$ *capture one or more of its essential aspects.*

4.4. Logical progression toward a unified paradigm

Assumption 1 regarding the true target, Definition 1 formalizing MIATTs, and the validated Principle 1 and Principle 2 supporting the practicability of LAF and UTTL collectively establish the scientific foundation for the unified paradigm of evaluation and learning with MIATTs.

The logical progression underlying the proposed framework is summarized in Table 3. As shown, Assumption 1 concerning the nature of the true target motivates Definition 1, which formally introduces MIATTs. This definition, in turn, supports the development of Principle 1 and Principle 2, demonstrating the practicability of LAF and UTTL. Together, these foundations lead to the proposal for a unified scientific paradigm of evaluation and learning.

Scientific foundation			**Proposal**
Assumption	**Definition**	**Principles**	
Assumption 1	MIATTs	Principle 1	Evaluation and learning with MIATTs
		Principle 2	

Table 3. Logical foundations underlying the unified paradigm of evaluation and learning with MIATTs

Note: *This table summarizes the logical progression of the proposed paradigm. Assumption 1 establishes the ontological premise that the true target may not exist as a well-defined object in the real world. Based on this, Definition 1 formalizes the concept of MIATTs. This formalization enables the development of Principle 1 and Principle 2, which validate the practicability of LAF and UTTL. These components collectively support the formulation of the unified paradigm for evaluation and learning with MIATTs presented in this work.*

5. Evaluation and learning of a predictive model with MIATTs

Building upon the scientific foundation established in Section 4, this section presents a systematic framework for formalizing the evaluation and learning processes of predictive models utilizing MIATTs. Section 5.1 introduces the procedure for generating MIATTs for a given instance, grounded in the formal definition provided earlier. Section 5.2 describes the construction of a predictive model that maps an instance to a predicted true target for a specific ML task in the context of MIATTs. Section 5.3 formulates metrics to quantify the discrepancy between the predicted true target and the corresponding MIATTs of an instance. Based on these metrics, Section 5.4 develops evaluation and learning procedures aimed at optimizing the predictive model to accurately approximate the desired true target. Finally, Section 5.5 offers a visual and conceptual synthesis of the entire evaluation and learning process in the context of MIATTs. Section 5.6 outlines recommendations for configuring the critical hyperparameters involved in the framework's formulation, leveraging the fundamental principles of LAF and UTTL to facilitate its practical application.

5.1. Generation of MIATTs

$MIATTs$ can be generated from the provided instance (i) and its corresponding label (l). Regarding Definition 1, generating the $MIATTs$ (GM) for i can be formally expressed as

$$\begin{aligned} MIATTs = GM\left(i, l; \theta^{GM}\right) = \{{t_n}^* | n \in \{1, \cdots, N\}\}, \ \ N \geq 2 \ \& \\ SF\left({t_n}^*\right) \subset SF\left(t^*\right) \ \& \ \bigcup_{n=1}^{N} SF\left({t_n}^*\right) \subseteq SF(t^*). \end{aligned} \tag{1}$$

Here, θ^{GM} denotes the hyperparameters for implementing the GM. For example, θ^{GM} can be the organization of a series of procedures involving humans and machines, which can be executed to generate the $MIATTs$ that can be subject to the condition $N \ \ N \geq 2 \ \& \ SF\left({t_n}^*\right) \subset SF\left(t^*\right) \ \& \ \bigcup_{n=1}^{N} SF\left({t_n}^*\right) \subseteq SF(t^*)$ for the i.

Note that the l corresponding to the i is not necessary to provide in formula (1), but it can be provided to simplify the implementation of the GM as it can provide some prior knowledge about the t^* for i.

5.2. Predictive model in context of MIATTs

In ML, a predictive model (PM) is usually regarded as a parameterized function that can map an input instance (i) into a predicted true target $(\tilde{t}\,)$, which can be formally expressed as

$$\tilde{t} = PM(i; \theta^{PM}). \tag{2}$$

Here, θ^{PM} denotes the parameters for constructing the PM.

