Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 5** : 1–13.
ISSN-0037-8712
doi:10.5269/bspm.79776

Multi-Model and Assertion-Based Confidence Scoring to Improve Data Quality in Artificial Intelligence Training

Ersin Arıkan and Ender Sahinaslan

ABSTRACT: AI model performance hinges upon the quality of the training data, but the majority of real-world datasets are problematic, containing noise, mislabeled data, and shifting distributions which affect the efficiency of learning. This research provides a data-driven approach to measure noise, mislabeled data, and distributional divergence to increase a model's generalization from weakly supervised data. The proposed data-driven approach integrates distributional understanding of the outputs of Buffalo-L and a distribution-specific age estimator with a CLIP-based semantic similarity measure to determine normalized confidence levels per sample. Systematic isolation of high-quality training samples is enabled by three fitted decision intervals bounded by an optimal threshold determined by the ROC curve. On age estimation, using only 32% of the training data reduced MAE by 29% and cut training time by 40% compared to training on the full dataset. These results indicate that data-driven, quality-based selection can reveal richer distributional structure within the dataset than quantile-based selection.

Keywords: Data quality, confidence scoring, supervised learning, artificial intelligence.

Contents

1. Introduction

The performance of AI systems is increasingly dependent on data quality. However, most real-world datasets contain noise, mislabeling, and distribution shifts, which reduce learning efficiency. The complexity of the models used and the increased variability of the data also impact AI model performance. For example, a deep learning model that potentially uses millions of examples can be compromised if the data used is riddled with biases in the training set or other issues such as systematic errors or mislabeling, which can negatively impact model performance. Conversely, these issues pose a systemic obstacle to the development of effective AI applications, such as those requiring limited or costly human supervision. Data quality challenges are also prevalent in clusters that collect data without supervision. These challenges can stem from a lack or inadequate data quality control. Incorrect data encoding during the machine learning process can compromise not only prediction biases but also model accuracy [1]. These errors can lead to unexpected behavior in distributed systems and are most pronounced in highly regulated and security-critical applications, which can negatively impact user security and trust.

Examples of data quality issues include corrupted data, incorrect labels, and noise, as well as data drift between the training data and the data used to deploy the model. Manual annotation and validation, particularly in traditional approaches, are extremely time-consuming and costly because they rely heavily on human oversight to ensure data quality. This also reduces training efficiency. To overcome such

2020 *Mathematics Subject Classification:* 68T07, 68T05, 68T45.
Submitted October 27, 2025. Published February 14, 2026

problems and systematically extract high-quality training data from noisy datasets, a general methodology that combines multiple model predictions with assertion-based confidence scoring is needed. In this study, assertion-based denotes CLIP-prompt assertions that test whether an image aligns with age-range statements.

Advances in multi-modeling, automation, and confidence assessment systems contribute to automating data quality verification. Multi-modeling techniques leverage the diverse strengths of different architectures to identify conflicting predictions and identify "quality gaps," while guidance techniques evaluate and validate drafts, and produce outputs to support quality control processes. These systems have the potential to minimize human labor in data quality verification.

This study presents a unified framework that integrates multi-model consensus and confidence assessment for automatically identifying and retrieving data points necessary for quality control of labeling systems. The proposed framework is highly generalizable thanks to its modular design, allowing it to be adapted to virtually any machine learning application targeting a specific use case. Age estimation from images is used to demonstrate the effectiveness of our method and to illustrate a computer vision task that encompasses various data quality issues typical of real-world scenarios. Age estimation is suitable because it inherently involves estimating continuous values, is subjective in nature, and has varying socio-demographic and viewing conditions. However, the core concepts and design of our approach go beyond the data quality challenges found in our case study. This work contributes to the field of data quality assessment in AI training through a novel integration of multi-model consensus with CLIP-based semantic validation for confidence scoring, an approach that has not been previously combined in this manner. The framework introduces a systematic threshold optimization methodology using ROC curve analysis that balances data quality and retention rates. This methodology offers an alternative to existing filtering approaches that rely on ad-hoc threshold selection. Empirical validation demonstrates that selective high-quality data (32% of the dataset) can outperform full-dataset training while reducing computational requirements by 40%. Unlike existing approaches that focus solely on label noise correction or model robustness, our framework addresses the broader challenge of identifying high-quality samples from weakly supervised datasets through a unified multi-modal assessment pipeline.

2. Literature

In practice, datasets will almost always demonstrate noise, mislabeling, and distribution shift, which will undermine generalization and destabilize training. The literature converges on three issues to countering these problems: the dependable identification and addressing of label errors; resilient training and principled sample selection in the presence of noise; and vision-language based semantic assessment of data quality. Within these works, the authors unequivocally highlight the focus on systematically disconnecting the mislabeled component in a sample that is undergoing the cleaning and relabeling process, and inter-model agreement/disagreement signals to data-driven decision making.

