Human in the Latent Loop (HILL): Interactively Guiding Model Training Through Human Intuition

Daniel GEISSLER [a]Lars KRUPP [a] Vishal BANWARI [a] David HABUSCH [a]
Bo ZHOU [a,b] Paul LUKOWICZ [a,b] Jakob KAROLUS [a,b]
[a] *German Research Center for Artificial Intelligence (DFKI), Germany*
[b] *University of Kaiserslautern-Landau (RPTU), Germany*

Abstract. Latent space representations are critical for understanding and improving the behavior of machine learning models, yet they often remain obscure and intricate. Understanding and exploring the latent space has the potential to contribute valuable human intuition and expertise about respective domains. In this work, we present HILL, an interactive framework allowing users to incorporate human intuition into the model training by interactively reshaping latent space representations. The modifications are infused into the model training loop via a novel approach inspired by knowledge distillation, treating the user's modifications as a teacher to guide the model in reshaping its intrinsic latent representation. The process allows the model to converge more effectively and overcome inefficiencies, as well as provide beneficial insights to the user. We evaluated HILL in a user study tasking participants to train an optimal model, closely observing the employed strategies. The results demonstrated that human-guided latent space modifications enhance model performance while maintaining generalization, yet also revealing the risks of including user biases. Our work introduces a novel human-AI interaction paradigm that infuses human intuition into model training and critically examines the impact of human intervention on training strategies and potential biases.

Keywords. Latent Space, Interactive Learning, Human Intuition

1. Introduction

The interplay between layers in deep learning models forms latent representations that capture the inherent properties and relationships within data. Through a progressive learning process, each layer extracts and refines specific features, transforming raw input into higher-level abstractions that ultimately guide the model's decisions [20]. This distributed learning approach enables deep models to encode complex patterns effectively [26]. Despite the central role of latent representations in decision-making, their structure and behavior are often opaque and obscure to human interpretation, making it difficult to assess or influence how the model encodes the information internally [37,29]. While much research focused on optimizing models through loss functions and architecture design, little attention has been given to directly examining and influencing how data is represented within the latent spaces of intermediate layers during training [34]. Exist-

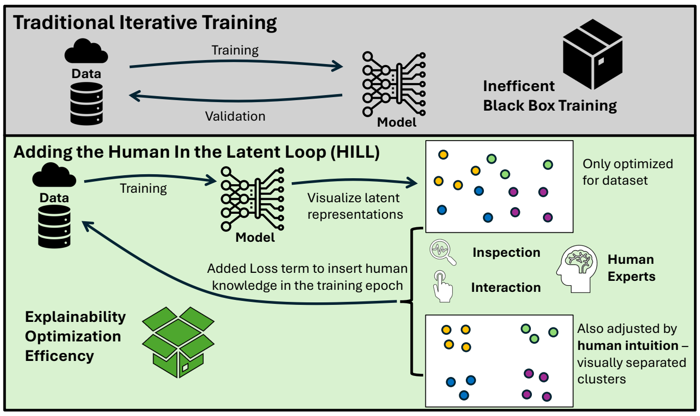

Traditional Iterative Training
Data
Training
Validation
Model
Inefficent
Black Box Training
Adding the Human In the Latent Loop (HILL)
Data
Training
Model
Visualize latent representations
Only optimized for dataset
Inspection
Interaction
Human Experts
Added Loss term to insert human knowledge in the training epoch
Explainability
Optimization
Efficency
Also adjusted by **human intuition** – visually separated clusters

Figure 1. Comparison of traditional iterative training, commonly resulting in a unsatisfying black-box training, with the approach of adding the Human In the Latent Loop (HILL) to insert the knowledge of human experts towards explainability, optimization and efficiency.

ing methods primarily focus on post hoc visualizations or indirect feature engineering, ignoring the potential of training-time, and interactive human feedback to enhance model performance and interpretability [10,50]. Consequently, there is a need for approaches that allow for more profound insight into the internal dynamics of model learning and create opportunities for humans to guide and refine these processes [13].

In this work, we propose a novel human-AI interaction paradigm at training time called Human In the Latent Loop (HILL), combining human intuition with latent space exploration and integrating it into the training loop (Figure 1). By visualizing and allowing modification of latent spaces during model training, users can directly influence the structure of the latent space to reflect human intuition, making use of their intrinsic domain knowledge and human discernment. Our approach relies on a knowledge distillation mechanism, where user-guided adjustments act as knowledge teacher to guide the model in refining its internal latent representation. As such, our approach does not manipulate the model nor the data but rather guides the model to refine the learning mechanism according to the user's input. We evaluated HILL in a user study tasking participants to find optimal models for two well-known machine learning (ML) datasets. The participants were supported by HILL visualizing the latent space and iteratively refined it based on their intuition to improve the model. We recorded the employed strategies and acquired feedback from participants through post-hoc questionnaires and interviews. Our results show that participants used diverse strategies, such as increasing cluster compactness or maximizing class separation, leading to improved accuracy and faster convergence compared to baseline training.

