Ensemble Learning for Multi-Task Road Safety Attributes Classification

Muhammad Shahid*[1], Marko Ševrović[1], and Monica Olyslagers[2]

[1]Faculty of Transport and Traffic Sciences, University of Zagreb, Zagreb, Croatia
[2]International Road Assessment Programme (iRAP), London, United Kingdom

Short Summary

Road traffic crashes claim 1.19 million lives annually and remain the leading cause of death for people aged 5–29 years. Safe road infrastructure is among the most effective countermeasures for reducing this burden. The International Road Assessment Programme (iRAP) offers a globally recognised framework that requires trained coders to manually identify up to 50 physical road attributes every 10-metres from survey images for road infrastructure assessment and improvements. While the iRAP methodology provides a robust and globally validated framework for infrastructure safety assessment, the current reliance on manual coding can be resource-intensive and may limit scalability, particularly in low- and middle-income contexts. Although computer vision approaches have been proposed for automated attribute classification, few studies have systematically examined ensemble learning to enhance performance across multiple safety attributes. This paper addresses this gap by evaluating ensemble voting strategies for road attributes classification task. Results show that ensemble methods outperform individual baseline models.
Keywords: automated road assessment, ensemble learning, iRAP, multi-task classification, vision transformers, road safety

1 Introduction

Road crashes remain a critical global public health challenge, claiming 1.19 million lives each year. The burden falls disproportionately on low- and middle-income countries, which account for 92 per cent of all road traffic fatalities despite having a fraction of the world's registered vehicles (WHO, 2023). A substantial share of road crashes is attributable to inadequate road infrastructure: poor geometric design, insufficient delineation, lack of pedestrian facilities, and absence of appropriate roadside treatments are well-established risk factors for crash occurrence and severity (Papadimitriou et al., 2019) . Designing safer roads and prioritising infrastructure investment therefore requires accurate, large-scale road infrastructure assessment.

The International Road Assessment Programme (iRAP) provides a Star Rating System to assess the inherent safety of roads, assigning ratings from 1 star (highest risk) to 5 stars (lowest risk) for different road users (iRAP Star Rating Manual, 2024). To perform road infrastructure assessment, trained and accredited human coders inspect geo-referenced road survey videos and record values for more than 50 physical road attributes for each 100-metre segment in accordance with the iRAP Coding Manual (iRAP Coding Manual, 2025). This attribute coding underpins iRAP's Star Rating models, which translate infrastructure characteristics into evidence-based risk scores. These scores are used by governments and road agencies worldwide to develop Safer Roads Investment Plans and prioritise cost-effective safety interventions. However, the manual coding process remains costly, time-consuming, and susceptible to coder variability, thereby limiting the scalability of formal road assessments, particularly in low- and middle-income countries.

Recent research has begun to address this bottleneck through automated computer vision approaches. Deep learning models trained on road survey imagery have demonstrated promising performance in classifying road attributes (Kačan et al., 2024; Shahid et al., 2025b). However, in our preliminary studies (Shahid et al., 2025b) we found that these road safety attributes are structurally diverse, ranging from easily recognisable geometric features to more complex transient, low-frequency scene elements that challenge any individual deep learning model. As a result, no single deep learning model reliably excels across this full diversity of classification tasks simultaneously.

Ensemble learning offers a promising way to overcome this limitation. By combining predictions from multiple independently trained models, ensemble methods can exploit complementary error

patterns and reduce task-specific failure modes (Ganaie et al., 2022). Despite the well-established efficacy of ensemble approaches in multi-task classification settings, their systematic application to iRAP road safety attribute classification has not been reported in the literature.
To address this gap, this paper makes the following contributions:

1. We present the first study to systematically apply and compare ensemble learning strategies for automated multi-task iRAP road safety attribute classification from road survey videos, covering 23 attributes across a large-scale dataset of 199,233 10-meter road segments.
2. We implement and evaluate various ensemble voting strategies spanning training-free approaches (Majority Vote, Confidence Vote), validation-calibrated methods (Weighted Average Probability, Task-Specific Best), and a meta-learning strategy (Stacking), evaluated against individual backbone models to quantify the benefit of each aggregation approach.
3. All strategies are assessed using comprehensive performance metrics providing a holistic view of classification quality.

2 Methodology

Dataset Details

We evaluate our approach on a multi-task dataset comprising 199,233 annotated 10-meter road segments extracted from survey videos covering 2,300 km of public roads in Bosnia and Herzegovina. The dataset was created by aligning coded iRAP attributes (50+ attributes per segment) from an official coding campaign with corresponding video frames using GPS endpoint information, as described in (Shahid et al., 2025b). We use a subset of $T = 23$ iRAP classes and split the segments into 186,796 training, 6,593 validation, and 5,844 test samples while keeping all segments from the same video in the same partition to avoid data leakage.