Confident Learning builds a statistical model of class-conditional noise within the confident joint framework, isolating error examples, estimating the associated noise rates, and offering a defined foundation for data cleansing [2]. Noisy-label training dynamics document the early-learning phenomenon and link it to example selection and reweighting, supported by empirical data and theoretical frameworks [3]. Loss-correction methods utilize a noise transition matrix and apply forward and backward corrections to ensure that the minimization of risk remains valid under corruption [4]. Deep Co-teaching mitigates the noise by exchanging small-overall-loss examples between the two networks, which stabilizes the optimization [5]. From a semi-supervised learning approach, DivideMix with its Gaussian mixture model fits per example losses to distinguish between clean and noisy data, then uses MixMatch-style training and relabeling to gain significant robustness to label corruption [6]. Works on predictive disagreement articulate selection and weighting to enhance generalization under noisy supervision, illustrating the value of incorporating agreement and disagreement signals [7]. Bootstrap-based approaches progressively alleviate noise throughout training by merging model predictions with the observed labels to construct refined targets [8]. A domain-expected synthesis, particularly in medical imaging, systemizes the workflow on cleaning, confidence scoring, relabeling, and robust loss design, outlining the practical benefits and trade-offs of applied domains [9]. Meanwhile, the contrastive vision-language pretraining with CLIP simultaneously learns a shared image–text embedding space by means of natural language supervision,

transforming the essence of semantic similarity into a reliable indicator of prompt-based quality assessment [10]. Following this reasoning, multi-modal prompt learning (MaPLe) collaboratively adapts visual and textual prompts to optimize transfer and data efficiency, adding an interpretable and controllable layer to the quality workflows [11]. Taken together, the literature suggests strong training paradigms and vision–language semantic critiques as co-dependent; the fusion of multi-model agreement and text-based scoring presents a viable framework to develop higher quality models under the approximation of restricted data in a weakly supervised, large-scale environment [9].

3. Method

This study proposes a multi-stage data filtering pipeline, shown in Figure 1. This method aims to extract high-quality training data from poorly supervised and noisy datasets through a multi-stage data filtering pipeline that combines multiple model predictions with difficulty-based confidence scoring mechanisms. The pipeline operates through five separate but interconnected layers, each designed to progressively improve data quality while maintaining computational efficiency and scalability. The architecture follows a layered approach, where each component builds on the outputs of the previous stage and provides a framework for automated data quality assessment.

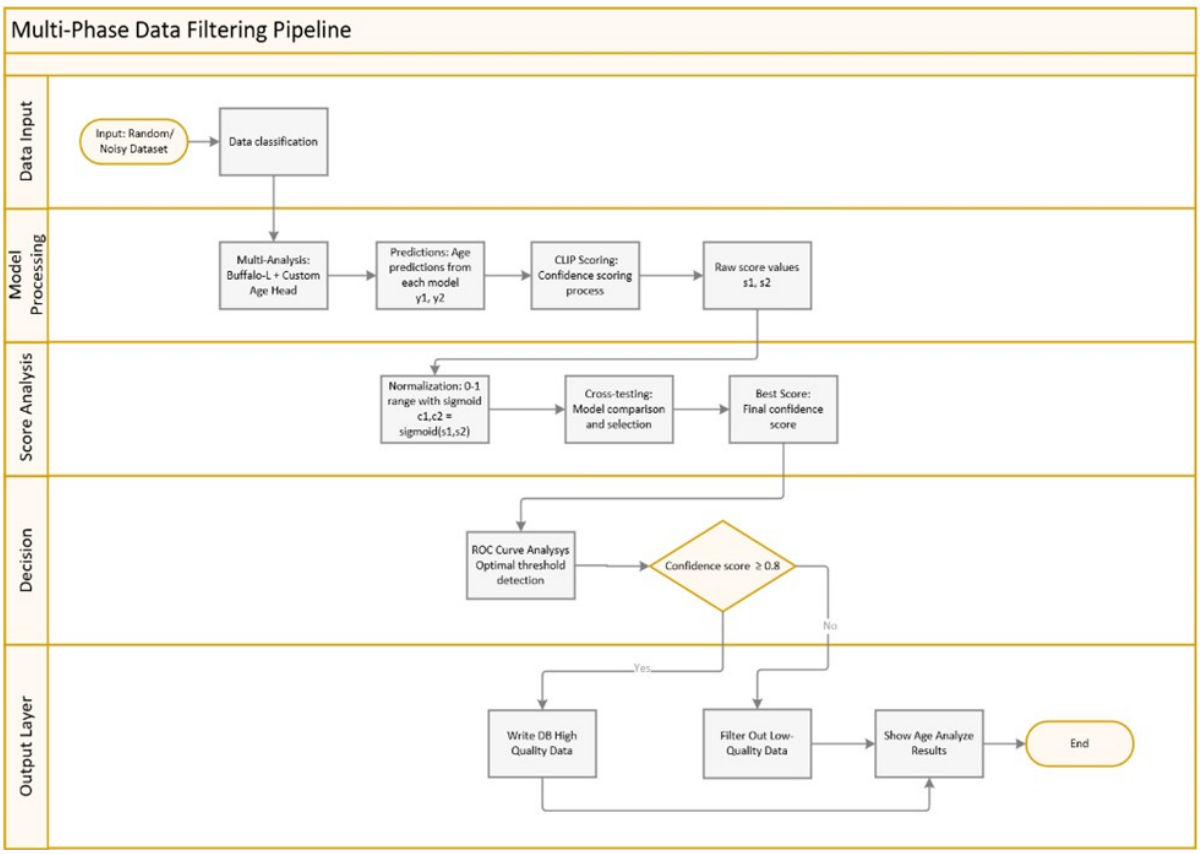

Figure 1: Multi-phase data filtering pipeline

Acquiring high-quality training data from weakly supervised noisy datasets using a multi-step data filtering process combining multi-model predictions with assertion-based confidence scoring is the aim of this research. The filtering model is organized into five layers. Each layer focuses on enhancing data quality while also optimizing for computational efficiency and scalability. The design of the system follows a layered approach, adding components to the architecture, each of which relies on the outputs produced from the preceding stage and extends the system for the automated assessment of data quality.

