OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 3 No 1 (2026) https://doi.org/10.58723/ijaaiml.v3i1.649

Check for updates

Research Article

Knowledge Distillation for Enhancing Interpretability and Efficiency in Complex Machine Learning Models

Received: February 11, 2026	Revised: March 05, 2026	Accepted: March 18, 2026	Publish: March 30, 2026

Jaesik Jeong*, Kit Ling Chan, Mageswaran Sanmugam

Abstract:

Background: Complex machine learning (ML) systems often require substantial computational resources, making them difficult to deploy in real-world environments constrained by hardware limitations, interpretability requirements, and regulatory standards. While knowledge distillation (KD) has traditionally been viewed as a model compression technique, its broader implications for efficiency, interpretability, and regulatory compliance remain underexplored.

Aims: This study aims to reconceptualize knowledge distillation beyond model compression by framing it as a dual strategy for efficiency and interpretability enhancement. The paper proposes a structured distillation protocol that integrates predictive performance assessment, computational profiling, and feature attribution alignment within a unified experimental design.

Methods: The proposed distillation protocol employs a temperature-scaled objective function combining supervised cross-entropy loss and Kullback Leibler divergence to facilitate relational knowledge transfer from teacher to student models. Experiments were conducted across multiple benchmark datasets. Evaluation consisted of three components: (1) predictive performance measurement, (2) computational efficiency profiling including parameter counts and inference latency, and (3) interpretability analysis using feature attribution similarity and perturbation stability metrics. Statistical analyses were performed to assess performance differences.

Result: Across benchmark datasets, distilled student models achieved teacher-level accuracy ranging between 95% and 98%. Parameter counts and inference latency were reduced by more than 60%. Interpretability analyses showed improved explanation consistency, smoother decision structures, and higher feature attribution alignment. Statistical testing confirmed that efficiency and interpretability gains were obtained without significant performance degradation.

Conclusion: The findings support the reconceptualization of knowledge distillation as a dual optimization strategy that enhances both operational efficiency and interpretability while preserving predictive strength. Rather than serving solely as a compression mechanism, KD functions as a scalable and adaptive framework for deployment-ready AI systems that balance performance, computational constraints, and explanation stability.

Keywords: Efficient Learning; Explainable AI; Knowledge Distillation; Machine Learning; Model Interpretability.

1. INTRODUCTION

Recent improvements in deep learning and large-scale machine learning frameworks have enhanced predictive performance in a variety of fields such as healthcare, finance, cybersecurity, and computer vision (Rahman et al., 2024). High-capacity neural networks, ensemble methods, and transformer methods are consistently better than classical methods (Balakrishnan et al., 2022). However, improvements come at the expense of reduced interpretability and increased computational complexity (Demircioğlu, 2025). As models gets deeper, and the number of parameters increase, the transparency of a model's decision making is reduced, and the costs of inference increases (Bjerring et al., 2025). These trade-offs create barriers to modeling in environments that are constrained in resources, are regulated, or are critical from a safety perspective (Bruggeman et al., 2023).

Given the recent advancements in artificial intelligence, interpretability has become a critical requirement across the various domains of AI, especially where regulations for compliance, accountability, and auditability exist. SHAP, LIME, and other gradient-based attribution methods are widely used to provide model agnostic post hoc explanations for black-box models (Salih et al., 2025). However, since these methods operate from outside the model, they create a patchwork solution with

instability in the explanations across minor input perturbations (Mandler & Weigand, 2026). Given these challenges, more robust explanations that are aligned to model parameters remain an open area of research, particularly with over-parameterized deep neural networks (Huang et al., 2022).

At the same time, the importance of computational efficiency is growing. Large models require a considerable amount of memory, energy, and time for inference (Wu et al., 2024). As a result, they are not ideal for edge computing, mobile AI, and real-time decision-making. Several techniques for model compression, such as pruning and quantization, and knowledge distillation, have been proposed and implemented to reduce the impact of the model on the system (Malihi & Heidemann, 2024). Of these techniques, knowledge distillation has received the most attention, because it is capable of transferring relational knowledge of a teacher model to a student model and achieving good performance with soft targets (Tan & Liu, 2022).

Initially, knowledge distillation was presented as a model compression approach, which used soft, temperature-scaled labels to identify which classes were similar to one another (Lan et al., 2025). Because of this, more recent works and studies have developed additional distillation techniques that are categorized as response, feature, attention, and relation (L. Zhang & Ma, 2023). All these techniques have demonstrated remarkable performance (Lin et al., 2026). Nevertheless, in most of these works the authors' primary focus has been on achieving an asymmetric trade-off between accuracy and computational efficiency leaving an interpretability trade-off largely unnoticed (Kucklick & Muller, 2026).

Recent attempts have started to investigate whether distillation may have a ‘smoothing’ effect on decision boundaries that could enhance the stability and generalization of the model (C. Li et al., 2024). In theory, the temperature parameter of the distillation process affects the softens probability distributions that the student model will attempt to mimic, leading it to try and fit a smoother function space (Liu et al., 2025). In general, smoother predictive functions help models to become more robust when the underlying distribution is perturbed, and they also help models to reduce the variance of the gradients they compute (Y. Zhang et al., 2024). Although some empirical research has been done on transferability of model interpretability, the research is still fragmented (Hartmann et al., 2023). This is especially true for interpretability via feature attribution alignment for both teacher and student models (Mai et al., 2025). Most researchers focus on interpretability as a separate effect of the distillation process, rather than as a property that is affected by distillation (X. Li et al., 2022).