Our contributions can be summarized as follows:

1. We introduce HILL, an interactive tool for latent space exploration that allows users to monitor and guide the model directly during training. HILL is open-source and available on GitHub (https://github.com/DFKIEI/HITL-ML).
2. We develop a loss function insertion approach inspired by knowledge distillation, that integrates user modifications into the training process, enhancing the structural organization of latent spaces.
3. We evaluate our framework in a user study, demonstrating its effectiveness while critically examining the aspects of human intuition and its potential manipulation.

2. Related Work

Human-in-the-loop (HITL) ML is an emerging paradigm that incorporates human expertise into the model development process to enhance performance while addressing inherent limitations [47]. By integrating human judgment and intuition, HITL systems provide adaptability and reliability in applications requiring specialized domain expertise [45].

HITL techniques have demonstrated success in various ML tasks, particularly in data preprocessing, annotation, and iterative labeling. For instance, humans can help transform unstructured data into structured formats, ensuring the creation of high-quality training datasets [52]. Human-guided workflows also enable the creation of rules for data extraction, improving consistency and reducing computational burdens [25]. In health research, HITL approaches have been used to address tasks like subspace clustering and k-anonymization, where human intervention reduces algorithmic complexity while maintaining accuracy [18]. Additionally, HITL systems in interactive ML frameworks empower users to iteratively refine model outputs, achieving improved performance in tasks such as classification, regression, and clustering [9].

Beyond data preparation, HITL has played a significant role in enhancing the interpretability of ML models. Deep reinforcement learning allows to generate human-readable explanations for model predictions in critical domains like medical diagnostics [40]. Similarly, concept-based methods enable humans to interact with interpretable latent spaces, facilitating actionable refinements and diagnostics for model behavior [22].

Integrating human feedback into the training loop presents challenges. One prominent issue is the variability in user strategies leading to inconsistent modifications [11]. Thus, achieving uniform improvements across user interactions to design reproducible workflows is complex [14]. Additionally, cognitive biases inherent in human decision-making may inadvertently introduce inaccuracies, requiring careful handling of human-algorithm dynamics [19].

Consequently, HITL systems must compromise between leveraging human intuition and preserving the computational rigor of ML algorithms. Moreover, ensuring transparency and accountability in workflows is essential to maintain trust in the system [30]. Scalability issues further challenge real-time integration of user inputs [43].

Recent advances in HITL research focus on active learning and real-time feedback mechanisms, enabling models to adapt more dynamically during training [4]. Central to this progress is the concept of human-AI symbiosis combining computational efficiency with domain-specific human intuition [31]. Facilitating seamless human-AI interactions allows humans to guide algorithms dynamically, ensuring that models align with user expectations [41] even in the absence of specialization [6]. Especially for latent space inspections and modifications, human expertise is relevant to properly comprehend the active learning progression and its emerging challenges [50].

Challenges such as trust [35] and cognitive barriers [51] still delay the integration of HITL-based AI systems into traditional workflows, necessitating a compromise between explainability, interactivity, and usability to enhance human-AI collaboration.

3. Designing for Integration of Human Intuition

By synergistically infusing human expertise into ML pipelines, we can unlock unprecedented levels of model interpretability and performance, ultimately bridging the gap between artificial and human intelligence as we strive to replicate the nuanced complexities of human cognition [12]. However, this integration requires careful consideration to balance human knowledge and intuition with the algorithmic nature of ML models [46]. Unlike explicit domain knowledge, which is often systematic and well-documented, human intuition can be subjective, biased, and rooted in individual experiences [49]. Misinterpreting intuition as knowledge or over-relying on it can lead to misguided decisions and suboptimal model behavior.

3.1. Pitfalls of Human Intuition

Every individual, whether an expert or novice in the field, brings unique experiences, beliefs, and perspectives, especially when designing and training complex deep learning models [44]. Such user-dependent biases, often unconscious, profoundly stem from the intricate nature of human decision-making, which often incorporates subjective interpretations, heuristics, perceptions, and personal experiences into the process. This unconsciousness can influence how data is processed, incorporated, and optimized, leading to inconsistent and potentially biased outcomes [1]. A prominent example can be observed in the training of large language models, where datasets represent the worldwide as well as local biases, imbalances, and user preferences that contribute to unintended outcomes, such as skewed representations or unfair predictions [8].

Beyond those dataset issues, the direct insertion of human knowledge into the algorithmic training process introduces additional challenges. Overreliance on human knowledge, especially from single or small user groups, can result in aggressive interventions that compromise the model's generalization capabilities, leading to overfitting to specific user preferences or beliefs [5]. This problem becomes exacerbated when users lack complete knowledge of the system or its intricacies, leading to suboptimal guidance through intuition rather than knowledge. However, the model integrates those inputs as ground truth, relying on the user's subjective biases and limitations.

Therefore, we aim to establish a clear distinction between guidance and manipulation. Human intuition should function as a subtle influence, offering directional cues without dictating the entire learning process. This differentiation is essential to prevent the model from becoming a mere extension of individual biases, which would undermine its objectivity and robustness. AI systems should leverage human intuition as a complementary force, to enhance their capabilities while maintaining the algorithmic integrity.

