أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Spoken Arabic Dialect Identification
NADI 2026 Shared Task 2

Musab Iskandar, Adel Alharbi, Khalid Alharbi,
Mohammed Ebrahim, Youssef Mohamed

musab.iskandar@kaust.edu.sa, adel.alharbi.1@kaust.edu.sa, khalid.alharbi.4@kaust.edu.sa,
mohammed.omar@kaust.edu.sa, youssef.mohamed@kaust.edu.sa

Abstract

Abstract

Spoken Arabic dialect identification (ADI) underpins dialect-aware speech technology, but systems that score well in-domain often collapse when applied to recordings from unseen sources, because dialect labels are entangled with channel, speaker, and recording-condition cues. We address NADI-2026 Shared Task 2, which makes this failure mode explicit: models are trained on ADI20-micro and evaluated on a blind test set drawn from undisclosed sources. We benchmark five backbones under a common pooling-plus-linear head and select an Arabic ASR encoder, `Cohere transcribe-arabic`, on the basis of out-of-domain accuracy rather than in-domain validation accuracy. We then sweep the cross-corpus augmentation taxonomy of Dey et al. against the same out-of-domain criterion, train the final system on a mixture of ADI20-micro and Casablanca so that more than one recording domain is seen during training, and average late-training snapshots. The system reaches 91.54% in-domain accuracy on ADI20-micro, 73.10% on the fully held-out MADIS-5 corpus, and 53% on the official blind test set (Cavg 0.11), placing fourth on the shared-task leaderboard. The 38-point drop from in-domain accuracy to the blind test, of which 20 points remain even after moving to a fully held-out out-of-domain corpus, is the central result: in-domain accuracy is not merely an optimistic proxy for cross-domain ADI performance, it is close to uninformative, and even a genuinely held-out out-of-domain corpus overstates blind performance substantially.

Date: August 24, 2026
Code: https://github.com/musab-iskandar/Arabic_Spoken_Dialect_Identification
Project Page: https://musab-iskandar/Arabic_Spoken_Dialect_Identification.github.io

Contents

1 Introduction **2**

2 Related Work **3**

3 Methodology **3**

3.1 Data 3

3.2 System 4

4 Results **5**

5 Discussion **6**

6 Limitations and Future Work **8**

7 Conclusion **9**

References **10**

1 Introduction

Spoken Arabic exhibits substantial dialectal variation across the roughly twenty Arabic-speaking countries, yet most speech technologies are trained predominantly on Modern Standard Arabic (MSA) or a narrow subset of dialects, leaving them brittle when applied to naturally occurring dialectal speech. Automatic spoken Arabic dialect identification (ADI) is a key enabling component for dialect-aware downstream systems, including voice assistants, call-centre routing and analytics, media monitoring, and dialect-conditioned ASR or machine translation, each of which benefits from knowing a speaker's regional variety before or during processing. Beyond its applied value, ADI is also a useful testbed for studying acoustic domain generalization: dialect labels are frequently entangled with recording conditions, channel characteristics, and speaker identity, which can produce misleadingly high in-domain accuracy without the model having learned genuine dialectal signal.

We address NADI-2026 Shared Task 2 [1]: given a raw speech utterance, predict its country-level Arabic dialect from a fixed inventory of twenty classes (ADI20) [2]. The task is explicitly cross-domain: the blind test set is drawn from sources and recording conditions that differ from the primary training and development data. A system that fits the training domain's acoustic shortcuts such as channel noise, speaker set, and recording equipment rather than dialect-discriminative content will perform well in-domain and poorly on the blind test set. Performance is measured primarily by top-1 accuracy, with Cavg (a NIST LRE-style cost metric computed over the full 20-class output space) as a secondary metric that penalizes mis-classification severity rather than treating all errors as equal.