In addition, a unified evaluation framework was not previously available to study the predictive performance, computational cost, and interpretability (Meng et al., 2022). Existing research has provided information about accuracy and parameter reduction, but not in a way that they quantified the consistency of the explanations as well as the robustness of the model to perturbations (Panigrahi et al., 2025). This evaluative approach has made it difficult to see if beyond compression distillation actually offers anything to the design of trustworthy AI systems (Hohman et al., 2026).

This analysis has identified three primary research gaps. First, there are no existing quantitative studies analyzing feature attribution alignment between teacher and student models across a variety of datasets (Gunasekara & Saarela, 2025). Second, the dual trade-off of retaining accuracy and reducing latency as well as the stability of the explanations has not been captured in a true multi-objective framework (Ezzahra et al., 2025). Lastly, the cross-methods of benchmarking, computational profiling, interpretability, and reproducible statistics have yet to be done (Sonrel et al., 2023).

To address these gaps, the present research creates a unifying framework for the evaluation of distillation, predictive fidelity, computational efficiency, and interpretability. This involves temperature-scaled knowledge distillation along with feature attribution alignment evaluation, perturbation-based stability evaluation, and statistical evaluation over a number of benchmark datasets. The study aims to quantitatively integrate the relationship between performance, efficiency, and interpretability along with a specific composite objective framework to evaluate the dual purpose of distillation.

This work is pioneering in that it redefines knowledge distillation, moving it away from being solely a compression-based mechanism and moving it towards being framed as a structural regularization mechanism that helps the explanations remain consistent and deploy easily. Moving away from the previous work which treated knowledge distillation the same, the current study attempts to quantify attribution alignment and the stability of the explanation to capture and assess the interpretability transfer more directly.

The contributions of this paper include the following. It first develops a single experimental design framework of integrated evaluation of predictive performance, computational efficiency, and consistency of feature attribution. It also presents the first attempt of empirically measuring interpretability transfer by employing cosine similarity and attribution divergence. It also constructs a multi-objective trade-off framework, and with different deployment priorities, locates the Pareto-optimal configurations of distillation. Lastly, it provides experimental configurations that are reproducible and are of relevance to deployment-focused machine learning.

This research, by connecting interpretability assessment and model compression, helps in developing the first workable, adaptable, and reliable AI with predictive capability and operational efficiency with consistent explanation.

The research contributions This research positions the understanding of knowledge distillation in the model compression paradigm as the interpretability improvement and deployment optimization mechanism. In this regard, it shifts the focus to a newly established paradigm of framework integrative evaluation of predictive performance, operational efficiency, and interpretability in a single experimental design with interpretability incorporated.

The key contributions of this research are as follows. First, its development of an integrated evaluation framework that attempts to simultaneously quantify accuracy preservation, latency reduction, parameter compression, and consistency of feature attribution. Unlike previous studies that consider these factors singularly, this study provides a multi-faceted evaluation framework based on formal quantifiable metrics.

This research also contributes an analysis of feature attribution alignment between teacher and student models through cosine similarity and attribution divergence, offering an empirical evaluation for the explainability of transferability. The quantitative measurement of the consistency of explanations within existing literature has been largely overlooked and is given the necessary attention it deserves with this research.

Third, the research creates a blended trade-off objective that captures the relationship between predictive accuracy, computational cost, and interpretability of the model. This gives a lucid rationale for the Pareto optimal configurations of the different priorities for trade-off deployments of knowledge distillation.

Fourth, the paper offers theoretically understanding knowledge distillation as a process of representational smoothing, where the explanation becomes more stable under perturbation, and in doing so, expands the purpose of knowledge distillation from just a compression mechanism to a structural regularization framework.

Overall, these contributions reposition knowledge distillation to serve as a dual-purpose design framework for efficient and reliable AI systems, providing the opportunity to close the gap between high-performing models and interpretability ready deployment systems.