Model Pipeline

Each 10-metre road segment is represented not by a single frame but by a short temporal sequence. Concretely, the input to the model is a batch of such sequences, $X \in \mathbb{R}^{B\times F\times C\times H\times W}$, where $F = 3$ frames are sampled at temporal offsets $[-4, -1, 0]$, corresponding to a spatial context window of $[-40\,\text{m}, -10\,\text{m}, 0\,\text{m}]$. This offset configuration was determined through a systematic ablation study comparing different frame offset combinations. To pass all frames through a single backbone in one forward pass, the batch is reshaped to $X' \in \mathbb{R}^{(B\cdot F)\times C\times H\times W}$, treating each frame as an independent image.
The classification target consists of $T = 23$ road attributes, each with $C_t \in \{3, \ldots, 7\}$ classes. Because the joint prediction space $\prod_{t=1}^{T} C_t$ is intractably large, the problem is decomposed into T independent classification tasks, each solved by a dedicated head attached to a shared backbone. Two backbone architectures are evaluated for this role BEiT and ConvNextV2 (Table 1).

Table 1: Pre-Trained models used with their configurations

Model	Size	Resolution	Params (Million)	Pre-training & Fine-tuning
BEiT (Bao et al., 2021)	Base	384×384	86.7	ImageNet-22k–ImageNet-22k
ConvNextV2 (Woo et al., 2023)	Base	384×384	88.7	ImageNet-22k–ImageNet-1K

Both backbones share the same feature extraction strategy and are kept frozen throughout training, acting as multi-scale feature extractors. For each input frame $\mathbf{x}_f \in \mathbb{R}^{3\times H\times W}$, feature maps from all intermediate backbone stages are spatially collapsed via global average pooling and concatenated channel-wise into a single vector of dimension $D = \sum_{i=1}^{N} C_i$, where C_i is the channel depth at stage i. The three per-frame vectors are then averaged to produce a single temporally-fused representation $\mathbf{z} \in \mathbb{R}^D$ for the segment:

$$\mathbf{z} = \frac{1}{F}\sum_{f=1}^{F} \text{GAP}(\text{Cat}[\text{Backbone}(\mathbf{x}_f)]) \tag{1}$$

where GAP denotes global average pooling and Cat denotes channel-wise concatenation of multi-scale features.

This fused vector $\mathbf{z}$ is then fed independently to each of the T task heads. Each head is a three-layer MLP with LayerNorm and GELU activations, producing the logits for its assigned attribute:

$$\hat{\mathbf{y}}_t = W_t^{(3)}\,\mathrm{GELU}\Big(\mathrm{LN}\Big(W_t^{(2)}\,\mathrm{GELU}\Big(\mathrm{LN}\Big(W_t^{(1)}\mathbf{z}\Big)\Big)\Big)\Big) \tag{2}$$

where $W_t^{(\cdot)}$ are task-specific weight matrices and LN is LayerNorm.
During training, each backbone and its attached heads are fine-tuned independently. The model is evaluated on the validation set after every epoch, and training stops early when validation macro-F1 fails to improve for three consecutive epochs; the best-performing checkpoint is retained. At inference, both saved checkpoints are loaded with all weights frozen and applied independently to the test set. Each model $m \in \{1, 2\}$ produces a per-task softmax probability vector $\mathbf{p}_m^{(t)} \in \Delta^{C_t}$ for every segment and every attribute $t \in \{1, \ldots, T\}$. These $M \times T$ probability vectors are the inputs to the ensemble aggregation layer, which combines them using the strategies described in Section 2 to produce the final attribute predictions $\hat{y}^{(t)}$ across all $T = 23$ tasks. The complete two-phase pipeline is illustrated in Figure 1.

Training Details

Both models are fine-tuned independently under an identical hyperparameter configuration to ensure a fair comparison and controlled ensemble diversity. Each model is trained for 20 epochs with early stopping based on validation macro-F1, with a patience of 3 epochs to prevent overfitting while retaining the best-performing checkpoint. Optimisation uses AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$ on a single NVIDIA GeForce RTX 4080 GPU. A decoupled weight decay of $\lambda = 5 \times 10^{-3}$ is applied to non-bias, non-normalisation parameters; bias and batch-normalisation parameters are excluded from weight decay. The learning rate is set to 2×10^{-5} for both the backbone and task-head parameters. Training employs mixed-precision arithmetic (FP16) via PyTorch's automatic mixed-precision (AMP) with gradient scaling to prevent underflow. Gradients are accumulated over 2 steps. A drop-path rate of 0.2 and a dropout rate of 0.4 are applied for regularisation.

Ensemble Strategies

In this paper we have evaluated various ensemble voting strategies. Let M=2 be the number of ensemble members, T the number of tasks, and $\mathbf{p}_m^{(t)} \in \Delta^{C_t}$ the softmax probability vector from model m for task t on a given sample. Majority Vote and Confidence Vote constitute the hard-voting strategies, as they operate on discrete class predictions, whereas Weighted Average Probability, Task-Specific Best, and Stacking are soft-voting strategies that operate directly on predicted class probabilities.