Three properties of the task shape what a system can and cannot do. First, the label space is closed and country-level: every utterance is assigned to exactly one of twenty countries, with no option to abstain, no Modern Standard Arabic class, and no mechanism for utterances that fall outside the inventory. Since Arabic dialects form geographic continua rather than discrete categories, neighbouring countries within the Levant, the Gulf, or the Maghreb may be acoustically near-indistinguishable at the utterance level, and the same speaker may shift towards MSA within a single recording. Some portion of the error rate is therefore irreducible under this label definition. Second, the blind test set covers an undisclosed subset of the twenty countries, but submissions must still score over the full twenty-class output space, so a system is exposed to false positives on classes that may not be present at all. Third, and most consequentially, the target domain is unknown at development time. Any out-of-domain corpus used for model selection is a guess about the nature of the shift rather than a measurement of it, and its ability to predict blind performance cannot be verified until after submission.

Our system follows from these constraints. We treat cross-domain behaviour, not in-domain accuracy, as the quantity to optimise, and we enforce this at every decision point: backbones are compared on out-of-domain accuracy, augmentations are swept against the same criterion with a different-seed control run to establish a run-to-run noise floor, and checkpoints are selected on out-of-domain performance rather than in-domain validation loss. We benchmark five backbones under a common pooling-plus-linear head, select an Arabic ASR encoder, and train it on a mixture of ADI20-micro and a second corpus, Casablanca, so that more than one recording domain is represented in training rather than only the organizer-provided source. We additionally audit each corpus for non-dialect shortcuts before training, testing whether signal-level properties such as SNR and silence ratio predict the dialect label above chance. The resulting system places 4 on the official leaderboard, but its more useful contribution is negative: we show that in-domain accuracy is anti-correlated with blind performance in this setting, and that even a held-out out-of-domain corpus overstates the blind score by roughly twenty points.

2 Related Work

Spoken Arabic dialect identification (ADI) has advanced primarily through the NADI shared task series. In NADI-2025 [3], the strongest submissions combined large self-supervised encoders with auxiliary data: the winning ELYADATA & LIA system [4] fine-tuned a Whisper-large-v3 [5] encoder in two stages, freezing its first sixteen layers and adding attentive pooling, to reach 79.8% accuracy on the ADI subtask. Its companion ADI-20 corpus [2] extends ADI-17 to twenty countries and ships open ECAPA-TDNN [6] and Whisper baselines in SpeechBrain. MarsadLab [7] instead adapted this same ECAPA-TDNN baseline directly, adding feature reweighting via a lightweight attention mechanism prior to the classification layer, discriminative learning rates, and progressive unfreezing; the resulting system placed third overall at 61.6% accuracy, only marginally above the 61.1% baseline, suggesting that architectural refinements on top of ECAPA-TDNN offer limited headroom compared to switching to a larger pretrained backbone. The ECAPA-TDNN, with its channel-attentive pooling and squeeze-and-excitation blocks, remains a standard ADI backbone, while Whisper offers robust weakly supervised representations but treats Arabic as a single language without distinguishing dialects. This line of work defines the state of the art, yet it is tuned for in-domain accuracy and leaves cross-source behaviour underexplored.

The central difficulty is domain shift: models exploit speaker and recording-condition cues that correlate with dialect in training but fail to transfer to unseen sources. Two complementary remedies are relevant here. Abdullah et al. [8] apply nearest-neighbour voice conversion (kNN-VC) to normalise speaker identity, mitigating speaker bias and improving cross-domain accuracy. The same technique proved decisive in the NADI-2025 Saarland-Groningen system [9], where voice-conversion augmentation raised blind cross-domain accuracy from 68.7% to 76.4%. On the training side, Dey et al. [10] pair domain diversification with a domain-generalization objective, finding that adversarial training via a gradient reversal layer best suppresses recording-condition information on unseen data, a mechanism first introduced for accented-speech recognition [11].

Prior work therefore either maximizes in-domain accuracy without explicit domain generalization, or applies these robustness techniques in isolation and largely outside the fine-grained, country-level Arabic setting. Building on the NADI-2025 pilot baseline, we combine strong pretrained backbones with speaker-normalising augmentation, evaluated explicitly under the blind, cross-source conditions where existing systems are weakest.

3 Methodology

3.1 Data

Primary training data. We train on `UBC-NLP/NADI_2026_ADI20_micro`, the organizer-released ADI20 training and validation split (67,015 train / 10,806 validation clips, 20 dialect classes). The corpus is balanced by duration rather than by clip count: each dialect contributes approximately 10 hours of speech to the training split. Clip counts per dialect therefore vary considerably (up to 2.4× between the largest and smallest class by row count), simply reflecting differences in average clip length across dialects rather than an underlying label imbalance.