Finally, the varying strategies individuals employ when interacting with ML models or visualizations further underscore the pitfalls of human intuition [33]. Users may interpret and manipulate data differently, for instance by clustering, spreading, or reordering data points according to subjective criteria, potentially introducing inconsistencies into the training process. Addressing these challenges requires designing systems that feedback the human input with algorithmic checks such as validation performance metrics, ensuring that models remain guided but not dominated by human intuition.

3.2. Training Guidance Through Human Intuition

Including the human into model training requires a symbiotic learning paradigm where algorithmic optimization and human expertise interact synergistically. Our approach draws inspiration from the widely-used knowledge distillation technique, traditionally involving a teacher-student setup [17]. In typical scenarios, a large, pretrained teacher model transfers its knowledge to a smaller student model during training, using a shared loss function to improve the student's performance. In this work, we replace the large model with a human expert, who acts as the teacher, guiding the model through intuitive adjustments and domain-specific insights. The general framework behind HILL can be found in Figure 2, outlining the idea of human guidance instead of manipulation.

A core aspect of our framework is interactive visualization, which bridges human intuition and model training by allowing users to observe and refine latent space representations. Unlike non-deterministic techniques like t-SNE [48] or UMAP [32], we use a deterministic transformation via fully connected layers to project high-dimensional latent data into a stable, interpretable 2D space. This transformation layer is frozen after the first epoch, ensuring that changes in the visualization reflect model updates rather than mapping variations.

As depicted in Figure 2, we measure the human-guided interventions and strategies based on three key metrics: the class center movement, considering the movement of whole class clusters and therefore shifting the center of the class; the spread of classes, representing the compactness change of each class such as moving all data points closer or further away from its designated center; and separation of clusters, measuring the distance of clear boundaries between individual class clusters. These modifications are incorporated into the training process through a weighted loss function. Rather than directly manipulating transformed data back into the original feature space, which usually requires a lossy and error-prone inverse of the dimension reduction, the human teacher's adjustments are encoded within $\mathscr{L}_{\text{human}}$, a loss function that includes the center alignment, spread, and separation. Combined with the standard cross-entropy loss $\mathscr{L}_{\text{CE}}$, the global loss $\mathscr{L}_{\text{global}}$ integrates both human guidance and algorithmic training, while the adjustable parameter α allows control for the level of human knowledge insertion. As a guideline, we experimented α and selected 0.5 as the balance in order to maximize the human guidance while preserving the cross-entropy-based classification nature. The additional $\lambda \cdot |1.0 - \text{scale}_{\text{model}}|$ term in the global loss ensures that the model maintains a consistent scale for features during training, preventing unintended distortions or penalization due to the range of the latent space visualizations and modifications. Throughout our experiments, fixing λ to 0.1 proved to regulate the strength of the model scale and stabilize the overall human guidance approach.

3.3. The Latent Loop Tool (HILL)

To effectively integrate the human as a teacher according to the framework from Figure 2, we developed HILL, an interactive tool designed to facilitate human-model collaboration. The tool enables users to inspect, analyze, and guide the latent space representations during training, ensuring a balance between learned model representations and domain-specific human knowledge according to $\mathscr{L}_{\text{global}}$. We implemented an iterative and interactive process, in which the model is trained epoch-wise through the tool, allowing the user to pause the training to visualize the latent representation and feedback the latent adaptations for the next training iteration. Noteworthy, the architecture of the tool is based on PyTorch [36], designed to allow individual, custom models, and datasets through two appropriate code interfaces to insert relevant data loaders and model architecture definitions.

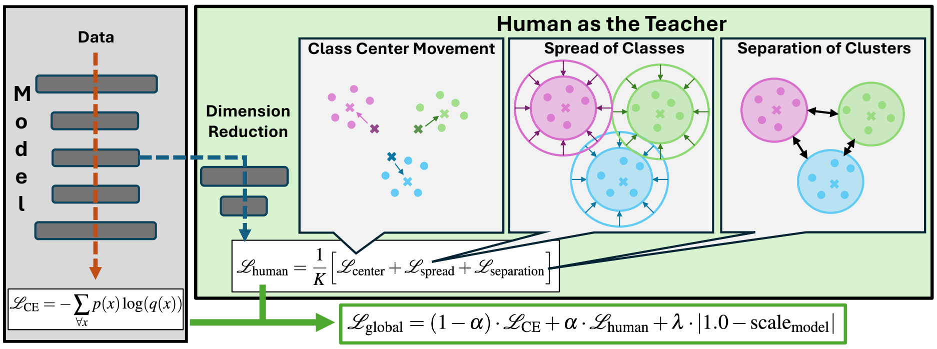

Data
Model
$\mathscr{L}_{\mathrm{CE}} = -\sum_{\forall x} p(x)\log(q(x))$
Human as the Teacher
Dimension Reduction
Class Center Movement
Spread of Classes
Separation of Clusters
$\mathscr{L}_{\mathrm{human}} = \frac{1}{K}\left[\mathscr{L}_{\mathrm{center}} + \mathscr{L}_{\mathrm{spread}} + \mathscr{L}_{\mathrm{separation}}\right]$
$\mathscr{L}_{\mathrm{global}} = (1-\alpha)\cdot\mathscr{L}_{\mathrm{CE}} + \alpha\cdot\mathscr{L}_{\mathrm{human}} + \lambda\cdot|1.0 - \mathrm{scale}_{\mathrm{model}}|$