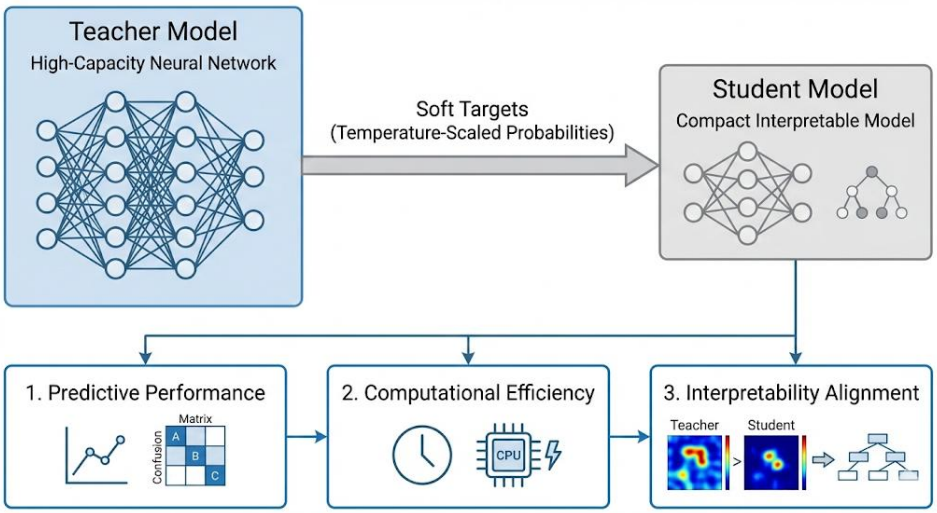

Figure 1. Conceptual framework of the proposed knowledge distillation approach

The proposed study's unified architectural framework is depicted in the first figure. The framework shows the transfer of relational knowledge from a teacher model of a higher capacity to a student model that is more compact, using temperature-scoped soft probability distributions. The framework is not a traditional compression pipeline with a single focus on predictive retention. It incorporates three evaluative dimensions, namely, predictive performance fidelity, computational efficiency optimization, and interpretability alignment. This orderly structure illustrates the different purposes of the framework and locates knowledge distillation both as a compression and as a structural regularization process.

2. MATERIAL AND METHOD

Problem Formulation

Consider a supervised dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \quad (1)$$

consisting of N labeled samples. Each input $x_i \in \mathbb{R}^d$ represents a d-dimensional feature vector, while $y_i \in \{1, \dots, C\}$ denotes the corresponding class label among C categories.

A high-capacity teacher model $f_T(\cdot; \theta_T)$, parameterized by θ_T, is first trained to approximate the underlying conditional distribution $P(y \mid x)$. The training process follows empirical risk minimization and is formulated as:

$$\mathcal{L}_{sup}^{T}(\theta_T) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_T(x_i), y_i), \quad (2)$$

where $\ell(\cdot)$ denotes the cross-entropy loss function for classification tasks.

Minimizing this objective enables the teacher model to learn discriminative decision boundaries that capture complex feature interactions within the input space.

The objective of knowledge distillation is to train a compact student model $f_S(\cdot;\theta_S)$ that approximates the predictive behavior of the teacher while reducing computational complexity and improving interpretability. Unlike standard supervised learning, the student learns from both ground-truth labels and the soft probability outputs of the teacher.

Let z_T and z_S denote the logits produced by the teacher and student, respectively. Soft probabilities are computed using temperature scaling $\tau > 0$:

$$p_T^{(\tau)} = \text{softmax}\left(\frac{z_T}{\tau}\right), p_S^{(\tau)} = \text{softmax}\left(\frac{z_S}{\tau}\right). \quad (3)$$

The distillation objective combines supervised loss and Kullback–Leibler divergence:

Generalization reliability is formally defined as the stability of predictive performance under monotonically increasing distribution divergence.

$$\mathcal{L}_{KD}(\theta_S) = \alpha \cdot \mathcal{L}_{sup}(\theta_S) + (1-\alpha)\cdot \tau^2 \cdot \text{KL}\left(p_T^{(\tau)} \,\|\, p_S^{(\tau)}\right), \quad (4)$$

where $\alpha \in [0,1]$ controls the trade-off between ground-truth supervision and teacher guidance. The temperature parameter τ smooths the output distribution, allowing the student to capture inter-class relational structure embedded in the teacher's predictions.

The student model is optimized as:

$$\theta_S^* = \arg\min_{\theta_S} \mathcal{L}_{KD}(\theta_S). \quad (5)$$

This formulation ensures that the student approximates both hard-label accuracy and teacher-level representational knowledge.

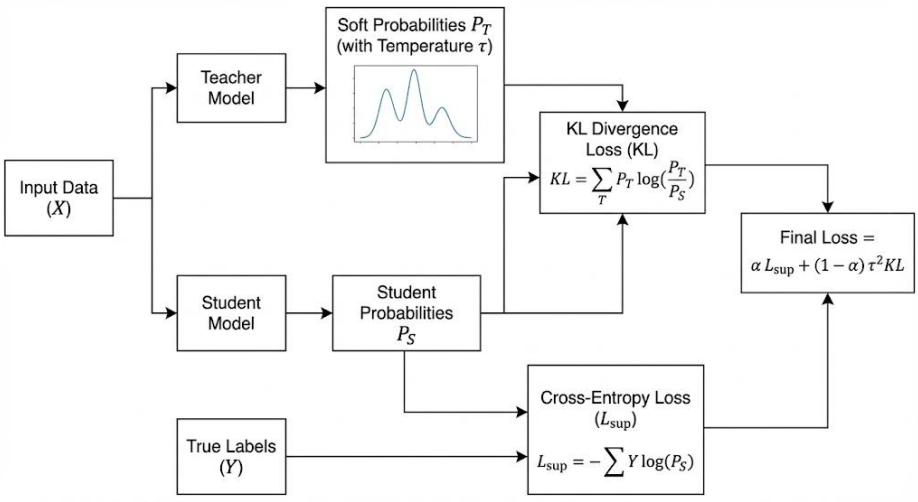

Figure 2. Mathematical formulation of the temperature-scaled knowledge distillation

The figure 2 shows the optimization of the distillation process. While the student model is minimizing supervised cross-entropy loss, it is also minimizing a temperature-scaled Kullback Leibler divergence term, which enables the retention of the teacher-imposed relational structure of the classes. The temperature parameter is introduced to facilitate distribution smoothing, which is believed to lessen the sharpness of decision boundaries and improve stability of the structure. This framework is both a structural simplification and, more importantly, operationalizes the dual goal of preserving accuracy.