Regarding formula (2), for a particular ML task, the essence is to find the optimized parameters ($\tilde{\theta}^{PM}$) for the PM. The found $\tilde{\theta}^{PM}$ can enable the PM to output the desired $\tilde{t}$ that has the minimum difference from the true target (t^*) associated with i. As a result, suitable metrics for measuring the difference between $\tilde{t}$ and t^* regarding i should be essentially established for finding the $\tilde{\theta}^{PM}$ for the PM. Usually, on the basis of established suitable metrics, computational procedures for the evaluation and learning of the PM are implemented for finding $\tilde{\theta}^{PM}$. In summary, based on the data of i and its corresponding t^* for a particular ML task, the construction of an appropriate PM for mapping i to the desired $\tilde{t}$ can be described as: establishing suitable metrics for the difference between $\tilde{t}$ and t^* regarding i to implement the computational procedures for the evaluation and learning of the PM to find $\tilde{\theta}^{PM}$ that can enable the PM to output the desired $\tilde{t}$ that has the minimum difference from the t^* associated with i.

In the context of MIATTs, for a particular ML task, the PM for mapping the i into the $\tilde{t}$ can still be formally expressed in the same way as formula (2). However, the construction of an appropriate PM for mapping the i into the desired $\tilde{t}$ will be different, as the data basis here is changed to the new form of the i and its corresponding $MIATTs$ instead of the traditional form of the i and its corresponding t^*. As a result, for the construction of an appropriate PM for mapping the i into the desired $\tilde{t}$ in the context of MIATTs, we need to establish new metrics for measuring the difference between $\tilde{t}$ and the $MIATTs$ regarding the i. Further, on the basis of the established new metrics, we also need to implement new computational procedures for the evaluation and learning of the PM for finding the $\tilde{\theta}^{PM}$ that can enable the PM to output the desired $\tilde{t}$ that has the minimum difference from the $MIATTs$ associated with the i.

5.3. Metrics for the difference between the predicted true target and MIATTs

The metrics (Ms) for the difference between the $\tilde{t}$ and $MIATTs$ for the i should be a series of computable equations that can quantitatively measure the discrepancies of the $\tilde{t}$ from the $MIATTs$ in various aspects. Establishing the computable equations for the Ms can be conducted on the basis of some underlying theoretical results deduced regarding the $\tilde{t}$ and the $MIATTs$. As a result, the establishment of the Ms has two key components: 1) Deducing (D) underlying theoretical results ($URTs$) regarding the $\tilde{t}$ and the $MIATTs$ for the i; 2) Establishing (E) computable equations for the Ms based on the deduced $URTs$. Formally, the two key components can be expressed as

$$URTs = D\left(\tilde{t}, MIATTs; \theta^D\right), \tag{3}$$

$$Ms = E\left(URTs; \theta^E\right). \tag{4}$$

Here θ^D and θ^E are the hyperparameters for the implementation of the D and for the implementation of the E, respectively.

doi.org/10.32388/UMHEFG.2

In implementing formula (3), we need to theoretically prove that the difference measured between the $\tilde{t}$ and the $MIATTs$ can reflect some aspects of the difference measured between the $\tilde{t}$ and the underlying t^{*} for the i. This theoretical proof is essential, as it can provide specific $URTs$ to offer a firm theoretical foundation for implementing formula (4).

In implementing formula (4), we need to experimentally validate whether the $URTs$ provided by formula (3) can be realized in practice. This experimental validation is also necessary, as it can provide computable Ms regarding a specific application.

In summary, the establishment of the metrics for the difference between the predicted true target and the MIATTs requires both theoretical proof and experimental validation, which eventually specifies the θ^{D} and θ^{E} in formulas (3) and (4).

5.4. Procedures for evaluation and learning with MIATTs

The evaluation procedure (EP) of the PM with $MIATTs$ aims to compute the numerical error (e) between the $\tilde{t}$ of the PM and the $MIATTs$ regarding the i. As the finally established Ms in the context of MIATTs is a series of computable equations, it can be used to compute the e. Thus, under the condition of the established Ms in the context of MIATTs, the EP of the PM with $MIATTs$ can be expressed as

$$e = EP\left(\tilde{t} = PM(i;\theta^{PM}), MIATTs|Ms\right). \tag{5}$$

The learning procedure (LP) of the PM with $MIATTs$ aims, under a specified learning strategy, to find the optimized parameters ($\tilde{\theta}^{PM}$) that can enable the PM to predict the $\tilde{t}$ that minimizes the e of the EP. Regarding formula (5), the LP of the PM with $MIATTs$ can be expressed as

$$\tilde{\theta}^{PM} = LP\left(arc\min_{\theta^{PM}}\left(e = EP\left(\tilde{t} = PM(i;\theta^{PM}), MIATTs|Ms\right)\right), \theta^{LP}\right). \tag{6}$$

Here θ^{LP} is the hyperparameters for implementation of the LP, such as specifying the learning strategy.