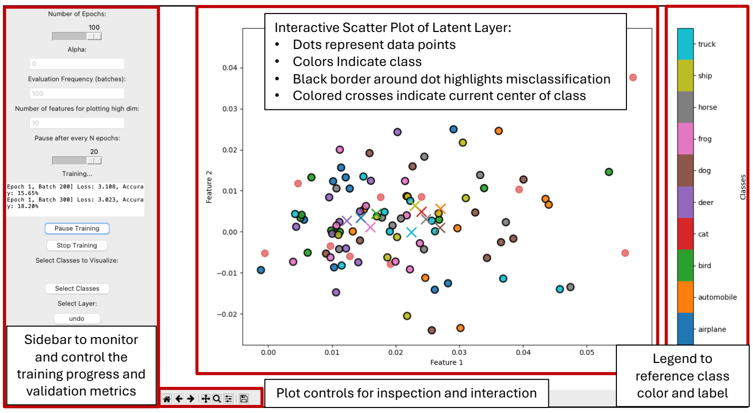
Figure 2. Utilizing a weighted loss function through α to balance the classic cross-entropy with the human as the teacher. The human input is gathered from center movement, spread of classes and separation of clusters, which are finally normalized over the total number of pairwise comparisons K. Additionally the scale of the model is added in the global loss to regulate the loss function further.

The tool's UI comprises three main components as shown in Figure 3: (1) a control panel as a sidebar on the left, which provides functionalities such as training pause/resume, epoch adjustments, hyperparameter adjustment for the weighted loss function, and a text box printing the current validation metrics; (2) a scatter plot visualization as the main part of the window, displaying a two-dimensional scatter plot projection of the reduced latent space for intuitive inspection and guidance; and (3) a class reference legend on the right side, allowing users to map visual representations to corresponding class labels. The scatter plot visualization strategy has been selected as it represents a human-interpretable projection of the latent space, enabling intuitive interventions by moving points and clusters in a well-established drag-and-drop manner [42]. Apart from color-coding the individual classes, we add the class cluster centers as colored crosses and mark misclassifications with a black border around the dots. As a main modification functionality, either individual dots can be moved or the whole cluster of a class can be moved by dragging the cluster center cross.

As illustrated in Figure 4, the interaction loop consists of three key stages: (1) initial state, where the model generates an unstructured latent representation as its initial representation as part of the traditional training; (2) the human-model interaction, where users intuitively guide the model's representation by clustering, spreading, or repositioning data points depending on the individual strategy; and (3) the adapted representation, where the model updates its internal structures to incorporate the human adjustments. This feedback loop allows human intuition to serve as a guiding force rather than a manipulative intervention, ensuring that the model does not merely overfit to user-imposed patterns but instead refines its latent space meaningfully.

4. Evaluation

We evaluate HILL using a mixed-methods approach collecting qualitative feedback through observing tool usage and interviewing participants as well as encouraging them to voice their thoughts (think-aloud) during the experiment. Additionally, we administer established questionnaires on system usability (UMUX-Lite [27]) and perceived workload (NASA-TLX [16,15]) complemented by custom questions (Table 1) aimed to gauge



Number of Epochs: 100

Alpha: 0

Evaluation Frequency (batches): 100

Number of features for plotting high dim: 10

Pause after every N epochs: 20

Training...

```
[Epoch 1, Batch 200] Loss: 3.108, Accuracy: 15.65%
[Epoch 1, Batch 300] Loss: 3.023, Accuracy: 18.20%
```

Interactive Scatter Plot of Latent Layer:
- Dots represent data points
- Colors Indicate class
- Black border around dot highlights misclassification
- Colored crosses indicate current center of class

Sidebar to monitor and control the training progress and validation metrics

Plot controls for inspection and interaction

Legend to reference class color and label

Figure 3. The user interface of HILL; sidebar on the left to control the model training through the tool; main window obtaining the interactive scatter plots with relevant controls; a legend on the right to reference class labels with colors.

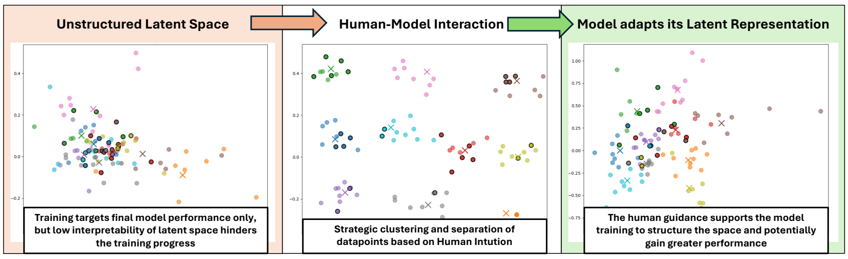


Figure 4. The extracted latent space of an exemplary training iteration. The model's latent space is initially unstructured and the model struggles to separate the classes properly, whereas after the insertion of human guidance, the model adapts its internal representation and classes can be distinguished with greater accuracy.

the users' understanding of HILL as well as its ability to support users in their task. Finally, we compare the traditionally trained models with the models trained during the user study, to outline changes in classification performance in comparison with the applied interaction strategies.