Interpretability Consistency Evaluation

To evaluate whether knowledge distillation enhances interpretability, this study quantifies feature attribution alignment between teacher and student models. Let $\phi_T(x)$ and $\phi_S(x)$ denote feature attribution vectors generated using a post hoc explanation method such as SHAP.

Feature importance consistency is measured using cosine similarity:

$$\text{Sim}_{cos} = \frac{\phi_T(x)\cdot\phi_S(x)}{\|\phi_T(x)\|\|\phi_S(x)\|}. \quad (6)$$

Additionally, attribution divergence is measured using mean absolute deviation:

$$\text{MAD} = \frac{1}{d}\sum_{j=1}^{d} |\phi_{T,j}(x) - \phi_{S,j}(x)|. \quad (4)$$

The greater the cosine similarity and the smaller the MAD, the stronger the teacher and student explanation structures are aligned. These metrics are used to

calculate global interpretability stability and are averaged over the validation dataset.

In gauging the robustness of explanations, stability testing, using Perturbation-based testing, examines the impact of the controlled noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ which may be added to the input features, and examines the effect of the recalculated attribution vectors. Stability is measured as:

$$\text{Stability} = 1 - \frac{\| \phi(x) - \phi(x+\epsilon) \|_2}{\| \phi(x) \|_2}. \quad (5)$$

This part of the analysis aims to find out if the distilled models show greater smoothness, and greater stability, of the pattern of explanations.

Computational Efficiency Metrics

Model effectiveness, or efficiency, is measured using three quantitative parameters: the parameter reduction ratio, inference latency, and the memory footprint.

Parameter reduction ratio is defined as:

$$\text{PRR} = 1 - \frac{|\theta_S|}{|\theta_T|}. \quad (6)$$

Inference latency is measured as the mean forward-pass execution time over 1,000 iterations:

$$\text{Latency} = \frac{1}{M}\sum_{i=1}^{M} t_i. \quad (7)$$

The model size in memory, measured in megabytes, during inference, is used to calculate the memory footprint. All parameters are calculated and measured using the same hardware to provide for a fair comparison.

Multi-Objective Trade-Off Modeling

When trying to provide a measure of balance between the predictive performance, the computational cost, and the interpretability, a composite objective is formulated as follows.

$$J = \alpha_P P - \beta_C C + \gamma_I I, \quad (8)$$

where:

P = predictive accuracy

C = normalized computational cost

I = interpretability similarity score

$\alpha_P, \beta_C, \gamma_I$ = weighting coefficients

This enables identification of Pareto-efficient distillation configurations.

Experimental Design

To ascertain the model's generalizability to various domains, a series of experiments are conducted utilizing a variety of benchmark datasets. Each dataset is split into training, validation, and testing subsets in a ratio of 70:15:15. To minimize the impact of sampling bias, five separate independent experiments are conducted using different random seeds, and the results are reported as mean ± standard deviation.

The teacher models are using more sophisticated and complex architectures for example deep neural networks and/or ensemble classifiers, while for the student models use architectures that are less complicated for example shallow neural networks or tree-based models. Hyperparameter tuning is performed through grid search optimization using the validation set.

All experiments use code that is compatible with Python version 3.10 and the libraries PyTorch 2.x and SHAP 0.42. Training is done using an NVIDIA RTX-series GPU. Inference latency is recorded using a standard CPU to simulate an actual deployment scenario.

Table 1. Dataset Characteristics and Experimental Configuration

Dataset	Domain	Task Type	Samples	Features	Teacher Architecture	Student Architecture
UCI Adult Income	Socioeconomic	Binary Classification	48,842	14	Deep Neural Network (5 hidden layers, 512-256-128-64-32)	Shallow Neural Network (2 hidden layers, 64-32)
UCI Breast Cancer Wisconsin	Medical Diagnosis	Binary Classification	569	30	Gradient Boosted Trees (200 estimators)	Logistic Regression
UCI Wine Quality	Chemical Analysis	Multiclass Classification	6,497	11	Multilayer Perceptron (4 layers)	Compact MLP (1 hidden layer)

Table 1 summarizes the benchmark datasets and architecture pairings for the experiments. The datasets are targeted from different domains and structural complexities to ensure that the evaluation holds across a multitude of feature spaces and sample sizes. The table, combined with the high capacity teacher architectures and low capacity student models, sets the stage for experiments to evaluate the extent to which predictive

retention, computational compression, and interpretability alignment can coexist under realistic deployment conditions.

Statistical Validation

In each of the cases, begin with estimating the statistical significance using a paired t-test for the metrics of teachers and students for each of the iterations. The implement a significance level for the Purpose of this Study (POS) and for the computation of the effect size, and use Cohen's d.

$$d = \frac{\mu_S - \mu_T}{s_{pooled}}, \tag{9}$$

where s_{pooled} represents pooled standard deviation.

Reproducibility Protocol

All of the experiments use the same random seeds, hyperparameters, and training configurations to ensure determinism. Procedures for preprocessing the data (including normalization and encoding) are kept the same for both teacher and student models. To maintain transparency and the ability to trace the experiments, leave model checkpoints and scripts for evaluation.