Secondary out-of-domain corpora. To measure cross-domain generalization during development, we therefore source two additional corpora ourselves:

- **Casablanca** [12]: covers 8 of our 20 target countries, with only validation and test splits (no train split) and a 1.57× label imbalance across dialects. Its `seg_id` field encodes a `program_id`, which we use as a pseudo-source identifier: same-program clips likely share the same speaker and recording conditions, so we treat program membership as a leakage

boundary when constructing our own splits from it, rather than splitting at the clip level.

- **MADIS-5** (`badrex/MADIS5-spoken-arabic-dialects`): covers 5 macro-region dialects (Gulf, Levantine, Maghrebi, Egyptian, and MSA), each labelled with its recording domain (Radio, TED Talks, TV Dramas, Theatre).

Shortcut auditing. For every corpus, we test whether non-dialect signal properties (SNR, RMS, silence ratio, duration) predict the dialect label above chance, as a proxy for the domain-shortcut failure mode this task is built around. We run this audit per corpus before any model training, so that dataset-level leakage is ruled out, or flagged, independently of downstream model performance.

Data scope by experimental phase. We use three different data scopes across the project, and are explicit about which applies where. For backbone benchmarking and the augmentation sweep, we subsample 25% of the ADI20-micro training set to keep iteration fast across the many configurations tested. For our final submitted system, we train on 80% of ADI20-micro together with 80% of Casablanca, combining both corpora so the model sees more than one recording domain during training rather than only the organizer-provided source. Evaluation uses the remaining 20% of ADI20-micro, the remaining 20% of Casablanca, and the full MADIS-5 set; MADIS-5 is evaluation-only, both because it is small and because it covers only 5 macro-region dialects rather than all 20 target classes. We excluded the full `ArabicSpeech/ADI20` corpus from training due to our limited compute resources.

3.2 System

3.2.1 Tools and infrastructure

We use SpeechBrain and HuggingFace `transformers/datasets` for modelling, and DuckDB for large-scale HuggingFace parquet counting. Early exploratory runs used Kaggle's dual-T4 environment as a stopgap; later training runs move to Vast.ai (H100 / RTX PRO 6000 Blackwell instances), managed via `nohup` for disconnect resilience.

3.2.2 Model

Backbone selection. We benchmark five candidate models on identical ADI20-micro splits (25% subsample; Section 3.1) using a pooling+linear classification head, evaluated both zero-shot and after fine-tuning: **ECAPA-TDNN** (the challenge baseline), **Whisper-Small** (partial encoder fine-tune after a linear-probe warm-up), **Whisper-Small + LoRA** (parameter-efficient adaptation, to test whether it reduces overfitting to ADI20-specific recording conditions), **Elyadata ADI-whisper-ADI20** [4] (the NADI-2025 winning checkpoint, used as a pretrained reference point), and **Cohere transcribe-arabic** (an Arabic-specific ASR encoder). Model selection is based on out-of-domain Casablanca accuracy rather than in-domain validation accuracy, consistent with our checkpoint-selection policy below.

Data augmentation. We adopt the cross-corpus augmentation taxonomy of Dey et al. [10] (noise, speed perturbation, bandwidth limiting, reverb/environment, enhancement, re-encoding, and codec round-trip) and sweep each augmentation individually against the OOD Casablanca metric, with a same-configuration, different-seed control run first to establish a noise floor so that augmentation effects can be judged against run-to-run variance rather than assumed real. We additionally explore nearest-neighbour voice conversion (kNN-VC) [8] as a speaker-normalising augmentation, sampling target-speaker pools across all dialects rather than per-dialect to avoid a known collapse failure mode. We also test test-time augmentation (TTA) at inference.

Final training configuration. We train the selected backbone on the data split described in Section 3.1 (80% of ADI20-micro combined with 80% of Casablanca) using the augmentation combination selected from the sweep above, and apply stochastic weight averaging (SWA) across late-training snapshots as a final step.