Table 1. Custom questions, targeting the experience with the system. All rated on a visual analog scale (VAS) from 0 to 100; strongly disagree to strongly agree.

Custom questions targeting the experience with the system.	
Q1	The system supported me in finding an optimal ML model.
Q2	The system distracting me during my task.
Q3	The system allowed me to understand the effectiveness of my model.
Q4	The system biased me.
Q5	I trust my model to perform.
Q6	I fully understand the visualization of my model.
Q7	The system made me understand the weaknesses of my model.

Our study consisted of two scenarios in which the participants were tasked to fine-tune a given prediction model. The two scenarios are based on the publicly available datasets CIFAR-10 [23] and PAMAP2 [39]. CIFAR-10 is an image dataset and contains low-resolution images across 10 classes, while PAMAP2 is a time-series dataset capturing human activities from wearable sensors across 13 classes. The model for CIFAR-10 consists of five convolutional layers with ReLU activation and dropout, interleaved with max-pooling layers, followed by four fully connected layers, including a final classification layer. For PAMAP2, the model consists of three 1D convolutional layers with dropout, followed by global max and average pooling and a fully connected classification head. The visualizations for both were generated based on the intermediate layer output after the convolutions before passing the data into the fully connected layers. At this stage, the latent representation obtains the greatest insight while still not being affected by the final classification. For both scenarios, participants were asked to use HILL to find an optimal model to predict the classes. We did not set a model performance goal to achieve nor did we advise participants on possible strategies. Participants were only given a short introduction into HILL, highlighting its features and interaction options as well as explaining the latent space visualization as presented in Section 3.3.

4.1. Procedure

After providing informed consent, we asked participants to provide demographics and their experience with ML models. A total of 14 participants (age: $\bar{x} = 27.1\,y$, $s = 3.8\,y$; 5 female, 9 male) took part in our study. Participants reported an average experience with ML at $\bar{x} = 52.9$ ($s = 28.8$, 0 to 100). Subsequently, we introduced the participants to HILL and randomly (counter-balanced) chose a starting scenario (either CIFAR-10 or PAMAP2). In order to to prevent excessive experiment durations, we pretrained both models for 25 epochs. With that, we cover two different scenarios, since CIFAR-10 still has great potential for improvement at this stage while PAMAP2 presents a challenge with less potential for improvement due to the complex nature of sensor data. We asked participants to voice their thoughts during the interaction and tasked them to improve the performance of the given model. We did not put any interaction constraints on this step and allowed participants to freely explore HILL and strategies they saw fit to improve the model performance. However, to receive comparable results, we fixed four interaction points in the training, respectively at epochs 25, 30, 35, and 40. After each training iteration (initiated by the participants at will), the updated model performance was displayed and the respective latent space visualization was updated. After each full pass of a scenario, we administered the UMUX-Lite, NASA-TLX, and our custom questions. Participants subsequently were tasked with the second, remaining scenario. At the end of the study, we interviewed the participants on their impressions of the system. The study duration did not exceed 60 min and was approved by the Ethics Team of the German Research Center for Artificial Intelligence.

4.2. Results

We analyzed the administered questionnaires and the recorded think-aloud protocols including the interviews to evaluate the user experience of HILL. We particularly focused on distilling the strategies to improve the model employed by the participants. A comparison of the model performance using HILL and baseline training without human intervention provides insight into the effectiveness of our approach.

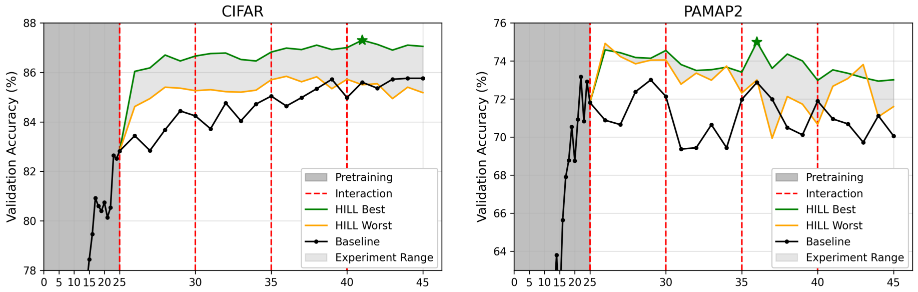


Figure 5. The model performance evaluation of HILL compared to the traditional training as baseline; Dark grey represents the pretraining phase; red dashed lines human interaction points; the light grey area represents the range of best (green) and worst (orange) participant utilizing HILL during the user study.

4.2.1. Model Performance

We present the validation accuracy for the two scenarios in Figure 5, comparing the traditional training (baseline) with HILL. We visualize the range of participants' experiments of HILL, including the best-performing experiment in *green* and the worst in *orange*. The evaluation was conducted on the unseen validation dataset, the same as during training, to ensure that the model does not overfit the test data.