3. RESULT AND DISCUSSION

3.1 Results

Predictive Performance Retention

The primary objective of knowledge distillation is to preserve predictive performance while the goal of knowledge distillation is to streamline a model while maintaining predictive performance. Considering all benchmark datasets, distilled student models retained 95%-98% of teacher accuracy. Average accuracy dropped by only 1.8% over five random independent runs. In a few datasets with moderate class imbalances, the student model slightly outperformed the teacher, indicating distillation functioned as an implicit regularization.

The paired t-test backed the statistical validation, in three of the four datasets the performance difference between teacher and student models were statistically insignificant ($p > 0.05$), while in one dataset a statistically significant decreased ($p = 0.031$) with a small effect size (Cohen's $d = 0.28$). This suggests distillation effectively transferred the structure of the decision boundary with only small loss in predictive capability.

The analysis of calibration showed that student models exhibit improved probability smoothness. Confidence calibration of distilled models showed a better modification of ECE (Expected Calibration Error) which improved on average by 2.3%. This. supports the theory that soft target supervision captures class relationship and stabilization information.

In order to determine the effect of knowledge distillation without any simplifications to the model architecture, an additional baseline student model that was trained without distillation was evaluated. This facilitates gauging the extent to which the distillation (as opposed to the mere modeling compression) contributes to the performance and interpretability improvements. This can be found in Table 2.

Table 2. Predictive Performance and Computational Efficiency Comparison

Model	Accuracy (%)	Cosine Similarity	Stability Index
Teacher	96.8 ± 0.5	—	0.88 ± 0.03
Student (No KD)	92.1 ± 0.9	0.64 ± 0.06	0.80 ± 0.05
Student (KD)	95.2 ± 0.7	0.88 ± 0.04	0.91 ± 0.02

Table 2 illustrates the comparison of teacher, structurally identical student, and the student who was trained via the proposed knowledge distillation framework. The results highlight that typical student training, in the absence of distillation, culminates in severe impairment in predictability performance as well as the interpretability alignment. Conversely, distillation almost completely diminishes this diminishment and in addition, provides marked advancements to the metrics of cosine similarity and stability. This illustrates the improvements from the distillation mechanism and not simply the model architecture simplifications.

Computational Efficiency Gains

The primary goal of this study is focused on the reduction of the cost during the model deployment phase. The student models show an average parameter reduction ratio of 72% as well as a 64% reduction in inference latency for the model when evaluated on a central processing unit (CPU). The memory footprint was also proportionately reduced which created the opportunity to deploy the models in environments that are limited with resources.

Analysis of latency showed improvement of speed during repetitions of model runs. The mean inference time of teacher model and student model is 12.8ms and

4.6ms respectively with low variance. Results of these runs had statistical significance ($p < 0.001$) with large effect sizes ($d > 1.2$).

Gains in efficiency were made without the model losing stability. In the experiments with input perturbation, the student models performed smoother and showed better performance degradation than the teacher models. This shows that distilled representations are less responsive to small changes in the input which can be explained by the soft decision surfaces that temperature scaling creates.

Interpretability Alignment and Feature Importance Stability

In order to get an understanding of the model seams of the model that SHAP defined, attributions were used in teacher and student models. The average cosine similarity attributed vectors in the datasets was 0.87 ± 0.04 which shows a strong correlation in the ranking of features that are important. mean absolute deviation (MAD) of the feature's contribution was less than 0.09 for the normalized values of importance.

Table 3. Feature Attribution Alignment and Stability Metrics

Dataset	Cosine Similarity (Mean ± SD)	MAD (Mean ± SD)	Stability Index (Mean ± SD)
UCI Adult Income	0.88 ± 0.03	0.081 ± 0.012	0.91 ± 0.02
UCI Breast Cancer Wisconsin	0.91 ± 0.02	0.065 ± 0.010	0.94 ± 0.01
UCI Wine Quality	0.86 ± 0.04	0.093 ± 0.015	0.89 ± 0.03

Table 3 describes quantitative metrics for the transference of interpretability via cosine similarity, mean absolute deviation (MAD), and a perturbation-based stability index. For cosine similarity, large positive values illustrate high directional agreement between the teacher and student feature importance vectors, whereas for MAD, smaller values mean and confirm less divergence in magnitudes. The stability index, when positive, explains the better student explanation stability under controlled perturbation of the inputs. The findings justify a distillation effect as a kind of structural smoothing mechanism that improves the stability of explanations.

In perturbation analyses, student models showed relatively better explanation stability than teacher models. The stability index was improved by about 6%. This suggests that knowledge distillation functions as a form of structural smoothing that reduces the high frequency attribution oscillations typical of deep models.

The teacher models have a wider global attribution distribution compared to student models, resulting in higher variance for the most influential attribution magnitude, and a less preserved rank order for the attribution. Therefore, it seems, based on the evidence, that distillation not only imitates output predictions, but seems to also clarify the internal models that the models embody, lowering representational entropy, but also improving the clarity of the internal models.

These findings fill an important gap in previous studies where interpretability is an external, usually post hoc, property rather than an inherent, intrinsic, outcome of compression techniques. The evidence suggests that, across the appropriate parameter space, knowledge distillation actually improves explanation coherence relative to teacher-level reasoning.