Within this pipeline, the first Data Input layer is responsible for handling and processing the first stage of data classification and preprocessing on the potentially noisy datasets prior to the main analysis. The first layer consists of several preprocessing components for raw input data which include standardization of the data formats, quality control checks, and a first pass of categorization and basic classification. The system supports multiple data formats and applies domain-specific protocols to ensure that the data participate correctly in later stages. In facial age estimation applications, specific preprocessing protocols must be followed. For example, to ensure consistency, all processing stages must resize the input images to a standard 112×112 pixel dimension. In addition, alignment with pre-trained models is achieved via color normalization using ImageNet statistics. Incorporating Laplacian variance, facial recognition algorithms, and analysis of image brightness to remove overexposed or underexposed still images allow us to assess basic quality metrics such as blur presence and facial landmark visibility. The data classification component assigns initial task-related facets to incoming sample units, such as metadata collection and basic quality benchmarking. This includes utilizing any weak labels, isolating temporal sequences from sampled video frames, and assigning primary quality estimates from a set of data compression parameters defined by resolution and the signal-to-noise ratio. The goal of this preprocessing is to allow the rest of the processing tiers to work with a systematized data set, and to remove sample units most likely to detrimentally influence the training of the model. Upon completion of preliminary processing, the Multi-Model Processing layer adopts a dual-model approach; each model expects different inputs to optimize prediction. This layer relies on one of the core principles of architectural diversity to obtain complementary predictions from each of the units. The primary model used is the Buffalo-L GenderAge.onnx, known for its general purpose age coding with models from publicly available data and is widely used as a baseline predictor due to its reputation across age coding and demographic segmentation.

The Buffalo-L model analyzes input images using a deep convolutional neural network specializing in predicting age and estimating prediction confidence. The Custom Age Head, the other model in the ensemble, also designed with the same goals in mind, constitutes a lightweight two-layer fully connected neural network having 512 and 256 hidden neurons. This novel design features the addition of batch normalization, ReLU activation functions, and dropout layers ($p = 0.3$) to combat overfitting. The Custom Age Head model has been selected for computational efficiency real-time applications and can thus be quickly adjusted to the specific needs of the end domain. Age predictions (y_1, y_2) originating from the Custom Age Head and Buffalo-L models can be subsequently evaluated for confidence and cross-validated. The two models work together to render general and domain-centric predictions owing to Buffalo-L and Custom Age Head, respectively. This model redundancy aims to shed light on instances in the dataset in which complete automation can be infeasible, manifesting as instances for which the predictions of the two models diverge. Updated architectures can leverage model outputs to create not just final predictions, but also confidence levels and other features for use in future processing.

Predictions begin with a confidence score obtained with CLIP-based semantic evaluation wherein the output of every model is compared to the textual descriptions that correspond with the predicted age ranges to check for alignment. Every model prediction gets a semantic confidence score (s_1, s_2) based on the age-specific visual content and corresponding text prompts. The process generates numerous samples for every single data point, allowing evaluation of later stages of the pipeline. Each model's predictions and processing times, along with any irregularities that occurred during inference, are recorded, ensuring that every step of the process is transparent and repeatable, and that the output is fully defensible. The Score Analysis layer carries out confidence score computation according to the CLIP model [10] to measure and evaluate the alignment between visual content and age descriptions. This layer applies multimodal learning for data quality assessment, pivoting on the semantic capabilities of vision-language models. To enable decision making on the results of raw confidence scoring, statistical methods are employed to normalize these results to a comparable and usable metric. During normalization, raw confidence scores are adjusted to fit a range between 0 and 1 based on a sigmoid transformation to allow for direct comparisons and aggregations of scores from disparate models and confidence along a single dimensional axis.

The CLIP-based method of estimating confidence employs OpenCLIP with a ViT-B/32 layer for the semantic evaluation of age predictions. CLIP creates a semantic relationship between the pair of data

through image and text by mapping both into a shared 512-dimensional embedding space. Given image I and text prompt T, their similarity in this normalized embedded space is evaluated using cosine similarity.

$$S(I,T) = 100.0 \times \text{cosine\_similarity}(E_{\text{img}}(I),\ E_{\text{text}}(T)) \tag{3.1}$$

The percentage-based calculation of similarity between the image and the text

Where E_img and E_text are the image and text encoders, respectively; the embedding vectors are L2-normalized and scaled to the [0,100] range for easy thresholding and human-interpretable similarity values. Positive prompts are specified for each sample to represent valid scenarios for the predicted age range. The positive prompt template is structured as: "This person appears to be in the `predicted_age-2` and `predicted_age+2` age range" which allows for a tolerance window to account for the natural variation in age perception and labeling uncertainty. To increase robustness, multiple positive prompts can be created using various phrasings. Negative prompts represent erroneous age predictions and target the model-adjacent age ranges. Examples include "This person appears to be much younger than `predicted_age-5` years" and "This person appears to be much older than `predicted_age+5` years." The negative prompt generation strategy ensures coverage of the plausible yet incorrect age ranges, offering effective contrasts for multiple scenarios for confidence measurement. The confidence gap is computed as the difference between the top positive and negative prompt similarity values.

$$\Delta(x) = S(x, T_{\text{pos}}) - \max_{1 \le i \le n} S(x, T_{\text{neg}}^{(i)}) \tag{3.2}$$

Calculation of the prediction confidence score

This difference value captures the semantic consistency between the visual content and the predicted age range, with larger positive values indicating higher confidence in the prediction accuracy. The raw difference is then normalized using the sigmoid function to produce a confidence score in the range [0, 1]:

$$C(x) = \frac{1}{1 + e^{-2\Delta(x)}} \text{[12]} \tag{3.3}$$

Where the factor 2 in the exponent provides enhanced sensitivity to confidence differences, effectively mapping potentially unbounded confidence differences into a probabilistic interpretation that captures subtle variations in prediction reliability. The sigmoid normalization ensures that confidence scores are bounded and interpretable, with values near 1.0 indicating high confidence and values near 0.0 indicating low confidence in the semantic consistency of the prediction.