Across both datasets, the first interaction had the most significant impact on performance. This can be attributed to the user's initial structuring of the latent space, which often resulted in a great boost of accuracy. Even the worst-performing HILL results were still capable of outperforming the baseline, demonstrating the robustness of the approach. Notably, the best-performing HILL runs significantly surpassed the baseline, providing strong evidence that human intuition can effectively enhance training. For CIFAR, the accuracy increased by 1.6% points to 87.3%, while for PAMAP2, the improvement was around 2.2% points to 75%. Additionally, HILL led to faster convergence, reducing the computational resources required to achieve satisfactory classification performance. In the case of PAMAP2, the worst HILL performance can be traced back to a participant who altered their strategy across interaction points. This change resulted in greater fluctuations in accuracy, particularly during the last two interactions.

4.2.2. Questionnaires

We calculated the SUS-parity score ($max = 100$) from the collected UMUX-Lite responses [28]. Our system showed good usability [2] with $\bar{x} = 72.4$ ($s = 9.23$). Likewise, the total NASA-TLX ($max = 120$) indicated a low workload for participants with $\bar{x} = 36.1$ ($s = 18.2$). Figure 6 depicts the individual subscales ($max = 20$) given our two scenarios. Physical and temporal demand is very low, while other subscales score slightly higher with bigger variance. In particular, frustration was lower for PAMAP2.

The results of our custom questions (Table 1) are visualized in Figure 7. For Q2 (”The system distracting me during my task.”) and Q4 (”The system biased me.”) we recorded low ratings from the participants across both scenarios indicating that the system was neither distracting the user ($\bar{x} = 12.6$, $s = 18.1$) nor biased their decision making ($\bar{x} = 23.1$, $s = 28.1$). For all other questions, we recorded split ratings, associated with whether participants were successful in improving the model with their strategies. This affected their model beliefs, such as its effectiveness (Q3), its weaknesses (Q7), and their trust in the model (Q5). Notably these beliefs also translated to an understanding of the

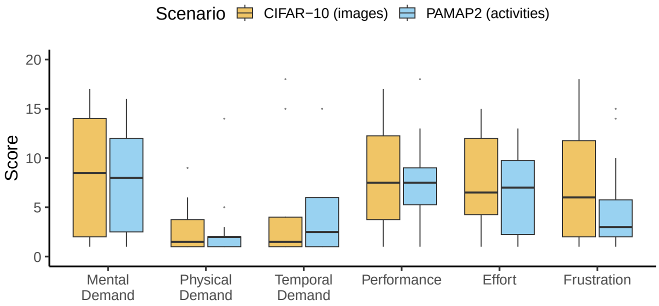


Figure 6. Scores for NASA-TLX subscales (0 to 20) given our two scenarios (CIFAR-10, PAMAP2).

visualization (Q6) and whether the system was able to support them in finding an optimal model (Q1). In particular, for Q1, we found that scores varied less for the PAMAP2 scenario, also linking to its lower frustration score (NASA-TLX).

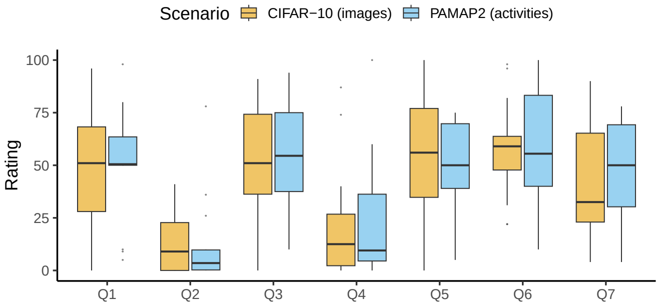


Figure 7. Ratings for our custom questions (Table 1) given our two scenarios (CIFAR-10, PAMAP2).

4.2.3. Think-aloud and Interviews

We transcribed the audio files (10 hours 30 minutes in total) collected from each participant using Whisper [38]. To analyze the interviews, we used the approach by Blandford et al. [3]. Three researchers coded an initial 20% of the data and agreed on an initial coding tree. The remaining data was split between the researchers. From a final discussion, the following themes surfaced: *Strategies*, *Human-Model-Interaction*, and *User Interface*. Since *Strategies* was the main topic for most participants during the study we subdivided that topic into six subtopics.

Strategies Most participants agreed on moving data points closer to their cluster center in order to remove misclassified outliers, therefore **Increasing Cluster Compactness**.

> *”Now my strategy is to cluster them near to the exact center.” (P13)*

Another commonly applied strategy involved **Maximizing Cluster Distance**, where participants moved whole clusters further apart from each other, in an effort to help the model distinguish and separate them properly.

> *”Let’s get the airplane away from the bird.” (P14)*

Contrary to moving clusters apart, some participants applied the **Merge Similar Classes** strategy where they decided to move classes connected on a high abstraction level closer.