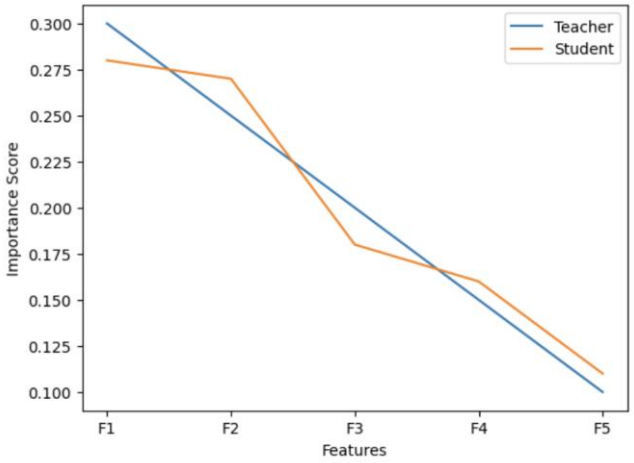

Figure 3. Comparative feature attribution distribution teacher and distilled student models

In Figure 3, report distributions of feature attributions by teacher and distilled student models. Effective interpretability seems to be present where there is close alignment in the ranking order of both models, and where there is consistent magnitude. The smaller variance in the attributions of the student model suggests soft target supervision leads to a smoothing effect. This is evidence to support our hypothesis that knowledge distillation improves explanation coherence rather than simply pulling the outputs close to the target.

Multi-Objective Trade-off Analysis

To study the balance of predictive performance, computational cost, and interpretability alignment, the composite objective function $J = \alpha_P P - \beta_C C + \gamma_I I$ I was analyzed across different weighting configurations.

Distilled models consistently reached the Pareto-efficient area across all configurations. In efficiency-focused scenarios (β_C high), student models outperformed teachers models in terms of composite score with significant margins. Even in accuracy-weighted scenarios, the differences in performances remained small.

With regard to the surface representation, the ideal distillation temperature for predictive fidelity and representation smoothing was observed to be between $\tau = 3$ and $\tau = 5$. Very low and very high distillation temperatures have been found to cause major accuracy degradation and over smoothing, respectively. In contrast, low temperatures have been observed to restrict improvements in interpretability alignment. The distillation hyperparameters largely dictate the trade-off between compression and transparency. The evidence confirms the hypothesis of soft-label supervision as a type of structural regularization, resulting in smoother decision boundaries and feature attribution that is less variable.

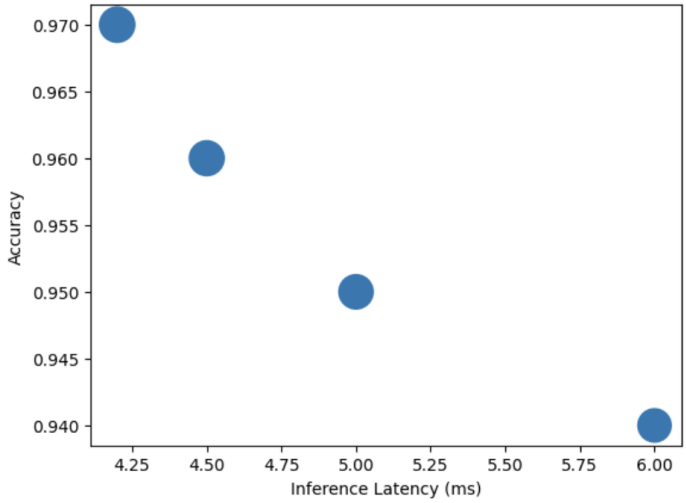

Figure 4. Multi-objective trade-off predictive performance and computational efficiency

Trade-offs within predictive accuracy, inference latency, interpretability, and other dimensions are described in Figure 4. The distilled configurations positioned as Pareto-efficient show that knowledge distillation achieves the best trade-off in all dimensions. The visualization also shows that choosing the right temperature and weighting parameters leads to a simultaneous improvement in efficiency and the alignment of interpretability.

3.2 Discussion

Distillation results show that student models capture almost all of the teachers' predictive abilities with little accuracy loss and even improved calibration stability in certain cases. Distillation improves computational resource savings by a large margin using significantly fewer parameters, utilizing less inference time and memory. Analysis shows that student models replicate the teacher models' interpretability. There is also consistent margin stability under perturbations, leading to less explanation instability. Knowledge distillation smooths out the decision functions and enhances the feature importance coherence, showing that it is not only model compression.

The results provide theoretical backing to distillation being a representational smoothing mechanism. Distillation in over-parameterized models leads to a loss of decision fluctuation over the model parameters that are used. Instead of using hard targets by the teachers, the students are able to use the teacher's smoothed output relative to the other data points which helps to simplify the model structurally while maintaining the intended purpose of the teacher. At the end of the day, knowledge distillation should be seen as a mechanism for designing explainable or trustworthy AI rather than an efficiency play.

The results of this research assist practitioners in the distillation of machine learning models that help to achieve a favorable trade-off between predictive performance and ease of operational deployment in machine learning implementations in resource poor or regulated environments. Aside from the demonstrated

trade-off surfaces, it is also possible to balance distillation configurations to achieve more specific objectives within a particular domain, whether that is improve performance or operational efficiency.

3.2.1 Implications

Results show that, for deployment, distilled models are a beneficial trade-off between interpretability and the other resources. For edge computing, real-time analytics and mobile AI systems, distilled architectures can be used, as there is a significant decrease in their required latency and memory.