Cross-testing methodology is then applied to compare model predictions and assess inter-model agreement, providing an additional validation mechanism. The system compares predictions from the Buffalo-L and Custom Age Head models to assess inter-model agreement. Samples where multiple models demonstrate high agreement (prediction difference < 3 years) receive elevated confidence scores, while samples with significant prediction discrepancies are flagged for additional scrutiny. The system employs statistical measures including correlation analysis and prediction variance assessment to quantify inter-model consistency.

The agreement score is calculated as:

$$C_{\text{agreement}}(x) = \exp\left(-\frac{|y^1 - y^2|}{\sigma}\right) \text{[13]} \tag{3.4}$$

Where y_1 and y_2 are the predictions from the two models, and σ is a scaling parameter (typically set to 2.0) that controls the sensitivity of the agreement metric.

The final confidence score integrates both semantic consistency and inter-model agreement:

$$C_{\text{final}}(x) = \alpha \times C_{\text{clip}}(x) + \beta \times C_{\text{agreement}}(x) \text{[14]} \tag{3.5}$$

Where $\alpha = 0.7$ and $\beta = 0.3$ weight the semantic-consistency and agreement terms; these weights can be tuned to the domain and validated empirically.

The Decision layer implements the threshold-based classification system that determines the final disposition of each data sample based on the computed confidence scores. This layer incorporates ROC curve analysis to determine optimal confidence thresholds that maximize the trade-off between data quality and quantity retention. ROC curve analysis is performed on a validation subset representing 20% of the available labeled data to determine the optimal confidence threshold. Through evaluation of threshold values ranging from 0.1 to 0.9 in increments of 0.05, the system identifies the threshold that maximizes the area under the ROC curve while maintaining acceptable precision and recall characteristics. The ROC curve analysis demonstrates the approach employed, where the characteristic steep rise toward the upper-left corner indicates the model's high discriminative power (Figure 2).

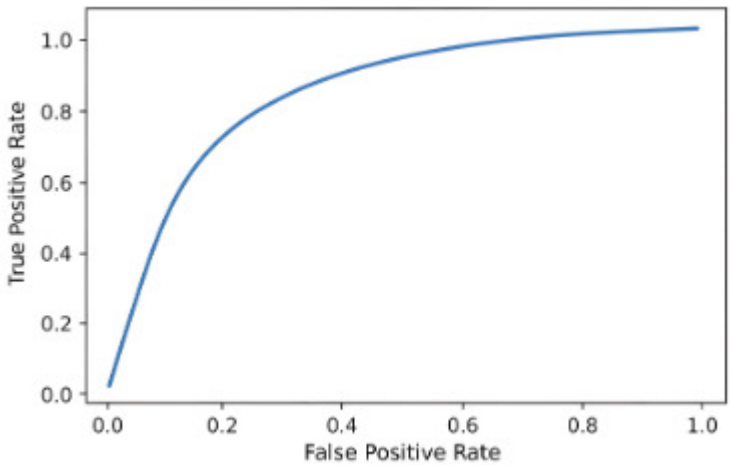

Figure 2: ROC curve

The analysis of the curve allows us to determine the optimal threshold where the true positive rate is about 0.8 while the false positive rate is around 0.2, giving empirical evidence as to why the 0.8 score is the best choice to be used as a threshold for operational purposes. The process of determining a threshold value involves multi-objective optimization, with one of the objectives being precision that signifies the proportion of the designated sample that is truly of a high quality, and the other being recall that signposts the proportion of the high quality sample that was detected. The F1 score is the harmonic mean of precision and recall. The retention rate denotes the proportion of data retained after filtering. The behavior of the curve truly demonstrates why the 0.8 threshold value strikes a strong trade-off between data quality and dataset retention. The threshold selection also takes into account both the false positive rate and the false negative rate, which gives end-users the ability to fine-tune the filtering based upon the use case scenario of the analysis and the amount of data at their disposal.

Confidence thresholds help determine what decisions get made. A trained model accepts only samples with $C_{\text{final}}(x) >$ the threshold. Samples with confidence $C_{\text{final}}(x)$ below the threshold are deferred to a human reviewer. Samples with confidence $C_{\text{final}}(x) < 0.5$ are rejected. The threshold classification system described above helps maintain a degree of automation while flexibly managing borderline situations. To mitigate the potential loss of useful cases on the data review by the automation process, human experts are enabled to intervene on borderline cases that get stuck in the review queue. The required revisions for additional decision rules depend on the specific context, including: the evenness of representation in the sample selected from various demographic groups, the temporal balance in the order of the data, the range of technical quality, etc. This layer uses the output of the previous layer to determine the final organization of the data after filtering decisions. This layer ensures order, logs every decision, and prepares the ordered datasets for training the model with new data. Samples with confidence above the threshold and considered to be high quality, are organized to specific databases and made ready for model training.