3.2.3 Evaluation

Checkpoint selection and metrics. All checkpoint selection is based on out-of-domain Casablanca performance, never on in-domain ADI20-micro validation accuracy, to keep model comparisons honest about cross-domain generalization, the project's central risk. We report top-1 accuracy separately in-domain and out-of-domain (held-out Casablanca, MADIS-5), and Cavg computed over the full 20-class output space, matching the competition's required output format regardless of which countries appear in a given evaluation set. We additionally track the in-domain/out-of-domain accuracy gap as an explicit indicator of generalization failure.

4 Results

Table 1 Backbone comparison, top-1 accuracy (%), identical split and protocol. Casablanca is out-of-domain here, as no selection run trains on it. MADIS-5 is scored over five macro-regions, the others over twenty classes.

Backbone	Steps	ADI20	Casa	MADIS
Whisper-small	3.0k	32.87	36.21	18.96
ADI-whisper-ADI20	3.0k	83.70	83.74	48.02
`cohere-transcribe-ar`	3.0k	78.12	81.39	80.22
`cohere-transcribe-ar`	5.2k	90.52	93.99	75.99

Table 2 reports the performance of the final system described in Section 3.2. It reaches 91.54% in-domain accuracy on ADI20-micro with a Cavg_{20} of 0.046. On the Casablanca held-out split it reaches 92.97%, rising to 95.53% when scoring is restricted to the eight countries actually present in that corpus, which indicates that part of the residual error consists of false positives on the twelve absent classes rather than genuine dialect confusion. On MADIS-5 it reaches 73.10%. On the official blind test set it scores 53% accuracy with a Cavg of 0.11.

The evaluation sets degrade monotonically with distance from the training domain, and the ordering is informative. Because 80% of Casablanca is part of the training data, its held-out figures are in-corpus accuracy rather than evidence of cross-domain transfer, which is why they sit slightly above the in-domain figure rather than below it. MADIS-5 is never used for training and drops roughly 18 points. The blind test set, which is both unseen and drawn from undisclosed sources, drops a further 20 points to 53%. The gap between our best development estimate and the official score is therefore 38.5 points in accuracy, and Cavg rises from 0.046 in-domain to 0.11 on the blind set.

Table 2 Final system performance across evaluation sets, ordered by increasing distance from the training domain. Casablanca is held-out in-corpus, since 80% of the corpus is in the training data; Casa_8 restricts scoring to the eight countries present in Casablanca. The blind test score is the official NADI-2026 Subtask 2 result.

Evaluation set	Accuracy	Cavg
ADI20-micro (in-domain)	91.54	0.05
ADI20-micro, macro	92.38	—
Casablanca (held-out, in-corpus)	92.97	0.01
Casablanca, 8 classes	95.53	—
MADIS-5 (out-of-domain)	73.10	—
Blind test (official)	**0.53**	**0.11**

5 Discussion

Section 4 reported 91.54% in-domain accuracy on ADI20-micro, 92.97% on the held-out Casablanca split, 73.10% on MADIS-5, and 53% with a Cavg of 0.11 on the official blind test set. Taken individually these numbers are unremarkable; taken as an ordered sequence they are the main result of this project, and this section explains why they fall the way they do.

The ordering is a distance-to-training-domain ordering, not a difficulty ordering. The four evaluation sets degrade monotonically as they move away from the data the model was fitted on, and each step down has a different cause. Casablanca sits *above* the in-domain figure rather than below it, which is initially surprising until one recalls that 80% of the corpus entered the training mixture: its held-out split is program-disjoint but not corpus-disjoint, so what it measures is in-corpus generalization across programs, not transfer across domains. MADIS-5 is the only set in the table the model has never seen in any form, and it drops roughly 18 points. The blind test set drops a further 20. That final drop is the interesting one, because nothing in our development pipeline predicted it and nothing in our development pipeline could have: MADIS-5 is scored over five macro-regions rather than twenty classes, so the roughly 18 points is a *lower bound* on the penalty attributable to leaving the training domain and the remaining 20 an *upper bound* on what was invisible to us. Either way, the most conservative development estimate we had still overstated the blind score by roughly twenty points.