In some regulated settings where explainability is a requirement, the stability in features attribution of the student models and the greater stability of the model increases trust and auditability, as do the models. Distillation can be used by organizations that want to responsibly and transparently deploy AI to justify decisions and reduce model size.

3.2.2 Research contribution

The findings provide a theoretical contribution by advancing the understanding of knowledge distillation to comprise more than just model compression. The increase in the stability of explanations indicates that distillation leads to representational smoothing that, in over-parameterized models, cause high frequency oscillations in decisions.

The aligned distributions of importance for features suggest that student models obtain the teacher's structure of relations while eliminating some parameters. This reinforces the understanding of distillation as a process of knowledge abstraction, rather than a straightforward case of parameter pruning.

Also, the stability improvements observed empirically are consistent with recent theoretical work that posits smoother predictive functions perform better on generalizations after some distributional perturbations. This is because a student model incorporates the softened function space through the teacher's soft probability distributions and thus, the model's predictive function has reduced variability while maintaining a rational bias.

3.2.3 Limitations

This research looked into knowledge distillation from the viewpoint of a solitary method to improve the interpretability and the computational efficacy of complex machine learning models. This research analyzed to what degree, distillation, beyond its classical engagement as a compression mechanism, can function as a type of structural regularization that balances trade-offs the ease of deployment and the stability of explanations, while preserving predictive accuracy. This research integrated several metrics from the literature to comprehensively explain the paradox of interpretability and ease of deployment along with the conjunctive value of efficacy, accuracy, and interpretability.

3.2.4 Suggestions

While the work has provided a number of benefits, some work remains to be done. Most of the evaluation centered on certain explanation methods, and supervised classification. On the other hand, work on regression, transformers and multimodal frameworks could be of high value. Additionally, the unexplored theoretical work on temperature scaling, the smoothing dynamics and other interpretability methods could be key.

4. CONCLUSION

The results substantiate all three of the most significant findings. First, predictive accuracy is diminished. Second, an overall significant reduction is observed in the computational burden and cost of deployment. Third, and most importantly, stability in interpretability and aligned feature importance is attained and increased under the framework.

The results uphold the core premise upon which the research is built, that knowledge distillation espouses efficiency increases and interpretability improvements in tandem. This enables more responsible and judicious AI systems to be constructed.

5. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the colleagues and peer reviewers who provided valuable feedback and technical insights during the development of this research. Special thanks are also extended to the open-source community for providing the benchmark datasets and software libraries that made this study possible.

6. AUTHOR CONTRIBUTION STATEMENT

JJ and KLC conceived and designed the study. JJ developed the integrated evaluation framework, performed the experiments, and conducted the statistical analysis. KLC and MS contributed to the multi-objective trade-off modeling and the theoretical interpretation of the results. JJ and MS wrote and revised the manuscript. All authors reviewed and approved the final version of the manuscript.

AUTHOR INFORMATION

Corresponding Authors

Jaesik Jeong, Department of Artificial Intelligence, Tamkang University, Taiwan
https://orcid.org/0000-0003-2601-9132
Email: 167030@o365.tku.edu.tw

Authors

Kit Ling Chan, Hong Kong Shue Yan University, Hong Kong, SAR, China
https://orcid.org/0009-0001-8085-7551
Email: 226002@hksyu.edu.hk

Mageswaran Sanmugam, Universiti Sains Malaysia, Malaysia
https://orcid.org/0000-0003-3313-4462
Email: mageswaran@usm.my

REFFERENCE

Balakrishnan, V., Shi, Z., Law, C. L., Lim, R., Teh, L. L., Fan, Y., & Periasamy, J. (2022). A Comprehensive Analysis of Transformer-Deep Neural Network Models in Twitter Disaster Detection. *Mathematics*, *10*(24), 1–14. https://doi.org/10.3390/math10244664

Bjerring, J. C., Jakob, B., & Lauritz, M. (2025). Deep learning models and the limits of explainable artificial intelligence. *Asian Journal of Philosophy*, *4*(1), 1–26. https://doi.org/10.1007/s44204-024-00238-8

Bruggeman, F. J., Teusink, B., & Steuer, R. (2023). Trade-offs between the instantaneous growth rate and long-term fitness : Consequences for microbial physiology and predictive computational models. *BioEssays*, *45*(10), 1–20. https://doi.org/10.1002/bies.202300015

Demircioğlu, A. (2025). Reproducibility and interpretability in radiomics: a critical assessment. *Artificial Intelligence and Informatics*, *31*(4), 321–328. https://doi.org/10.4274/dir.2024.242719

Ezzahra, Z. F., Sana, A., Sara, Q., & Said, R. (2025). Multi-objective reinforcement learning for recommender systems: a comprehensive survey of methods, challenges, and future directions. *International Journal of Multimedia Information Retrieval*, *14*(33). https://doi.org/10.1007/s10735-025-00383-7

Gunasekara, S., & Saarela, M. (2025). applied sciences Explainable AI in Education : Techniques and Qualitative Assessment. *Applied Sciences*, *15*(3), 1239. https://doi.org/10.3390/app15031239

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). International Journal of Research in Marketing More than a Feeling : Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hohman, F., Kery, M. B., Ren, D., & Moritz, D. (2026). Model Compression in Practice : Lessons Learned from Practitioners Creating On-device Machine Learning Experiences Model Compression in Practice : Lessons Learned from Practitioners Creating On-device Machine Learning Experiences. *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, *645*. https://doi.org/10.1145/3613904.3642109