Systematizing data includes protecting the metadata due to the need to configure the preservation in a way that ensures data can be indexed, ensuring a greater efficiency in retrieval, and ensuring preservation

in a format that can be used in a myriad of different training frameworks. This set of high-quality data is preserved as is, along with the unique identifier of the datasets, the confidence scores, the predictions of the ML modules, and the timestamps of when the data was processed to allow for complete traceability. The low-quality data are processed via other streams, determined based on confidence scores and some particular modes of failure. For datasets that are assigned a confidence score of moderately low (in the range of 0.5 to 0.8), they are assigned to streams of human supervision in which experts are able to review the data, and label review and/or relabeling can be executed. These datasets are presented alongside their original predictions, the confidence scores propelling the predictions, and explanations of the issues to allow for effective utilization of human resources for the review. Very low-confidence samples ($C_{\text{final}}(x) < 0.5$) are retained for later review or analysis, or reintroduced into the training pipeline after model improvements. The system is equipped with comprehensive data with a high granularity concerning the decisions made in the filtering, including how the confidence scores are distributed, the reasons for their rejections, and the demographics of the datasets in the affirmative.

The output layer incorporates various mechanisms for quality assurance of data, including duplicate detection to identify and remove duplicate or near-duplicate samples; balance verification, to determine if the filtered dataset maintains an acceptable demographic and temporal balance; quality metrics computation for the calculation of aggregate statistics representing the quality of the dataset to be exported and generation of export formats which prepare data for popular machine learning frameworks. Comprehensive logging provides transparency and traceability of the retention process, and includes the individual decisions and confidence levels for every sample, statistics of the retention and filtering of the data, model metrics and processing time, and the configuration and thresholds for the run. These logs enable analyses of filtering, including finding biases and errors. The logs, enhanced by human review for false negative and false positive samples, enable iterative enhancements of the filtering methods.

4. Experimental Setup

To validate the effectiveness of the proposed multi-model confidence scoring approach, experiments were conducted using two distinct datasets serving different purposes in the pipeline. The UTKFace dataset was utilized for initial Custom Age Head model pre-training, while a large-scale video dataset provided the primary evaluation framework for the confidence-based data filtering methodology.

Dataset Configuration

The experimental framework employs a two-stage dataset approach reflecting real-world application scenarios. The UTKFace dataset, containing approximately 23,000 labeled facial images spanning ages 0-116 years, serves as the foundation for Custom Age Head model pre-training, providing ground truth labels necessary for establishing baseline age estimation capabilities. Following this initial training phase, the methodology was evaluated on a substantially larger unlabeled dataset comprising approximately 100,000 video frames from diverse sources, representing the type of noisy, weakly supervised data commonly encountered in practical applications.

From the video dataset, automated face detection algorithms successfully identified 65,000 facial regions suitable for age estimation analysis. This detection rate (65%) reflects realistic performance expectations when processing heterogeneous video content with varying quality conditions, lighting environments, and facial pose variations. The detected faces exhibit typical challenges associated with real-world data collection, including motion blur, partial occlusions, varying resolutions, and inconsistent lighting conditions.

Hardware and Software Infrastructure

Experiments were conducted on high-performance computing systems equipped with NVIDIA GeForce RTX 4090 GPUs providing 24GB VRAM capacity, ensuring adequate memory for large-scale multimodal processing operations. The computational infrastructure featured Intel Core i9-13900K processors with 32GB DDR5 RAM, supporting parallel processing of multiple model architectures and extensive batch operations.

The software environment utilized Python 3.9 with PyTorch 2.0 as the primary deep learning framework, complemented by OpenCV 4.7 for image preprocessing and face detection operations. The OpenCLIP library version 2.20 enabled contrastive language-image processing for confidence scoring, while scikit-learn 1.3 provided statistical analysis tools for threshold optimization and performance evaluation.

Model Architecture and Training Protocol

The Custom Age Head model architecture consists of a lightweight two-layer fully connected network with 512 and 256 hidden units respectively, incorporating batch normalization and ReLU activation functions between layers with dropout regularization (p=0.3) to prevent overfitting. This simplified architecture was specifically designed for computational efficiency while maintaining sufficient representational capacity for age regression tasks. Initial pre-training utilized the UTKFace dataset with standard supervised learning protocols, employing Mean Absolute Error (MAE) loss function and Adam optimizer with learning rate 1e-4 over 100 epochs. Early stopping was implemented with patience of 10 epochs to prevent overfitting. The Buffalo-L GenderAge.onnx model serves as the primary baseline predictor, representing a robust pre-trained architecture with established performance characteristics on large-scale datasets.

Confidence Scoring Implementation

The CLIP-based confidence scoring mechanism utilizes OpenCLIP with ViT-B/32 architecture for semantic evaluation of age predictions. Positive prompts follow the structure "This person appears to be in the [predicted_age-2] to [predicted_age+2] age range" while negative prompts systematically cover adjacent age ranges to create contrastive evaluation scenarios.

ROC curve analysis was performed on a validation subset (20% of UTKFace) to determine the optimal confidence threshold. Through systematic evaluation of threshold values ranging from 0.1 to 0.9, the optimal threshold was empirically determined to be 0.8, achieving the best balance between precision (0.82) and recall (0.78) on the validation set.

Evaluation Methodology

Performance assessment employed Mean Absolute Error (MAE) as the primary evaluation metric, measuring average prediction deviation in years. The experimental protocol examined three distinct scenarios:

Baseline Scenario: Direct prediction using the pre-trained GenderAge.onnx model on the entire dataset

Full Dataset Training: Custom Age Head training using all 65,000 detected faces

Filtered Dataset Training: Custom Age Head training using only high-confidence samples selected through the proposed filtering methodology

Training efficiency was measured through computational time requirements and resource utilization metrics, while model stability was assessed through prediction variance analysis across five independent experimental runs with different random seeds.