Training-Free Strategies

1. Majority Vote (MV). Majority voting is the most widely used hard-voting combination rule. Each model independently predicts the most probable class label for a given sample, and the ensemble selects the class that receives the plurality of votes across all M models. Formally, each model casts a hard vote $\hat{c}_m^{(t)} = \arg\max_c p_{m,c}^{(t)}$, and the ensemble selects the plurality class (Kokkinos and Margaritis, 2014):

$$\hat{y}_{\mathrm{MV}}^{(t)} = \arg\max_c \sum_{m=1}^{M} \mathbf{1}\Big[\hat{c}_m^{(t)} = c\Big] \tag{3}$$

Majority voting is training-free and requires no validation data; its limitation is that it ignores the confidence of individual predictions, treating all votes as equally reliable regardless of model certainty (Ganaie et al., 2022).
2. Confidence Vote (CV). Confidence voting is a hard-voting variant that resolves disagreements by deferring to the most confident model rather than the plurality (Stanovov et al., 2020). When all models agree, the shared prediction is used; when they disagree, the model with the highest peak softmax probability is selected:

$$\hat{y}_{\mathrm{CV}}^{(t)} = \hat{c}_{m^*}^{(t)}, \quad m^* = \arg\max_m \max_c p_{m,c}^{(t)} \tag{4}$$

This rule is particularly effective when base models are well-calibrated, so that high confidence reliably signals correct predictions.

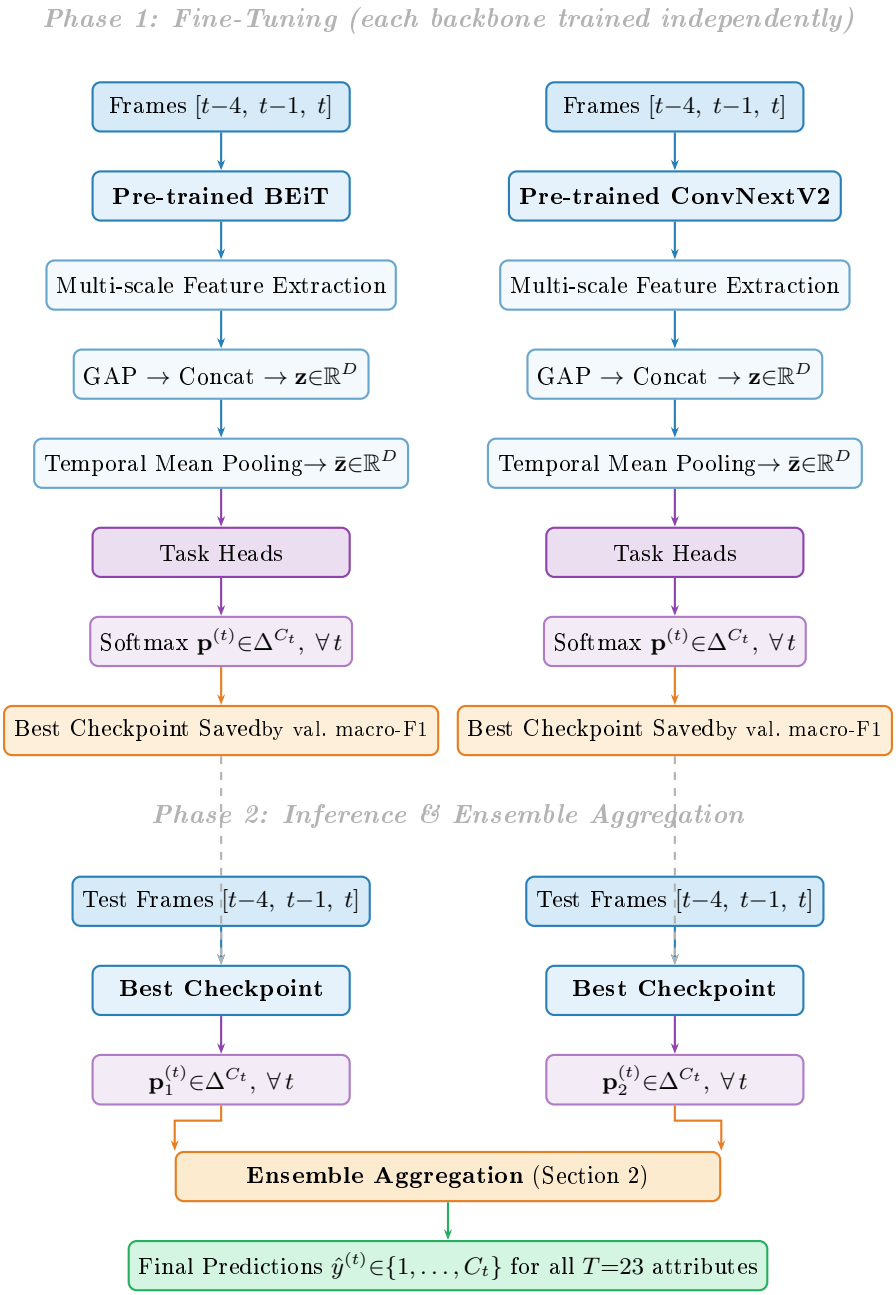


Figure 1: Two-phase pipeline for multi-task iRAP road safety attribute classification. Both models independently produce per-task softmax distributions $\mathbf{p}_m^{(t)}$, which ensemble strategies aggregate into final predictions across all T=23 attributes.