> *”(...) feature one, you have animals, and then kind of sort them. And on the other side, like, cars, automobiles, like transportation things.” (P14)*

In cases where participants were shown dense, well-separated clusters with a clear distinction from others, they commonly employed the **Keep Arrangement** strategy, only moving points from classes where this was not the case.

> *"I think the automobile is great. So I'll not move this at all." (P4)*

When the selected strategy did not produce satisfying improvements, a few participants **Changed Strategy** and decided to continue training by inverting their previous approach.

> *"I would like to try something completely different. Before I purposefully tried to separate classes that are similar. But what happens if I do the exact opposite?" (P6)*

Human-Model-Interaction. Apart from diverse strategies, most of the participants gave qualitative feedback about their impressions working directly with the model during training. Most feedback was positive, showing that HILL helps understanding the inner workings of the trained models.

> *"It's not so much a black box anymore" (P2)*

Yet, some participants wondered about the lack of influence they had on the training.

> *"I wanted to see how the tool reacts, if it helps with training at all or not. So I tried a few different things and I think for the first task this worked somewhat, for the second task really not." (P6)*

User Interface Overall, participants appreciated features and layout of the interface. However, some required more feedback if their actions had the desired impact.

> *"The system is easy to use, but I'm obviously doing something wrong." (P2)*

5. Discussion

A key challenge in the field of HITL ML is that users bring their own preconceived notions into the interaction, which can subtly shape both their engagement with the system and the resulting model behavior. As indicated by our interviews (Section 4.2.3), participants often approached latent space modifications based on their own mental models of the data, sometimes **reinforcing subjective structures rather than optimizing for pure algorithmic generalization**. This phenomenon links to developer blindness, where those designing or interacting with such systems impose their own interpretations into the model, overlooking alternative perspectives or emergent patterns [21].

Unlike related approaches that modify training data directly (cf. Section 2), HILL operates within the latent space, only including it in the loss function, hence, making it less susceptible to overfitting caused by human biases as evident from our analysis on model performance (Section 4.2.1). However, guidance should not turn into manipulation, since allowing too much user influence can lead to the model conforming to subjective strategies rather than generalizable improvements [7]. Especially when evaluating the model performance, users need to ensure to exclude their own beliefs and intuitions to properly **balance human intuition with algorithmic training**.

Human users interpret data differently than algorithms, often applying concepts based on their knowledge about the classes, such as grouping or separating "birds" and "planes" since they are correlated with the sky. This can enrich training by **introducing structured insights from human cognition**, which purely data-driven approaches overlook. Our results show that human-induced adjustments improved model performance without causing overfitting (Section 4.2.1), suggesting that human guidance can high-light relevant decision boundaries that standard optimization fails to capture. Future work should explore adaptive strategies where the model learns which human interventions are beneficial and adjusts its reliance accordingly.

A recurring topic in user feedback was the need for more **transparent algorithmic responses to the user**, apparent in our questionnaires (Section 4.2.2) and interviews (Section 4.2.3). Some participants struggled to gauge whether their interactions were meaningful, which aligns with findings from our think-aloud analysis. Addressing this requires more explicit feedback mechanisms beyond pure performance metrics, such as visualizing the differences between latent spaces before and after intervention. Providing explanatory cues, for instance highlighting which modifications or classes contributed to accuracy improvements, or even utilizing support from Large Language Models [24], could further bridge the gap between human intuition and model behavior.

Traditional training optimization techniques focus on minimizing a loss function defined over the training dataset, which encapsulates the differences between predicted and actual outcomes within that specific context. While these methods are effective for achieving reasonable performance on the given data without human intervention, they operate within the confines of the information provided by the dataset. Human intuition, on the other hand, draws upon a vast reservoir of world knowledge, experiences, and contextual understanding that extends far beyond the boundaries of any single dataset. This intuitive knowledge includes common sense reasoning and domain-specific expertise. As such, **human intuition has the ability to recognize patterns or relationships that may not be explicitly represented in the data** but are crucial for real-world applications.

6. Conclusion

In this work, we introduced ”Human in the Latent Loop” (HILL), facilitating a novel human-AI interaction paradigm that interactively infuses human intuition into model training through latent space representations. HILL demonstrated that human interventions improve model performance and convergence while maintaining generalization. Our evaluation revealed diverse interaction strategies, highlighting both the strengths as well as the risks of potential human biases. By integrating human intuition into the model training process, particularly through interactive methods such as manipulating latent space clusters, HILL can effectively bridge the gap between data-driven optimization and human-centric understanding. This approach enables the model to benefit from insights that traditional methods might overlook, potentially leading to solutions that not only perform well on the dataset but also make sense in broader, real-world contexts.

Acknowledgement

This work is supported by the European Union’s Horizon Europe research and innovation program (HORIZON-CL4-2021-HUMAN-01) through the ”SustainML” project (grant agreement No 101070408).

References

[1] Tobias Baer and Vishnu Kamalnath. Controlling machine-learning algorithms and their biases. *McKinsey Insights*, 2017.

[2] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[3] Ann Blandford, Dominic Furniss, and Stephann Makri. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers, 2016.