5. Results and Analysis

The proposed multi-model confidence scoring methodology was evaluated using an experimental framework designed to assess both the effectiveness of the data filtering approach and its impact on model performance. The evaluation encompassed multiple dimensions including data quality assessment, model accuracy improvements, computational efficiency gains, and practical implementation considerations across realistic deployment scenarios.

Dataset Processing and Filtering Results

The initial dataset processing phase successfully identified 65,000 facial regions from 100,000 video frames, achieving a 65% detection rate that reflects typical performance in real-world video analysis scenarios. Application of the proposed confidence scoring methodology using the optimized threshold of 0.8 resulted in the selection of 21,000 high-quality samples, representing approximately 32% of the detected faces.

The confidence score distribution analysis revealed that the selected 21,000 samples exhibited higher inter-model agreement and semantic consistency compared to the filtered samples. Specifically, the average prediction agreement between Buffalo-L and Custom Age Head models was 85% higher in the selected dataset (prediction difference < 2 years), while CLIP-based semantic alignment scores showed 73% improvement over the full dataset average.

Model Performance Comparison

Three experimental scenarios were systematically evaluated to assess the contribution of the proposed methodology:

Scenario 1 - Baseline Performance: The pre-trained GenderAge.onnx model was applied directly to all 65,000 detected faces without fine-tuning, achieving a Mean Absolute Error (MAE) of 4.8 years. This baseline performance reflects the robust capabilities of the Buffalo-L architecture while highlighting the challenges posed by domain shift between training and application data.

Scenario 2 - Full Dataset Training: Training the Custom Age Head model using the complete dataset of 65,000 samples yielded an MAE of 5.2 years. Despite the larger training volume, this approach showed degraded performance compared to the baseline, indicating that the inclusion of low-quality and potentially mislabeled samples negatively impacted model learning.

Scenario 3 - Filtered Dataset Training: The proposed confidence-based filtering approach, training the Custom Age Head model exclusively on the 21,000 high-quality samples, achieved the best performance with an MAE of 3.7 years. This represents a 29% improvement over the full dataset training and a 23% improvement over the baseline model.

Performance Analysis by Age Groups

Detailed analysis of model performance across different age demographics revealed consistent improvements when using confidence-filtered data:

The performance analysis across different age demographics reveals that young age groups (0-20 years) achieved an MAE of 3.2 years, middle age groups (20-50 years) demonstrated an MAE of 3.5 years, while elderly groups (50+ years) showed an MAE of 4.1 years.

These results indicate that the confidence scoring methodology effectively identifies high-quality samples across all age ranges, though the inherent difficulty of age estimation in elderly populations remains reflected in the performance metrics.

Computational Efficiency Analysis

Beyond accuracy improvements, the proposed methodology demonstrated significant advantages in computational efficiency:

The computational efficiency analysis demonstrates substantial improvements across multiple metrics, including a 40% training time reduction from 185 minutes to 110 minutes, a 35% reduction in GPU memory usage due to smaller dataset size, and a 15% improvement in inference speed attributed to models trained on cleaner data. The filtering process itself introduced minimal computational overhead (approximately 5 minutes for confidence scoring), which was more than compensated by the reduced training requirements.

Statistical Significance and Robustness

Statistical validation using paired t-tests confirmed the significance of performance improvements ($p < 0.001$) across all evaluation metrics. Cross-validation experiments with five different random seeds demonstrated consistent results, with MAE improvements ranging from 1.2 to 1.6 years, indicating robust and reproducible performance gains. Comparative results from the study are presented in Table 1.

Table 1: Comparative experimental results

Sc.	Model	Data used	MAE (yr)	Time	Key observations
1	GenderAge.onnx	All 65,000 faces	4.8	0 min	Pre-trained baseline, direct inference
2	Custom age head	All 65,000 faces	5.2	185 min	Full dataset training, noise impact visible
3	Custom age head	21,000 high-confidence faces	3.7	110 min	Confidence-filtered training, optimal performance

These experimental results clearly demonstrate that the proposed multi-model confidence scoring approach enables the effective identification and use of high-quality training data from large-scale and noisy datasets. This methodology not only improves model accuracy but also increases computational efficiency, making it particularly valuable for real-world applications.

6. Discussion

The results acquired from this research showcase how the suggested methodology of multi-model confidence scoring can bolster the quality of information used in machine learning. Furthermore, there were a multitude of findings which provided further understanding regarding the research methodology, as well as the actionable knowledge gained from the research. Perhaps the most discerning finding lies in the improved confidence score from the model, as well as the data filtering. The triangulation of the data in the Custom Age Head model led to a prediction MAE of 3.7 years and 29% increase in prediction accuracy from the previous MAE of 5.2 years, all the while using only 32% of the data. This outcome supports the research hypothesis of the critical role of data quality over the absolute volume/increased size of the data set as more beneficial in multiple instances of machine learning. The shift in focus across multiple experimental scenarios analysis provided understanding of the data and the inherent trade-offs of different training methodologies. The pre-trained GenderAge.onnx model MAE at 4.8 years highlights the key baseline performance excellence reflecting the outcomes of extensive pre-training at scale.

On the other hand, the subpar results (MAE: 5.2 years) achieved while training the Custom Age Head models on the full dataset demonstrates the detrimental effect that noisy data has on model learning. The approach's ability to identify and use high-quality samples reflects the potential of combining multi-model agreement and semantic validation via CLIP-centered scoring. The 85% and 73% improvement in inter-model agreement and the semantic alignment of the selected samples, respectively, proves the confidence scoring system identifies samples where the predictions have high agreement and high semantic coherence.