Validation-Calibrated Strategies

Let $\mathcal{V}$ denote a labelled validation set different from both training and test sets. Backbone weights are frozen; $\mathcal{V}$ is used only to derive scalar mixing coefficients.

3. Weighted Average Probability (WAP). Weighted soft voting extends the average probability rule by assigning higher weight to models with stronger overall validation performance (Dogan and Birant, 2019). Global per-model accuracy a_m on $\mathcal{V}$ serves as the mixing weight:

$$\hat{y}_{\text{WAP}}^{(t)} = \arg\max_c \sum_{m=1}^{M} a_m \cdot p_{m,c}^{(t)}, \quad a_m = \text{Acc}_m(\mathcal{V}) \tag{5}$$

4. Task-Specific Best (TSB). In a multi-task setting, different models may excel on different subtasks due to varying task affinities and the effects of negative transfer (Standley et al., 2020). TSB exploits this by selecting, for each task independently, the single best model according to

per-task validation accuracy:

$$\hat{y}_{\mathrm{TSB}}^{(t)} = \hat{c}_{m^*(t)}^{(t)}, \quad m^*(t) = \arg\max_m \mathrm{Acc}_m^{(t)}(\mathcal{V}) \tag{6}$$

Meta-Learning Strategy

5. Stacking (ST). Stacked generalisation, introduced by Wolpert (1992), is a meta-learning approach in which a second-level learner is trained to combine the outputs of the base models. Unlike voting rules, stacking learns a meta-level combiner that adapts to base model outputs and is particularly effective when base learners are diverse (Džeroski and Ženko, 2004). We trained a per-task logistic regression meta-learner g_t on concatenated base model probability vectors from $\mathcal{V}$. The meta-feature vector for task t concatenates the M softmax probability vectors, yielding a meta-input of dimension $M \times C_t$:

$$\mathbf{u}^{(t)} = \left[\mathbf{p}_1^{(t)}; \mathbf{p}_2^{(t)}; \ldots; \mathbf{p}_M^{(t)}\right] \in \mathbb{R}^{M \cdot C_t} \tag{7}$$

The meta-learner g_t is trained to minimise the cross-entropy loss on $\mathcal{V}$:

$$\hat{y}_{\mathrm{ST}}^{(t)} = g_t\left(\mathbf{u}^{(t)}\right), \quad g_t = \arg\min_g \sum_{(\mathbf{u},y)\in\mathcal{V}} \mathcal{L}_{\mathrm{CE}}\left(g\left(\mathbf{u}^{(t)}\right), y^{(t)}\right) \tag{8}$$

StandardScaler normalisation is applied to $\mathbf{u}^{(t)}$ before fitting g_t to prevent feature scale dominance.

Evaluation Metrics

All ensemble strategies described above were evaluated using a comprehensive set of metrics that assess discriminative performance across multiple complementary dimensions. The definitions and equations for accuracy, precision, recall, and F1-score are provided in our previous work (Shahid et al., 2025a). For completeness, we briefly restate the notation used throughout the evaluation. Let N be the number of test samples, T the number of attribute classification tasks, C_t the number of classes for task t, $y_i^{(t)}$ the ground-truth label, and $\hat{y}_i^{(t)}$ the predicted class for the i-th sample on task t. The remaining metrics and their mathematical formulations are defined as follows:

Cohen's κ. Measures agreement between predictions and ground truth, correcting accuracy for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

where p_o is the observed accuracy and p_e is the expected agreement under the hypothesis that predictions and labels are independent:

$$p_e = \sum_c \frac{n_c^{\mathrm{pred}} \cdot n_c^{\mathrm{true}}}{N^2} \tag{10}$$

with n_c^{pred} and n_c^{true} denoting the marginal counts of class c in predictions and ground truth, respectively.

3 Results and Discussion

Table 2 reports comprehensive metrics across the six-metric evaluation suite for both individual backbone models and all ensemble strategies. Figure 2 provides a complementary multi-metric view of all ensemble strategies via radar charts.