From formulas (5) and (6), we can note that the Ms established for measuring the difference between the $\tilde{t}$ and $MIATTs$ plays the decisive role in implementing the EP and LP of the PM with $MIATTs$, and the implementation of the EP also supports the implementation of the LP.

With the $\tilde{\theta}^{PM}$ found via formulas (5) and (6), we can obtain the evolved predictive model that is able to map the i into the desired true target ($\bar{t}$). This mapping can be formally expressed as

$$\bar{t} = PM\left(i;\tilde{\theta}^{PM}\right). \tag{7}$$

5.5. Summary of the Process Pipeline

The overall process pipeline for formalizing the unified scientific paradigm for evaluation and learning of predictive models with MIATTs can be visually summarized in Figure 2, which comprises four key components:

1. **MIATT Generation:** Based on the formal definition of MIATTs, generate appropriate MIATTs for each instance through human annotation and/or machine-assisted methods.
2. **Model Construction:** Design a parameterized predictive model that maps an instance to a predicted true target within the context of a specific ML task using MIATTs.
3. **Metric Formulation:** Establish evaluation metrics—supported by theoretical analysis and empirical validation—to quantify the discrepancy between the predicted true target and the corresponding MIATTs for each instance.
4. **Model Optimization:** Develop procedures for evaluation and learning based on the defined metrics. These procedures are used to optimize the model parameters, aiming to minimize the discrepancy between the predicted true targets and the MIATTs, thereby producing an improved predictive model. In the optimization process, the evaluation procedure provides supports for the learning procedure.

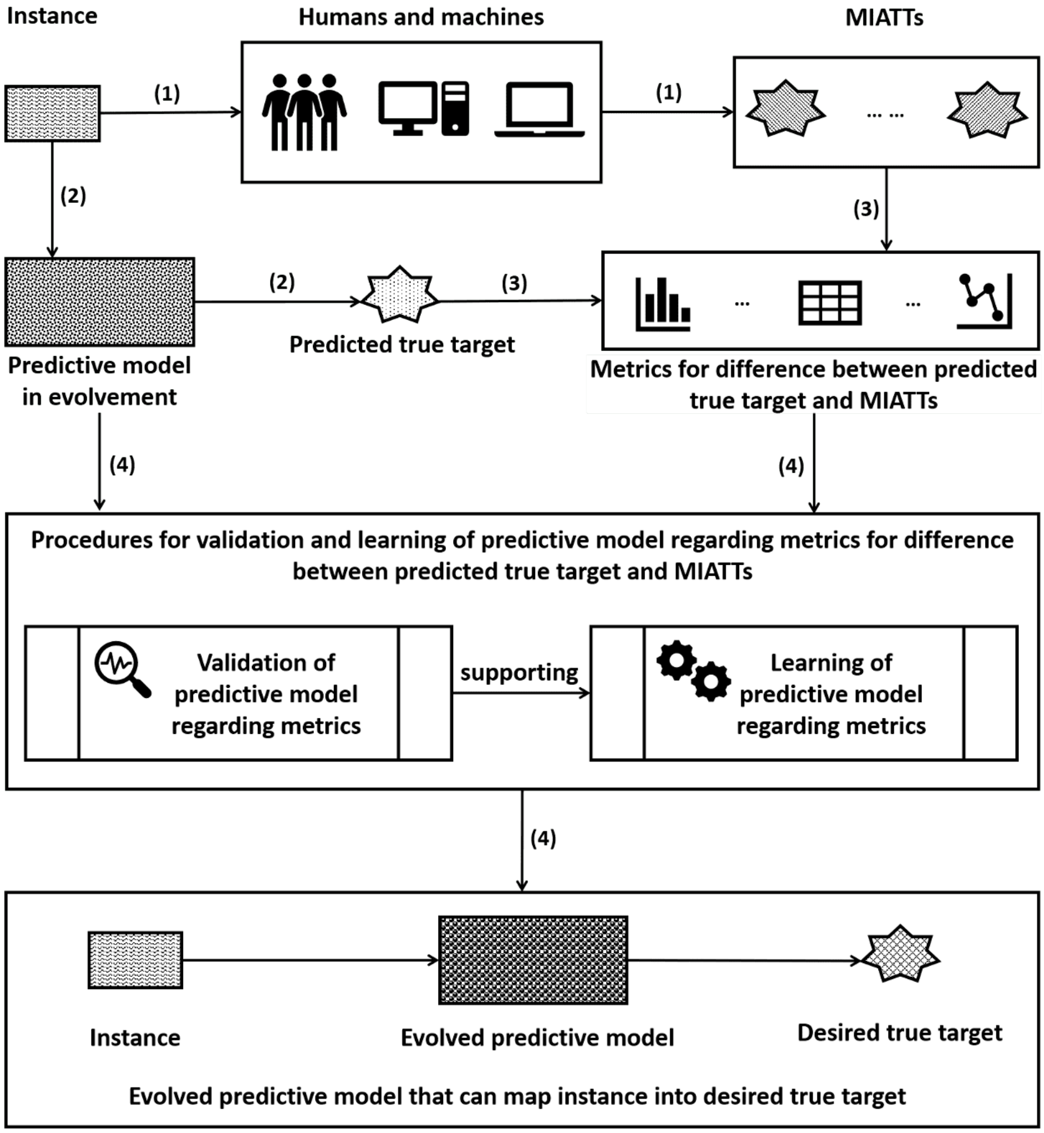
Instance
Humans and machines
MIATTs
(1)
(1)
... ...
(2)
(3)
(2)
(3)
...
...
Predicted true target
Predictive model in evolvement
Metrics for difference between predicted true target and MIATTs
(4)
(4)
Procedures for validation and learning of predictive model regarding metrics for difference between predicted true target and MIATTs
Validation of predictive model regarding metrics
supporting
Learning of predictive model regarding metrics
(4)
Instance
Evolved predictive model
Desired true target
Evolved predictive model that can map instance into desired true target