Selecting on out-of-domain accuracy was correct, and it was not sufficient. The backbone comparison in Table 1 shows the two selection criteria disagreeing outright. At matched steps, ADI-whisper-ADI20 – the NADI-2025 winning checkpoint – leads Cohere transcribe-arabic by 5.6 points on in-domain ADI20 accuracy while trailing it by 32 points on MADIS-5. Had we selected on in-domain validation accuracy, as is standard practice, we would have shipped the weaker cross-domain model and would have had no signal indicating the mistake. The same tension appears *within* a single backbone: training the Arabic encoder from 3.0k to 5.2k steps buys 12 points in-domain and costs 4 points out-of-domain. In-domain accuracy in this task is not merely an optimistic proxy for blind performance, it is anti-correlated with it over the range we explored, which is a stronger claim than we expected to be able to make. Yet the criterion that got the backbone right did not close the gap, and the reason is the second finding below.

Unexpected finding 1: a corpus stops being a selection signal the moment it enters training. Casablanca was genuinely out-of-domain during backbone selection and behaved exactly as a selection corpus should. We then mixed its training portion into the final run to buy accuracy, and continued to select checkpoints on it. The rule kept its name and silently changed its meaning: Casablanca accuracy now tracks in-corpus fit, so selecting on it reproduces precisely the in-domain bias the whole design was built to avoid. This is visible in the numbers. Across our eight Arabic-encoder development runs, MADIS-5 accuracy peaks at step 400 in six runs and by step 800 in the remaining two, reaching 79.9–86.0%, then falls by a mean of 4.78 points (max 8.3) by the step each run actually selected – while in-domain and Casablanca accuracy climb monotonically over the same interval. The best-generalizing checkpoint was produced in every run and discarded in every run. We estimate several points of blind accuracy were left on the table by this single bookkeeping error, and it is the clearest lesson of the project.

Unexpected finding 2: the closed label space is a measurable, separable error source. Restricting Casablanca scoring to the eight countries actually present raises accuracy from 92.97% to 95.53%, so 2.56% of clips are assigned to classes that cannot occur in that corpus. Tunisian is the dominant false attractor: absent from Casablanca, but a high-prior class in the training mixture. This is not dialect confusion in the usual sense – it is a prior mismatch between the training distribution and the evaluation inventory, and since the blind set also covers an undisclosed subset of the twenty classes, a comparable share of blind errors is likely of the same kind. It also bounds what class rebalancing can buy, and explains our choice of $\alpha = 0.5$ rather than full prior equalisation: the classes that full balancing would upweight are exactly the ones most likely to fire spuriously off-domain. Cavg moves in the opposite direction under the same restriction, rising rather than falling, because absent classes contribute only false-alarm terms to the 20-class average and deflate it – a reminder that accuracy and Cavg reward different behaviours on an inventory the system cannot see.

Unexpected finding 3: the augmentation pipeline never ran. Our configuration requested four waveform augmentations. A post-hoc audit found that the dispatch table was never constructed: every call raised a `NameError`, absorbed by a broad exception handler that returned the unmodified waveform. The submitted system therefore applied no waveform augmentation at all. SpecAugment and the per-batch duration cropping run on separate code paths and were unaffected, so the system is not unregularised, but every waveform-level result reported here is an un-augmented baseline. What makes this worth reporting rather than quietly fixing is that the handler did emit a one-line warning per augmentation, which went unread, and the failure was invisible in every metric we tracked – loss curves, in-domain accuracy, and out-of-domain accuracy all looked as expected, because a silently-identity augmentation degrades nothing, it merely fails to help. It also means the augmentation sweep results should be read as a null experiment, and that the headline numbers have whatever headroom a working augmentation path would have provided.

Where this leaves us relative to the benchmark. We placed fourth, and the top four systems fall within roughly four points of each other against a 38.5-point in-domain-to-blind gap. The between-system spread is an order of magnitude smaller than the domain-transfer penalty every system paid. Whatever separates the leaderboard entries, it is small next to the problem none of them solved.

6 Limitations and Future Work

We state the limitations of this system plainly, in rough order of how much we believe each one cost us, and pair each with a concrete next step.