The results show that Majority Vote (MV), which relies solely on hard label plurality, closely mirrors the weaker BEiT backbone (macro-F1: 0.6027 vs. 0.6023) and fails to exploit ConvNextV2's superior probability estimates, confirming that hard-decision aggregation loses calibration information critical for imbalanced multi-class tasks. Confidence Vote (CV) recovers this information by weighting votes by predicted class probability, matching ConvNextV2's overall performance (macro-F1: 0.7020) while providing slight recall improvements (0.7258 vs. 0.7259).

Among validation-calibrated strategies, Weighted Average Probability (WAP) achieves the highest macro-F1 (0.7063), macro-recall (0.7283), Cohen's κ (0.6611), and MCC (0.6664) of all deployable methods. Stacking (LR) attains the highest overall accuracy (0.8673) and macro-precision (0.7076) but at the cost of reduced recall (0.6956) and macro-F1 (0.6906), reflecting its tendency

Table 2: Overall model result comparison between pre-trained backbones and Ensemble strategies.

Strategy	**Acc.**	**Macro-F1**	**Macro-P**	**Macro-R**	κ	**MCC**
BEiT	0.7729	0.6023	0.6187	0.6407	0.4822	0.5019
ConvNext-V2	0.8606	0.7042	0.6994	0.7259	0.6600	0.6647
Majority Vote	0.7730	0.6027	0.6188	0.6423	0.4824	0.5023
Confidence Vote	0.8579	0.7020	0.6974	0.7258	0.6580	0.6634
Weighted Avg. Prob.	0.8601	**0.7063**	0.7036	**0.7283**	**0.6611**	**0.6664**
Task-Specific Best	0.8608	0.7032	0.6993	0.7224	0.6572	0.6608
Stacking (LR)	**0.8673**	0.6906	**0.7076**	0.6956	0.6459	0.6528

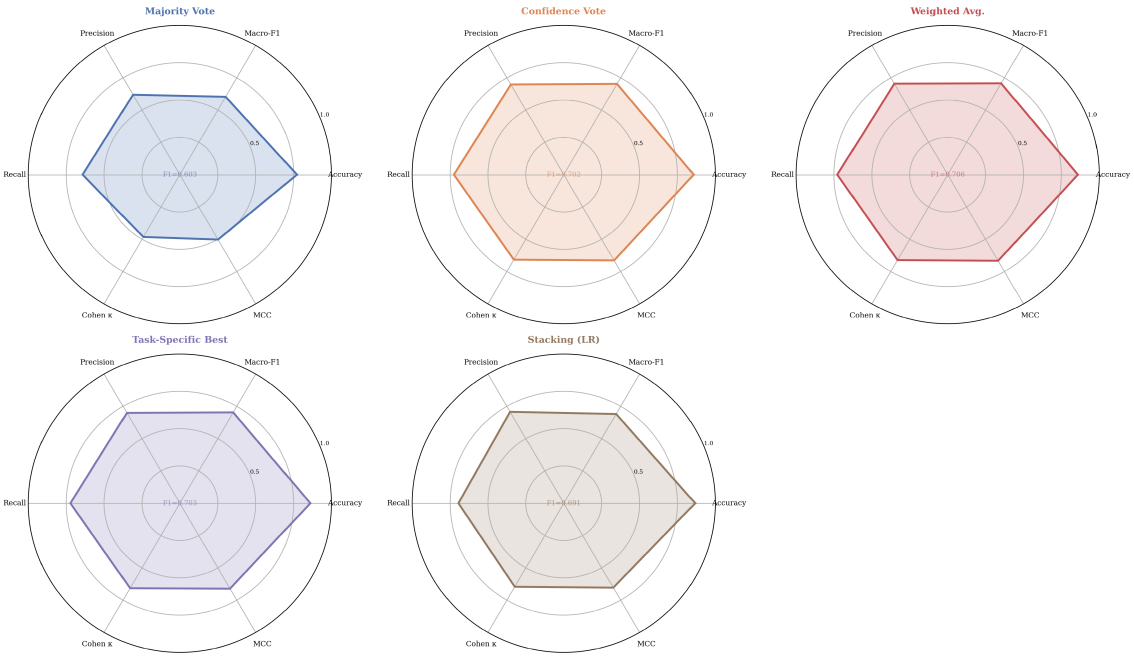


Figure 2: Per-strategy multi-metric radar charts across evaluation dimensions. Each polygon represents one ensemble strategy; the central label reports macro-F1.

to predict conservative majority-class outputs on low-frequency attributes. Across all five strategies, validation-calibrated soft-voting methods (WAP, TSB) consistently outperform training-free baselines.