Figure 2. Summarized process pipeline for the formalization of the unified scientific paradigm for evaluation and learning of predictive model with MIATTs.

5.6. Practical implementation leveraging principles of LAF and UTTL

Formulas (1) through (7) constitute the formal foundation of the proposed framework for evaluating and learning predictive models using MIATTs. To apply this framework in practice, it is essential to specify the hyperparameters associated with these formulas, particularly those in formulas (3) to (6), which represent the core components for the evaluation and learning of predictive models within the MIATT context.

Grounded in the scientific foundation presented in Section 4, we suggest using LAF as the hyperparameter for implementing formula (3). In doing so, Principle 1—establishing the practicability of LAF—serves as the

doi.org/10.32388/UMHEFG.2

theoretical basis for employing MIATTs in evaluation tasks. Building on this, formula (4), which involves defining task-specific evaluation metrics (e.g., for image classification or segmentation), can be instantiated by translating the properties of MIATTs into measurable quantities tailored to the given ML task. These metrics then directly inform the implementation of the evaluation procedure in formula (5).

Following the formulation of the evaluation procedure, we can further leverage UTTL as the hyperparameter for implementing formula (6). This draws upon Principle 2, which establishes the feasibility of learning from MIATTs under UTTL. Taken together, the integration of LAF and UTTL into formulas (3) through (6) ensures that the evaluation and learning procedures are both theoretically grounded and practically executable.

6. Conclusion and Future Work

The unified scientific paradigm of evaluation and learning with multiple inaccurate true targets (MIATTs) proposed in this article provides a novel and flexible framework for addressing machine learning (ML) tasks in domains where accurate true targets (ATTs) are difficult, costly, or even impossible to obtain. By integrating the Logical Assessment Formula (LAF) [9] and Undefinable True Target Learning (UTTL) [24], this framework relaxes conventional assumptions about true target accuracy and offers an alternative path forward grounded in the assumption that the true target for a given ML task is not assumed to exist as a well-defined object in the real world. This epistemological shift allows for a broader class of problems to be formally addressed, particularly in areas such as medical image analysis, robotics perception, and AI alignment [25], where data ambiguity and subjective annotations are prevalent.

Prior studies have shown that LAF can approximate conventional evaluation effectively in complex prediction tasks, while UTTL has been successfully instantiated within multi-target learning frameworks [9][24][25]. These developments collectively underpin the core scientific foundation supporting the practicability of evaluation and learning with MIATTs. While prior empirical validations of LAF and UTTL provide indirect support for the proposed framework of evaluation and learning with MIATTs [10][21][22][23], targeted experimental validation within specific application domains is essential to more directly assess its robustness and generalizability. Moreover, the development of standardized datasets and validation protocols tailored for MIATT-based evaluation and learning will be critical for promoting broader adoption and ensuring reproducibility within the ML community. Future work will focus on addressing these issues to further advance the applicability of the unified scientific paradigm of evaluation and learning with MIATTs.