L1. Checkpoint selection used a corpus that was in the training mixture. This is the most serious methodological flaw in the work. Casablanca was a valid out-of-domain selection signal during backbone comparison, but once its training portion entered the final mixture its accuracy became a measure of in-corpus fit while we continued to treat it as a transfer measure. Across eight development runs, MADIS-5 accuracy peaked at step 400–800 and then fell by a mean of 4.78 points by the step each run actually selected. *Next step:* designate one corpus that is never trained on – MADIS-5, or better, a newly sourced corpus reserved for this purpose – as the sole checkpoint-selection signal, and enforce the separation in code so that any dataset appearing in the training manifest is mechanically excluded from the selection manifest. This costs nothing and should be the first change made.

L2. The early out-of-domain peak is unexplained. We observe that out-of-domain accuracy peaks within the first few hundred steps and then decays while in-domain accuracy climbs, but we have not established whether that peak is a genuine optimum or an artefact of an under-trained encoder whose representation has not yet specialised to the training domain. *Next step:* run a small number of full-length trainings with dense evaluation on a never-trained corpus, and compare early-stopped checkpoints against late checkpoints on a second held-out corpus. If the peak is real, early stopping on an out-of-domain signal recovers several points at no compute cost; if it is an artefact, the finding rules out a tempting but wrong intervention.

L3. The augmentation path was never executed. An unnoticed `NameError` absorbed by a broad exception handler meant every requested waveform augmentation returned the unmodified signal, so all waveform-level results are an un-augmented baseline and the augmentation sweep is a null experiment; SpecAugment and duration cropping were unaffected. *Next step:* repair the dispatch table, remove the broad handler in favour of fail-fast behaviour, and add a unit test asserting that each augmentation changes the waveform. Then re-run the Dey et al. cross-corpus taxonomy sweep (noise, speed perturbation, bandwidth limiting, reverb, enhancement, re-encoding, codec round-trip) against a properly held-out criterion, with a same-configuration different-seed control run first to establish the run-to-run noise floor.

L4. Speaker identity is a shortcut the system does nothing to suppress. Dialect labels in all our corpora are entangled with speaker and recording chain, and our only deliberate robustness measure was per-batch duration cropping over $\{1, 2, 3, 5, 8, 12\}$ seconds. *Next step:* apply nearest-neighbour voice conversion as a speaker-normalising augmentation, sampling target-speaker pools across all dialects rather than per-dialect to avoid the known collapse failure mode. This raised cross-domain accuracy from 68.7% to 76.4% in the NADI-2025 Saarland-Groningen system and targets exactly the shortcut we leave unaddressed. A complementary step is a gradient-reversal domain-adversarial head over the pseudo-source identifier (Casablanca `program_id`, MADIS-5 recording domain), which suppresses recording-condition information directly rather than hoping augmentation washes it out.

L5. Single-run, single-submission evidence. Every backbone row is one truncated run and every conclusion rests on one blind score. We have no error bars on the backbone ranking, and it could reorder with longer training or a different seed; the 32-point MADIS-5 gap that drove backbone selection is large enough that we are confident in its direction, but the smaller effects are not

separable from run-to-run variance. *Next step:* repeat the backbone comparison with at least three seeds per configuration trained to convergence, and report means with spreads rather than point estimates.

L6. Compute budget constrained data scope. We excluded the full `ArabicSpeech/ADI20` release and trained on ADI20-micro only, for 6,000 steps on a single GPU at roughly four hours per run. *Next step:* scale to the full ADI20 corpus with the selection protocol fixed per L1, since the value of more data is contingent on selection being honest – more in-domain data with an in-domain selection rule would likely widen the in-domain-to-blind gap rather than close it.

L7. Evaluation-set heterogeneity. MADIS-5 is scored over five macro-regions via a fixed 20-to-5 lookup, so its accuracy is not comparable in difficulty to the 20-way rows beside it, and our decomposition of the 38.5-point gap yields a bound rather than a point estimate. We also cannot rule out that the backbone saw Casablanca during its own pretraining, since Casablanca appears in its release benchmarks, which would mean part of the transfer we attribute to the encoder is exposure. *Next step:* source a genuinely 20-way, genuinely unseen evaluation corpus covering the full country inventory, and where possible confirm pretraining data composition with the model provider.