[4] Chengliang Chai and Guoliang Li. Human-in-the-loop techniques in machine learning. *IEEE Data Eng. Bull.*, 43(3):37–52, 2020.

[5] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 429–440, 2021.

[6] Vivek Choudhary, Arianna Marchetti, Yash Raj Shrestha, and Phanish Puranam. Human-ai ensembles: When can they work? *Journal of Management*, 51(2):536–569, 2025.

[7] Pedram Daee, Tomi Peltola, Aki Vehtari, and Samuel Kaski. User modelling for avoiding overfitting in interactive knowledge elicitation for prediction. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pages 305–310, 2018.

[8] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872, 2021.

[9] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45, 2003.

[10] Daniel Geißler, Bo Zhou, and Paul Lukowicz. Latent inspector: An interactive tool for probing neural network behaviors through arbitrary latent activation. In *IJCAI*, pages 7127–7130, 2023.

[11] Daniel Geissler, Bo Zhou, and Paul Lukowicz. Strategies and challenges of efficient white-box training for human activity recognition. *arXiv preprint arXiv:2412.08507*, 2024.

[12] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. Towards human-guided machine learning. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 614–624, 2019.

[13] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[15] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[16] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[17] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[18] Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain informatics*, 3(2):119–131, 2016.

[19] Johannes Jakubik, Jakob Schöffer, Vincent Hoge, Michael Vössing, and Niklas Kühl. An empirical evaluation of predicted outcomes as explanations in human-ai decision-making. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 353–368. Springer, 2022.

[20] Christian Janiesch, Patrick Zschech, and K. Heinrich. Machine learning and deep learning. *Electronic Markets*, 31:685 – 695, 2021.

[21] Johanna Johansen, Tore Pedersen, and Christian Johansen. Studying the transfer of biases from programmers to programs. *arXiv preprint arXiv:2005.08231*, 2020.

[22] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[24] Lars Krupp, Jonas Bley, Isacco Gobbi, Alexander Geng, Sabine Müller, Sungho Suh, Ali Moghiseh, Arcesio Castaneda Medina, Valeria Bartsch, Artur Widera, et al. Llm-generated tips rival expert-created tips in helping students answer quantum-computing questions. *arXiv preprint arXiv:2407.17024*, 2024.

[25] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[27] James R Lewis, Brian S Utesch, and Deborah E Maher. Umux-lite: when there's no time for the sus. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2099–2102, 2013.

[28] James R Lewis, Brian S Utesch, and Deborah E Maher. Investigating the correspondence between umux-lite and sus scores. In *Design, User Experience, and Usability: Design Discourse: 4th International Conference, DUXU 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings, Part I*, pages 204–211. Springer, 2015.

[29] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and D. Dou. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64:3197 – 3234, 2021.

[30] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[31] Bahar Mahmud, Guan Hong, and Bernard Fong. A study of human–ai symbiosis for creative work: Recent developments and future directions in deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–21, 2023.

[32] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[33] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI.* Simon and Schuster, 2021.

[34] Sugeerth Murugesan, Sana Malik, F. Du, Eunyee Koh, and T. Lai. Deepcompare: Visual and interactive comparison of deep learning model performance. *IEEE Computer Graphics and Applications*, 39:47–59, 2019.

[35] Rohan Paleja, Michael Munje, Kimberlee Chang, Reed Jensen, and Matthew Gombolay. Designs for enabling collaboration in human-machine teaming via interactive and explainable systems. *arXiv preprint arXiv:2406.05003*, 2024.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[37] Tehreem Qamar and N. Bawany. Understanding the black-box: towards interpretable and reliable deep learning models. *PeerJ Computer Science*, 9, 2023.

[38] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[39] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[41] Gonesh Chandra Saha, Sanjay Kumar, Avinash Kumar, Hasi Saha, TK Lakshmi, and Niyati Bhat. Human-ai collaboration: Exploring interfaces for interactive machine learning. *Tuijin Jishu/Journal of Propulsion Technology*, 44(2):2023, 2023.

[42] Bruno Schneider. Visual integration of model and data spaces in classification problems. 2023.

[43] Burr Settles. Active learning literature survey. 2009.

[44] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Are bias mitigation techniques for deep learning effective? *arXiv e-prints*, pages arXiv–2104, 2021.

[45] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich,

Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies*, 67(8):639–662, 2009.

[46] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. Evolution and impact of bias in human and machine learning algorithm interaction. *Plos one*, 15(8):e0235502, 2020.

[47] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019.

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[49] Hao Wang, Snehasis Mukhopadhyay, Yunyu Xiao, and Shiaofen Fang. An interactive approach to bias mitigation in machine learning. In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 199–205. IEEE, 2021.

[50] Jiafu Wei, Ding Xia, Haoran Xie, Chia-Ming Chang, Chuntao Li, and Xi Yang. Spaceediting: Integrating human knowledge into deep neural networks via interactive latent space editing. *arXiv preprint arXiv:2212.04065*, 2022.

[51] Christine Wolf and Jeanette Blomberg. Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[52] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In *Proceedings of the second workshop on data management for end-to-end machine learning*, pages 1–4, 2018.