Acknowledgements

The author, Yongquan Yang, especially thanks his Yang Family, Chengdu, China, for providing him with financial support and mental encouragement to do this research.

References

1. ^LeCun Y, Bengio Y, Hinton G (2015). "Deep Learning." Nature. **521**:436–444. doi:10.1038/nature14539.
2. ^Welsby P, Cheung BMY (2023). "ChatGPT." Postgrad Med J. **99**:1047–1048. doi:10.1093/postmj/qgad056.
3. a, b, c, d, e, f Yang Y (2024). "Moderately Supervised Learning: Definition, Framework and Generality." Artif Intell Rev. **57**:37. doi:10.1007/s10462-023-10654-6.
4. a, b Chang HH, Zhuang AH, Valentino DJ, Chu WC (2009). "Performance Measure Characterization for Evaluating Neuroimage Segmentation Algorithms." NeuroImage. doi:10.1016/j.neuroimage.2009.03.068.
5. a, b Taha AA, Hanbury A (2015). "Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool." BMC Med Imaging. **15**:29. doi:10.1186/s12880-015-0068-x.
6. a, b M H, MN S (2015). "A Review on Evaluation Metrics for Data Classification Evaluations." Int J Data Mining Knowl Manag Process. **5**:01–11. doi:10.5121/ijdkp.2015.5201.
7. a, b Deng W, Zheng L (2021). "Are Labels Always Necessary for Classifier Accuracy Evaluation?" In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15069–15078.
8. a, b Jung HJ, Lease M (2012). "Evaluating Classifiers Without Expert Labels." doi:10.48550/arxiv.1212.0960.
9. a, b, c, d, e, f, g, h, i Yang Y (2024). "Logical Assessment Formula and Its Principles for Evaluations With Inaccurate Ground-Truth Labels." Knowl Inf Syst. doi:10.1007/s10115-023-02047-6.
10. a, b, c, d, e, f Yang Y, Bu H (2024). "Validation of the Practicability of Logical Assessment Formula for Evaluations With Inaccurate Ground-Truth Labels: An Application Study on Tumour Segmentation for Breast Cancer." Comput Artif Intell. **2**:1443. doi:10.59400/cai.v2i2.1443.
11. a, b Yang X, Song Z, King I, Xu Z (2023). "A Survey on Deep Semi-Supervised Learning." IEEE Trans Knowl Data Eng. **35**:8934–8954. doi:10.1109/tkde.2022.3220219.
12. a, b Van Engelen JE, Hoos HH (2020). "A Survey on Semi-Supervised Learning." Mach Learn. **109**:373–440. doi:10.1007/s10994-019-05855-6.
13. a, b Han K, Sheng VS, Song Y, Liu Y, Qiu C, Ma S, Liu Z (2024). "Deep Semi-Supervised Learning for Medical Image Segmentation: A Review." Expert Syst Appl. **245**:123052. doi:10.1016/j.eswa.2023.123052.
14. a, b Supervised Learning (n.d.). In: Cognitive Technologies. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 21–49. doi:10.1007/978-3-540-75171-7_2.

15. a, b*Supervised Learning (2011). In: Web Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 63–132. doi:10.1007/978-3-642-19460-3 3.*

16. a, b*Song H, Kim M, Park D, Shin Y, Lee J-G (2023). "Learning From Noisy Labels With Deep Neural Networks: A Survey." IEEE Trans Neural Netw Learning Syst.* ***34****:8135–8153. doi:10.1109/TNNLS.2022.3152527.*

17. a, b*Natarajan N, Dhillon IS, Ravikumar PK, Tewari A (2013). "Learning With Noisy Labels." In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.*

18. a, b*Fatima S, Ali S, Kim H-C (2023). "A Comprehensive Review on Multiple Instance Learning." Electronics.* ***12****:4323. doi:10.3390/electronics12204323.*

19. a, b*Foulds J, Frank E (2010). "A Review of Multi-Instance Learning Assumptions." Knowl Eng Rev.* ***25****:1–25. doi:10.1017/s026988890999035x.*