Promising directions. Beyond repairing the above, three extensions seem worth pursuing. First, calibration and abstention-aware decision rules: the closed label space forces a prediction on every clip, but a system that knows which of its 2.56% absent-class false positives it is unsure about could shift probability mass toward plausible neighbours and improve Cavg even at fixed accuracy. Second, hierarchical or region-aware objectives that make the loss sensitive to the geographic continuum – confusing two neighbouring Levantine countries should not cost the same as confusing Morocco with Oman. Third, explicit shortcut auditing as a standing part of the pipeline rather than a one-off pre-training check: testing whether SNR, RMS, silence ratio, and duration predict the label above chance on *every* new corpus, before it is trusted as a training source or a selection signal.

7 Conclusion

We submitted an Arabic ASR encoder (Cohere transcribe-arabic) with mask-aware mean pooling and a 20-way linear head, trained on ADI20-micro mixed with Casablanca and selected at every decision point on out-of-domain rather than in-domain accuracy, placing fourth on NADI-2026 Subtask 2 with 53% accuracy and a Cavg of 0.11, against 91.54% in-domain on ADI20-micro and 73.10% on the never-trained MADIS-5 corpus. The selection criterion was the right instinct and it worked where we applied it consistently: it chose an encoder that in-domain validation accuracy would have rejected, and which beat the NADI-2025 winning checkpoint by 32 points on a held-out corpus while trailing it by 5.6 points in-domain. But it was not sufficient. Of the 38.5-point in-domain-to-blind gap, at most 18 points are explained by leaving the training domain, leaving roughly twenty points invisible to every development signal we had, and we trace part of that residual to a specific mechanism: Casablanca stopped being a valid selection signal the moment its training portion entered the mixture, yet we kept selecting on it, steering away from checkpoints whose out-of-domain accuracy peaked within the first few hundred steps and then decayed while in-domain accuracy climbed. The takeaway a reader should carry away is therefore not our score but the constraint behind it: in cross-domain spoken dialect identification, in-domain accuracy is close to uninformative about blind performance and in our runs actively anti-correlated with it, a genuinely held-out out-of-domain corpus still overstated our blind result

by roughly twenty points, and a corpus ceases to be a measurement of transfer the instant it becomes part of what the model is fitted to – however much accuracy mixing it in appears to buy.

References

[1] Peter Sullivan et al. "NADI-2026: The Second Multi-dialectal Arabic Speech Processing Shared Task". In: *Proceedings of the Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)*. Budapest, Hungary: Association for Computational Linguistics, 2026.

[2] Haroun Elleuch et al. "ADI-20: Arabic Dialect Identification Dataset and Models". In: *Interspeech 2025*. 2025, pp. 2775–2779.

[3] Bashar Talafha et al. "NADI 2025: The First Multidialectal Arabic Speech Processing Shared Task". In: *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 720–733.

[4] Haroun Elleuch et al. "ELYADATA & LIA at NADI 2025: ASR and ADI Subtasks". In: *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 762–766.

[5] Alec Radford et al. "Robust Speech Recognition via Large-Scale Weak Supervision". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023, pp. 28492–28518.

[6] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". In: *Interspeech 2020*. 2020, pp. 3830–3834.

[7] Md. Rafiul Biswas et al. "MarsadLab at NADI Shared Task: Arabic Dialect Identification and Speech Recognition Using ECAPA-TDNN and Whisper". In: *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 752–756.

[8] Badr M. Abdullah et al. "Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification". In: *Interspeech 2025*. 2025, pp. 2790–2794.

[9] Badr M. Abdullah et al. "Saarland-Groningen at NADI 2025 Shared Task: Effective Dialectal Arabic Speech Processing Under Data Constraints". In: *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*. Suzhou, China: Association for Computational Linguistics, 2025, pp. 745–751.

[10] Spandan Dey, Md Sahidullah, and Goutam Saha. "Cross-Corpora Spoken Language Identification with Domain Diversification and Generalization". In: *Computer Speech & Language* 81 (2023), p. 101489.

[11] Sining Sun et al. "Domain Adversarial Training for Accented Speech Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 4854–4858.

[12] Bashar Talafha et al. "Casablanca: Data and Models for Multidialectal Arabic Speech Recognition". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 21745–21758.