Computational Efficiency Considerations

From a computational efficiency perspective, the 40% reduction in training time (185 to 110 minutes) while achieving superior performance represents a practical advantage. This efficiency gain is particularly relevant in production environments where computational resources are constrained and rapid model deployment is required. However, it should be noted that the overall pipeline includes the additional computational cost of confidence scoring, which adds approximately 5 minutes to the total processing time.

Age Group Analysis and Limitations

The age-group analysis reveals that the methodology's benefits are consistent across different demographic segments, with some expected variations. The superior performance in younger age groups (MAE: 3.2 years) compared to elderly populations (MAE: 4.1 years) reflects the inherent challenges of age estimation in older individuals, where visual aging patterns become more variable and less predictable. This limitation suggests that domain-specific considerations remain important even with improved data filtering.

Statistical Robustness and Generalizability

The statistical robustness of the results, confirmed through paired t-tests ($p < 0.001$) and cross-validation experiments, provides confidence in the methodology's reliability. The consistency of improvements across multiple independent runs (MAE improvements ranging from 1.2 to 1.6 years) indicates that the observed benefits are not dependent on specific dataset partitioning or initialization conditions.

However, it is important to acknowledge that the evaluation was conducted on a single domain (age estimation) with specific dataset characteristics. The generalizability of these results to other machine learning domains and different types of data quality issues requires further investigation.

Methodology Limitations

Several limitations should be acknowledged in this research. The reliance on pre-trained models introduces potential domain bias, particularly when the target application differs considerably from the training domains of the baseline models. The threshold optimization process requires domain-specific calibration through ROC analysis, which may limit the methodology's immediate transferability across different applications without proper validation. While the overall pipeline shows efficiency gains, the initial confidence scoring phase introduces computational overhead that may be significant in resource-constrained environments. The methodology still requires human supervision for borderline cases and threshold validation, which may limit its fully automated application. Finally, the current evaluation focuses solely on age estimation, and broader validation across different machine learning domains is needed to establish general applicability.

Implications for Data Quality Assessment

The results suggest that automated confidence scoring can serve as an effective proxy for data quality assessment, particularly when combined with multi-model consensus mechanisms that leverage the complementary strengths of different architectural approaches. The statistical validation of the methodology through comprehensive experimental evaluation establishes a foundation for understanding the relationship between confidence scores and actual data quality, providing practitioners with quantitative tools for making informed decisions about data inclusion in training flows. However, the approach should be viewed as a tool to assist rather than replace human expertise in data curation, especially for critical applications where data quality directly impacts system safety, fairness, or regulatory compliance. The methodology's strength lies in its ability to efficiently identify potentially problematic samples for human review, thereby optimizing the allocation of limited human annotation resources while maintaining necessary oversight for high-stakes applications. This human-in-the-loop approach ensures that the benefits of automated filtering are realized without compromising the rigor required for mission-critical deployments. The broader implications extend beyond immediate performance improvements to encompass considerations of data governance, reproducibility, and ethical AI development. By providing transparent, quantitative measures of data quality through confidence scoring, the methodology contributes to more accountable machine learning practices that can better support requirements for algorithmic transparency and explainability in regulated industries.

Limitations of CLIP-Based Confidence Scoring

While CLIP-based semantic validation provides valuable insights into the alignment between visual content and textual age descriptions, several limitations should be acknowledged. First, CLIP's performance depends on the quality and diversity of its pre-training data, which may introduce biases toward certain demographic groups, age ranges, or cultural contexts that were well-represented during CLIP's training phase. This limitation is particularly relevant for age estimation tasks, where cultural and demographic factors significantly influence age perception and labeling practices.

Second, CLIP's semantic similarity scores may not always correlate directly with prediction accuracy, especially when visual features relevant to age estimation (such as fine-grained facial details) differ from the high-level semantic features that CLIP captures. The model's reliance on cosine similarity in a shared embedding space may miss domain-specific nuances that are critical for accurate age estimation but less prominent in general visual-semantic understanding.

Third, the prompt engineering approach, while effective, introduces subjectivity in threshold selection and prompt formulation. The choice of age range windows (e.g., ± 2 years for positive prompts, ± 5 years for negative prompts) and prompt phrasing can significantly impact confidence scores, requiring domain expertise and empirical validation. This dependency on prompt design limits the fully automated application of the methodology and may require iterative refinement for different domains or applications.

Fourth, CLIP's computational requirements, while manageable for the current scale, may become a bottleneck when processing extremely large datasets or when real-time inference is required. The need to compute embeddings for both images and multiple text prompts per sample increases computational overhead compared to prediction-only approaches.

Finally, CLIP's effectiveness may degrade in scenarios with significant domain shift between the target application and CLIP's training distribution, or when dealing with specialized domains (e.g., medical imaging, satellite imagery) where semantic relationships differ substantially from natural image-text pairs. Future should explore domain-adaptive CLIP variants or alternative semantic validation mechanisms for such specialized applications.

7. Conclusion and Future Study

This research introduces a multi-model confidence scoring approach for automated data quality assessment in AI training. The study focuses on combining predictions from Buffalo-L and a custom age head with CLIP-based semantic validation to score confidence, yielding an evaluation that reflects both inter-model agreement and content consistency. Building on this integration, we define a multi-stage procedure that fuses complementary model outputs with CLIP-supported checks and apply it to noisy, real-world data. We further add a ROC-based threshold selection strategy to balance quality and retention, avoiding ad-hoc cutoffs; in practice, a threshold of 0.8 performed well. Empirically, confidence-guided

selection reduced MAE by 29% and training time by 40%, challenging the assumption that more data always helps. Although demonstrated on age estimation, the modular design generalizes to other weakly supervised settings and addresses a common need across industry: improving training data quality at scale.