Per-Attribute Ensemble Gains

Figure 3 presents a heatmap of per-attribute macro-F1 scores for all ensemble strategies, with attributes ordered from hardest (top) to easiest (bottom) by mean F1 across strategies. The heatmap shows two distinct attribute groups: a structurally ambiguous group (e.g., *Roadside severity driver-side object* attr_12: F1 $\approx$ 0.2996–0.4580, *Road condition* attr_5: F1 $\approx$ 0.4000–0.5121) where all strategies converge at low values, and a saturated group (e.g., *Bicycle facility* attr_22: F1 $\approx$ 0.9968–0.9998, *Delineation* attr_9: F1 $\approx$ 0.8665–0.9781) where all strategies reach near-ceiling performance. The MV column shows that hard-voting systematically underperforms soft-voting strategies across virtually every attribute. WAP and TSB produce the most uniformly dark columns (High F1 Score) in the mid-range, while Stacking (ST) shows relative weakness on transient attributes (e.g., attr_19: ST 0.6217 vs. TSB 0.7387, attr_3: ST 0.7062 vs. WAP 0.6919) despite its high overall accuracy.

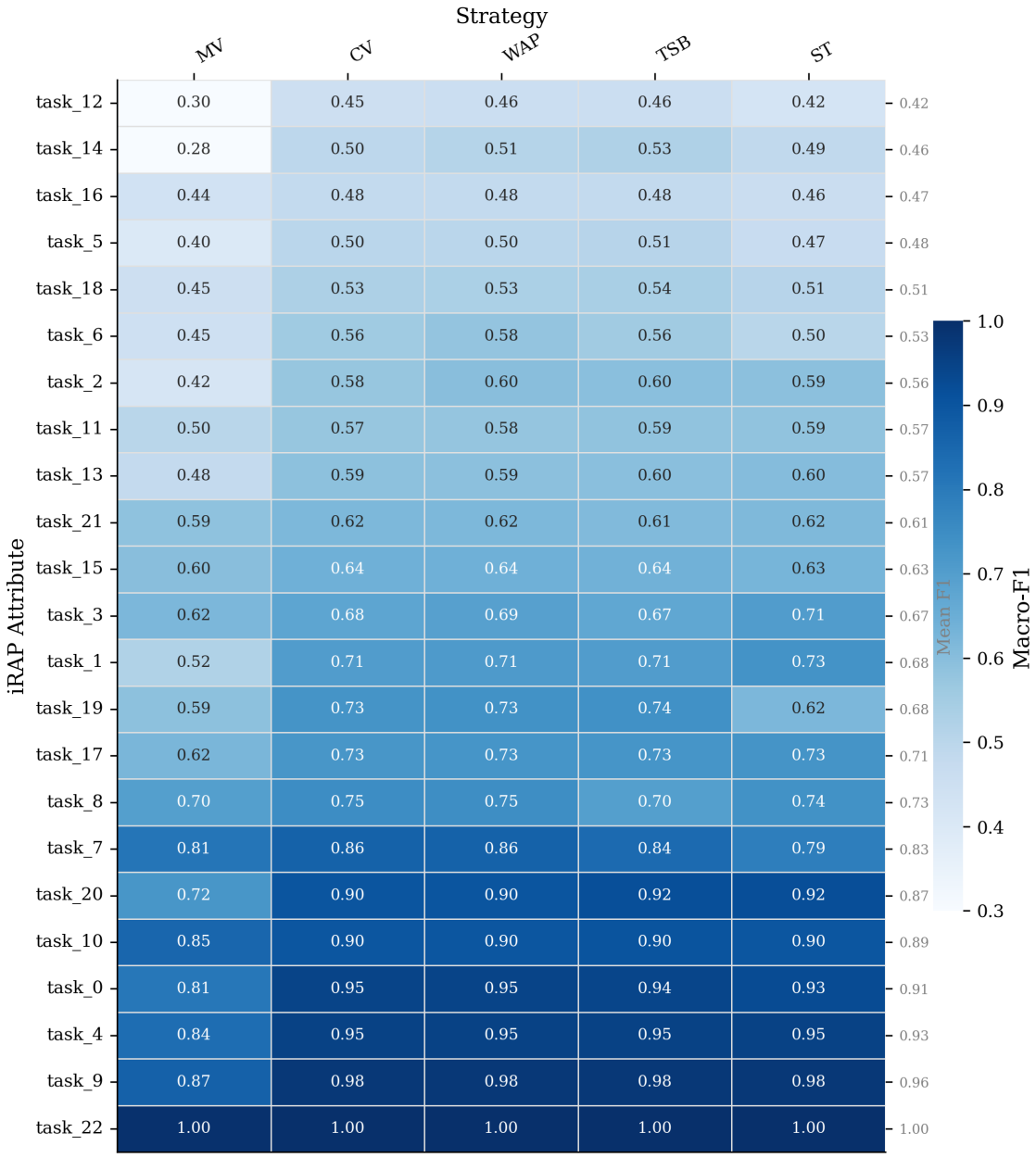


Figure 3: Per-attribute macro-F1 heatmap for all ensemble strategies. Attributes are sorted from hardest (top) to easiest (bottom) by mean F1 across ensemble strategies.