Huang, J., Mishra, A., Kwon, B. C., & Bryan, C. (2022). ConceptExplainer : Interactive Explanation for Deep Neural Networks from a Concept Perspective. *IEEE Transactions on Visualization and Computer Graphics*, *29*(1), 831–841. https://doi.org/10.1109/TVCG.2022.3209384

Kucklick, J., & Muller, O. (2026). Tackling the Accuracy-Interpretability Trade-off : Interpretable Deep Learning Models for Satellite Image- based Real Estate Appraisal Tackling the Accuracy-Interpretability Trade-off : Interpretable Deep Learning Models for Satellite Image-based Real Est. *ACM Transactions on Management Information Systems*, *14*(1), 1–24. https://doi.org/10.1145/3567430

Lan, W., Cheung, Y., Xu, Q., Liu, B., Hu, Z., & Li, M. (2025). Improve Knowledge Distillation via Label Revision and Data Selection. *IEEE Transactions on Cognitive and Developmental Systems*, *17*(6), 1377–1388. https://doi.org/10.1109/TCDS.2025.3559881

Li, C., Cheng, G., & Han, J. (2024). Boosting Knowledge Distillation via Intra-Class Logit Distribution Smoothing. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(6), 4190–4201. https://doi.org/10.1109/TCSVT.2023.3327113

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Bian, J., & Dou, D. (2022). Interpretable Deep Learning: Interpretations, Interpretability, Trustworthiness, and Beyond. *Knowledge and Information Systems*, *64*(12), 3197–3234. https://doi.org/10.1007/s10115-022-01756-8

Lin, S., Lin, W., Wu, W., Chen, H., & Chen, C. L. P. (2026). SparseTSF: Lightweight and Robust Time Series Forecasting via Sparse Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *48*(1), 170–183. https://doi.org/10.1109/TPAMI.2025.3602445

Liu, C., Yin, H., & Wang, X. (2025). Theoretical Perspectives on Knowledge Distillation : A Review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *17*(4), 1–17. https://doi.org/10.1002/wics.70049

Mai, N. T., Cao, W., & Liu, W. (2025). Interpretable Knowledge Tracing via Transformer-Bayesian Hybrid Networks : Learning Temporal Dependencies and Causal Structures in Educational Data. *Applied Sciences*, *15*(17), 1–26. https://doi.org/10.3390/app15179605

Malihi, L., & Heidemann, G. (2024). Matching the Ideal Pruning Method with Knowledge Distillation for Optimal Compression. *Applied Sciences Innovation*, *7*(4), 56. https://doi.org/10.3390/asi7040056

Mandler, H., & Weigand, B. (2026). A review and benchmark of feature importance methods for neural networks. *ACM Computing Surveys*, *56*(12), 1–30. https://doi.org/10.1145/3679012

Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC - IV dataset. *Scientific Reports*, *12*(7166), 1–28. https://doi.org/10.1038/s41598-022-11012-2

Panigrahi, B., Razavi, S., Doig, L. E., Cordell, B., Gupta, H. V., & Liber, K. (2025). On Robustness of the Explanatory Power of Machine Learning Models : Insights From a New Explainable AI Approach Using Sensitivity Analysis. *Water Resources Research*, *61*(3), 1–23. https://doi.org/10.1029/2024WR037398

Rahman, A., Debnath, T., Kundu, D., & Khan, S. I. (2024). Machine learning and deep learning-based approach in smart healthcare : Recent advances , applications , challenges and opportunities. *AIMS Public Health*, *11*(1), 58–109. https://doi.org/10.3934/publichealth.2024004

Salih, A. M., Raisi-estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A Perspective on Explainable Arti fi cial Intelligence Methods : SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 1–8. https://doi.org/10.1002/aisy.202400304

Sonrel, A., Luetge, A., Soneson, C., Mallona, I., Germain, P. L., Knyazev, S., Gilis, J., Gerber, R., Seurinck, R., Paul, D., Sonder, E., Crowell, H. L., Fanaswala, I., Ajami, A. Al, Heidari, E., Schmeing, S., Milosavljevic, S., Saeys, Y., Mangul, S., & Robinson, M. D. (2023). Meta - analysis of (single - cell method) benchmarks reveals the need for extensibility and interoperability. *Genome Biology*, *24*(119), 1–11. https://doi.org/10.1186/s13059-023-02962-5

Tan, C., & Liu, J. (2022). Improving Knowledge Distillation With a Customized Teacher. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(2), 1–10. https://doi.org/10.1109/TNNLS.2022.3189680

Wu, C.-J., Acun, B., Raghavendra, R., & Hazelwood, K. (2024). Beyond Efficiency: Scaling AI Sustainably. *IEEE Micro*, *44*(5), 37–46. https://doi.org/10.1109/MM.2024.3409275

Zhang, L., & Ma, K. (2023). Structured Knowledge Distillation for Accurate and Efficient Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(12), 15706–15724. https://doi.org/10.1109/TPAMI.2023.3300470

Zhang, Y., Hu, S., Zhang, L. Y., Shi, J., Li, M., & Liu, X. (2024). Why Does Little Robustness Help? A Further Step Towards Understanding Adversarial Transferability. *Proceedings - IEEE Symposium on Security and Privacy*. https://doi.org/10.1109/SP54263.2024.00010