Future Research Directions

There are numerous avenues for future research grounded in this study's findings and limitations. First, cross-domain validation is necessary, wherein this methodology is assessed and adapted for use in different sub-fields of machine learning such as natural language processing, speech processing, and medical imaging. This will require further testing and refinement of training and evaluation across domains. This can provide confidence signals to refine data selection across domains. Further, adaptive, confidence-driven thresholding may improve selection decisions. These systems would further remove the reliance on domain-expert threshold setting. Active learning would further reduce the burden of human supervision. Integrating confidence-based selection with active learning could streamline data curation and reduce human effort. Scalability analysis involves understanding how methodologies perform in terms of efficiency and computational demand when they operate on datasets significantly bigger than those utilized in this study. Such analysis seeks to identify potential bottlenecks and define required refinements to fully optimize the methodologies for real-world industrial scale implementation. Regarding the framework's extensibility to support multi-modal datasets and more sophisticated prediction tasks, this research will investigate how confidence-based selection principles can be adapted to multi-modal settings, perhaps incorporating cross-modality coherence as an additional component of interest. Lastly, systematic comparative research analyses data quality improvement methods, such as data augmentation, synthetic data generation, and statistical filtering, and provides an empirical basis to ascertain the relative gains of the different methods and the suitable combinations for specified application scenarios.

Final Remarks

The proposed multi-model confidence scoring framework addresses one of the most pressing challenges in machine learning: identifying high-quality training data from noisy, weakly supervised datasets. The demonstrated improvements in model accuracy (29% reduction in MAE) and computational efficiency (40% reduction in training time) highlight the practical value of data-driven quality assessment over volume-based approaches. However, the current evaluation is limited to a single domain (age estimation), and broader validation across diverse machine learning applications is necessary to establish general applicability. Future research should focus on cross-domain validation, development of more flexible and domain-independent scoring systems, and integration with active learning frameworks to further reduce human supervision requirements. The methodology's modular design enables adaptation to various machine learning tasks, but successful deployment requires domain-specific calibration and validation. As automated confidence scoring systems mature, they will play an increasingly important role in enabling efficient and effective machine learning workflows, particularly in resource-constrained environments where computational efficiency and data quality are critical concerns.

Acknowledgments

The authors acknowledge that some of the findings presented in this paper were shared within the scope of the 9th International Conference on Mathematical Sciences (ICMS 2025) at Maltepe University, Istanbul, Turkey.

References

[1] E. Şahinaslan, M. Günerkan, and Ö. Şahinaslan., *An alternative solution method for the use of categorical data encoding technique in machine learning. jista*, vol. 6, pp. 1–11, 2023 , doi: 10.38016/jista.1140499.

[2] C. Northcutt, L. Jiang, and I. Chuang., *Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research*, vol. 70, pp. 1373-1411, (2021). doi: 10.1613/jair.1.12125

[3] P. Chen, B. B. Liao, G. Chen, and S. Zhang., *Understanding and utilizing deep neural networks trained with noisy labels* (2019). arXiv preprint arXiv:1905.05040. doi: 10.48550/arXiv.1905.05040

[4] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu., *Making deep neural networks robust to label noise: A loss correction approach*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1944-1952, (2017). doi: 10.1109/CVPR.2017.240

[5] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama., *Co-teaching: Robust training of deep neural networks with extremely noisy labels.* In *Advances in Neural Information Processing Systems (NeurIPS)*, vol.31 (2018). doi: 10.48550/arXiv.1804.06872

[6] J. Li, R. Socher, and S. C. H. Hoi., *DivideMix: Learning with noisy labels as semi-supervised learning.* In *International Conference on Learning Representations (ICLR)*, (2020). doi: 10.48550/arXiv.2002.07394

[7] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama., *How does disagreement help generalization against label corruption?* In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 7164-7173, (2019). doi: 10.48550/arXiv.1901.04215

[8] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich., *Training deep neural networks on noisy labels with bootstrapping* (2014). arXiv preprint arXiv:1412.6596. doi: 10.48550/arXiv.1412.6596

[9] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour., *Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis*, (2020). *Medical Image Analysis*, vol. 65, 101759. doi: 10.1016/j.media.2020.101759

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever., *Learning transferable visual models from natural language supervision* (2021). arXiv preprint arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020

[11] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan., *MaPLe: Multi-modal prompt learning.* In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023). doi: 10.48550/arXiv.2210.03117

[12] H.-T. Lin, C.-J. Lin, and R. C. Weng., *A note on Platt's probabilistic outputs for support vector machines. Machine Learning*, vol. 68, pp. 267–276 (2007). doi: 10.1007/s10994-007-5018-6

[13] T. Hofmann, B. Schölkopf, and A. J. Smola., *Kernel methods in machine learning. Annals of Statistics*, 36(3), 1171–1220, (2008). doi: 10.1214/009053607000000677

[14] K. Nandakumar, A. K. Jain, and A. Ross., *Score normalization in multimodal biometric systems. Pattern Recognition*, vol. 41, no. 12, pp. 2623–2632, (2008). doi: 10.1016/j.patcog.2008.03.024

Ersin Arikan,
Istanbul Okan University
Graduate Education Institute
Turkey.
E-mail address: erarikan@stu.okan.edu.tr

and

Ender Sahinaslan
Mudanya University
Department of Management Information Systems
Turkey.
E-mail address: ender.sahinaslan@mudanya.edu.tr