Where Ensemble Learning Improves Results: Attribute-Level Analysis

A central contribution of this work is identifying *which* attributes benefit from ensemble learning and *which strategy* provides the greatest gain. Three distinct improvement patterns emerge from the per-attribute analysis (Table 3 and Figure 3):

High-Frequency Saturated Attributes. Attributes such as *Bicycle facility* (attr_22), *Delineation* (attr_9), and *Number of lanes* (attr_0) achieve macro-F1 > 0.90 with individual models and all ensemble strategies alike, as they exhibit visually salient, low-ambiguity features that both backbones recognise reliably. For these attributes, the marginal gain from ensemble combination is negligible ($\Delta < 0.01$), as all base learners already approach saturation.

Medium-Saliency Geometric Attributes. Attributes such as *Lane width* (attr_1), *Sight distance* (attr_8), *Property access points* (attr_17), and *Quality of curve* (attr_3) exhibit moderate individual-model accuracy where the two backbones make complementary errors. WAP and TSB provide the most consistent gains ($\Delta \approx 0.02$–0.05) on this group. Stacking achieves the highest single-attribute F1 on *Lane width* (0.7339) and *Quality of curve* (0.7062), surpassing both individual models.

Table 3: Per-attribute macro F1 scores: individual models and ensemble strategies.

		Individual Models		Ensemble Learning Methods				
ID	Attribute Name	BEiT	ConvNext	MV	CV	WAP	TSB	ST
attr_0	Number of lanes	0.8078	0.9450	0.8074	0.9457	**0.9462**	0.9434	0.9263
attr_1	Lane width	0.5206	0.7048	0.5229	0.7070	0.7118	0.7069	**0.7339**
attr_2	Curvature	0.4211	0.5956	0.4188	0.5757	**0.5996**	0.5955	0.5947
attr_3	Quality of curve	0.6176	0.6784	0.6211	0.6801	0.6919	0.6738	**0.7062**
attr_4	Median Type	0.8368	0.9521	0.8358	0.9508	0.9508	0.9511	**0.9530**
attr_5	Road condition	0.4000	**0.5121**	0.4008	0.4997	0.5013	0.5098	0.4676
attr_6	Vehicle parking	0.4398	0.5554	0.4454	0.5571	**0.5796**	0.5556	0.5027
attr_7	Roadworks	0.8104	0.8352	0.8137	**0.8642**	**0.8642**	0.8418	0.7909
attr_8	Sight distance	0.7009	0.7245	0.6978	**0.7487**	0.7472	0.6978	0.7375
attr_9	Delineation	0.8665	0.9765	0.8687	0.9759	0.9761	0.9763	**0.9781**
attr_10	Street lighting	0.8535	**0.9022**	0.8530	0.8977	0.8987	0.9020	0.8988
attr_11	Roadside severity – driver-side dist.	0.5010	0.5823	0.5048	0.5745	0.5768	0.5851	**0.5865**
attr_12	Roadside severity – driver-side obj.	0.2996	0.4577	0.2996	0.4509	0.4567	**0.4580**	0.4239
attr_13	Roadside severity – pass.-side dist.	0.4760	0.5937	0.4790	0.5903	0.5931	0.5970	**0.6045**
attr_14	Roadside severity – pass.-side obj.	0.2801	**0.5272**	0.2811	0.4968	0.5119	0.5253	0.4870
attr_15	Paved shoulder – driver-side	0.5987	0.6407	0.5982	0.6420	**0.6422**	0.6398	0.6295
attr_16	Paved shoulder – passenger-side	0.4351	0.4810	0.4352	0.4820	**0.4823**	0.4808	0.4640
attr_17	Property access points	0.6253	0.7262	0.6236	0.7261	0.7272	0.7272	**0.7320**
attr_18	Land use — driver-side	0.4514	0.5370	0.4510	0.5314	0.5348	**0.5384**	0.5081
attr_19	Land use — passenger-side	0.5956	0.7384	0.5921	0.7320	0.7350	**0.7387**	0.6217
attr_20	Sidewalk — driver-side	0.7264	0.9187	0.7246	0.9002	0.9014	0.9155	**0.9206**
attr_21	Sidewalk — passenger-side	0.5927	0.6140	0.5897	**0.6178**	**0.6178**	0.6137	0.6158
attr_22	Bicycle facility	0.9967	0.9987	0.9968	0.9993	0.9993	0.9991	**0.9998**
Macro F1 Avg.		0.6023	0.7042	0.6027	0.7020	**0.7063**	0.7032	0.6906

Low-Frequency Scene Attributes. Attributes with sparse positive labels, *Roadside severity driver-side object* (attr_12), *Roadside severity passenger-side object* (attr_14), *Vehicle parking* (attr_6), and *Road condition* (attr_5), present the hardest classification challenge (macro-F1 $\approx$ 0.2996–0.5121) due to class imbalance and transient or fine-grained scene content. TSB provides the greatest relative gains for attr_12 (0.4580 vs. MV: 0.2996) and attr_14 (0.5253 vs. MV: 0.2811), while WAP is the most reliable across the broader group, achieving 0.5796 on attr_6 and 0.5013 on attr_5.