20. a, b*Multiple Instance Learning (2016). In: Multiple Instance Learning. Cham: Springer International Publishing. pp. 17–33. doi:10.1007/978-3-319-47759-6 2.*

21. a, b, c, d, e*Yang Y, Li F, Wei Y, Chen J, Chen N, Alobaidi MH, Bu H (2024). "One-Step Abductive Multi-Target Learning With Diverse Noisy Samples and Its Application to Tumour Segmentation for Breast Cancer." Expert Syst Appl.* ***251****:123923. doi:10.1016/j.eswa.2024.123923.*

22. a, b, c, d, e*Yang Y, Yang Y, Chen J, Zheng J, Zheng Z (2024). "Handling Noisy Labels Via One-Step Abductive Multi-Target Learning and Its Application to Helicobacter Pylori Segmentation." Multimed Tools Appl. doi:10.1007/s11042-023-17743-2.*

23. a, b, c, d, e*Yang Y, Yang Y, Yuan Y, Zheng J, Zhongxi Z (2020). "Detecting Helicobacter Pylori in Whole Slide Images Via Weakly Supervised Multi-Task Learning." Multimed Tools Appl.* ***79****:26787–26815. doi:10.1007/s11042-020-09185-x.*

24. a, b, c, d, e, f, g, h, i*Yang Y (2024). "Undefinable True Target Learning." doi:10.32388/KBK3P8.*

25. a, b, c, d, e, f*Yang Y (2023). "Discovering Scientific Paradigms for Artificial Intelligence Alignment." http://dx.doi.org/10.13140/RG.2.2.15945.52320.*

26. ^*Bian K, Priyadarshi R (2024). "Machine Learning Optimization Techniques: A Survey, Classification, Challenges, and Future Research Issues." Arch Computat Methods Eng. doi:10.1007/s11831-024-10110-w.*

27. ^*Wang Q, Ma Y, Zhao K, Tian Y (2022). "A Comprehensive Survey of Loss Functions in Machine Learning." Ann Data Sci.* ***9****:187–212. doi:10.1007/s40745-020-00253-5.*

28. ^*Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, Shahzad M, Yang W, Bamler R, Zhu XX (2023). "A Survey of Uncertainty in Deep Neural Networks." Artif Intell Rev.* ***56****:1513–1589. d*

oi:10.1007/s10462-023-10562-9.

29. ^*Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017). "A Survey of Deep Neural Network Architectures and Their Applications." Neurocomputing.* ***234****:11–26. doi:10.1016/j.neucom.2016.12.038.*
30. ^*Warfield SK, Zou KH, Wells WM (2004). "Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation." IEEE Trans Med Imaging. doi:10.1109/TMI.2004.828354.*
31. ^*Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood From Incomplete Data Via the EM Algorithm." J R Stat Soc B Methodol. doi:10.1111/j.2517-6161.1977.tb01600.x.*
32. ^*Warfield SK, Zou KH, Kaus MR, Wells WM (2002). "Simultaneous Validation of Image Segmentation and Assessment of Expert Quality." In: Proceedings - International Symposium on Biomedical Imaging. doi:10.1109/ISBI.2002.1029201.*
33. ^*Warfield SK, Zou KH, Wells WM (2002). "Validation of Image Segmentation and Expert Quality With an Expectation-Maximization Algorithm." In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/3-540-45786-0_37.*
34. ^*Martin-Fernandez M, Bouix S, Ungar L, McCarley RW, Shenton ME (2005). "Two Methods for Validating Brain Tissue Classifiers." In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 515–522. doi:10.1007/11566465_64.*
35. ^*Bouix S, Martin-Fernandez M, Ungar L, Nakamura M, Koo MS, McCarley RW, Shenton ME (2007). "On Evaluating Brain Tissue Classifiers Without a Ground Truth." NeuroImage. doi:10.1016/j.neuroimage.2007.04.031.*
36. ^*Joyce RJ, Raff E, Nicholas C (2021). "A Framework for Cluster and Classifier Evaluation in the Absence of Reference Labels." In: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security. New York, NY, USA: ACM. pp. 73–84. doi:10.1145/3474369.3486867.*
37. ^*Zhou Z-H (2018). "A Brief Introduction to Weakly Supervised Learning." Natl Sci Rev.* ***5****:44–53. doi:10.1093/nsr/nwx106.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.