4 Conclusions

This paper presented the first systematic evaluation of ensemble voting strategies for automated iRAP multi-task road safety attribute classification across 23 attributes. Results show that no single strategy dominates uniformly: Weighted Average Probability (WAP) achieves the highest macro-F1 (0.7063), macro-recall (0.7283), Cohen's κ (0.6611), and MCC (0.6664) among deployable strategies, reflecting well-calibrated soft-voting that effectively combines the complementary strengths of BEiT and ConvNextV2; Task-Specific Best (TSB) matches WAP closely; and Stacking (LR) attains the highest overall accuracy (0.8673) and macro-precision (0.7076) by learning conservative, precision-oriented combinations, though at the cost of reduced recall (0.6956) and macro-F1 (0.6906). Training-free baselines, Majority Vote and Confidence Vote, are consistently outperformed once a modest validation set is available for calibration, with Majority Vote closely mirroring the weaker BEiT backbone alone (macro-F1: 0.6027 vs. 0.6023).

Per-attribute analysis (Table 3, Figure 3) reveals three different patterns. Saturated high-saliency attributes (*Bicycle facility*, *Delineation*, *Number of lanes*) reach macro-F1 > 0.90 across all strategies, leaving negligible room for ensemble gains. Medium-saliency geometric attributes (*Lane width*, *Sight distance*, *Property access points*) benefit most consistently from WAP and TSB ($\Delta \approx 0.02$–0.05), while Stacking achieves peak single-attribute scores on select tasks. Structurally ambiguous, low-frequency attributes (*Roadside severity*, *Vehicle parking*, *Road condition*) remain the hardest to classify (macro-F1 $\approx$ 0.2996–0.5121).

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101119590. UK participants in Horizon Europe Project IVORY are supported by UKRI grant numbers EP/Y036026/1 (The International Road Assessment Programme – iRAP) and EP/Y036034/1 (Agilysis).

Conflicts of interest

All authors declare no conflicts of interest.

Data Availability

The data used in this study are not publicly available due to confidentiality agreements with the data provider.

References

Bao, H., Dong, L., Piao, S., and Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Dogan, A. and Birant, D. (2019). A weighted majority voting ensemble approach for classification. In *2019 4th international conference on computer science and engineering (UBMK)*, pages 1–6. IEEE.

Džeroski, S. and Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54(3):255–273.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., and Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering applications of artificial intelligence*, 115:105151.

iRAP Coding Manual (2025). *iRAP Coding Manual Version 5.5—Drive on Right Edition*. International Road Assessment Programme (iRAP), London, U.K. Available at `https://resources.irap.org/Specifications/iRAP_Coding_Manual_Drive_on_Right.pdf`.

iRAP Star Rating Manual (2024). *iRAP Star Rating and Investment Plan Manual Version 1.5*. International Road Assessment Programme (iRAP), London, U.K. Available at `https://resources.irap.org/Specifications/iRAP_Star_Rating_and_Investment_Plan_Manual_English.pdf?_gl=1*tesgp9*_ga*MTYyMzA5Nzg4NS4xNzczMzE3ODY0*_ga_HK6PSM29PR*czE3NzQ0NDE3NzMkbzMkZzAkdDE3NzQ0NDE3NzMkajYwJGwwJGgw`.

Kačan, M., Ševrović, M., and Šegvić, S. (2024). Dynamic loss balancing and sequential enhancement for road-safety assessment and traffic scene classification. *Ieee transactions on intelligent transportation systems*, 25(11):15628–15640.

Kokkinos, Y. and Margaritis, K. G. (2014). Breaking ties of plurality voting in ensembles of distributed neural network classifiers using soft max accumulations. In *Ifip international conference on artificial intelligence applications and innovations*, pages 20–28. Springer.

Papadimitriou, E., Filtness, A., Theofilatos, A., Ziakopoulos, A., Quigley, C., and Yannis, G. (2019). Review and ranking of crash risk factors related to the road infrastructure. *Accident Analysis & Prevention*, 125:85–97.

Shahid, M., Gregurić, M., Hassani, A., and Ševrović, M. (2025a). Optimizing car collision detection using large dashcam-based datasets: A comparative study of pre-trained models and hyperparameter configurations. *Applied Sciences*, 15(13):7001.

Shahid, M., Ševrović, M., Olyslagers, M., and Hassani, A. (2025b). A spatio-temporal multi-task framework for irap road attribute classification from street-level videos. In *2025 8th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)*, pages 1–6. IEEE.

Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., and Savarese, S. (2020). Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR.

Stanovov, V., Akhmedova, S., and Kamiya, Y. (2020). Confidence-based voting for the design of interpretable ensembles with fuzzy systems. *Algorithms*, 13(4):86.

WHO (2023). Road traffic injuries. Available at `https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries`.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. (2023). Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142.