Screening Yes, Differential No: AFM Tear-Film Disease Classification at Three Granularities

Anonymous ACL submission

Abstract

Tear fluid is a non-invasive matrix for diagnosing both ocular and systemic disease, and Atomic Force Microscopy (AFM) resolves the dried-residue surface at the nanometre scale, producing morphological patterns that vary across pathologies. Machine-learning work on this modality is frequently undermined by data leakage, in which multiple scans from the same patient end up on both sides of the train/test boundary. We present a patient-stratified benchmark for disease classification from AFM tear-film images across three diagnostic granularities: binary screening, a three-class grouping, and the full five-class differential. With no patient overlap between training and evaluation, EfficientNet-V2-S reaches a test macro accuracy of 94.1% (macro-F1 0.894) on binary screening, 82.4% (0.664) on the three-class task, and 51.0% (0.542) on the five-class task. Performance therefore drops as granularity increases, and the classical, lightweight-CNN, and modern-CNN families all converge to a narrow band at five-class, evidence that, at this cohort size, sample count and morphological variance bound the task more tightly than architecture. The benchmark establishes a baseline for the modality and makes the case that patient-level evaluation is a prerequisite for valid reporting.

1 Introduction

Tear fluid is an attractive diagnostic matrix because it is accessible and carries proteins, lipids, and metabolites at concentrations useful for biomarker work. It can be collected without needles, which suits large-scale screening and repeated sampling for both ocular conditions such as dry eye and systemic disease such as diabetes or multiple sclerosis. Biochemical assays have long been the standard readout, but high-resolution imaging now offers a complementary view: the morphology of the dried residue itself.

Atomic Force Microscopy resolves this residue at the nanometre scale. As tears evaporate, their constituents arrange into dendrites and crystalline patches whose geometry reflects the underlying physiological state. Prior work has hypothesised that changes in these patterns track changes in composition and pre-analytical conditions (Glinská et al., 2019; Kondrakhova et al., 2025), but most of that work stops at descriptive analysis or small classification tasks and does not examine what modern deep-learning models actually recover from the images.

The difficulty in asking that question rigorously is evaluation integrity. Any single patient typically contributes several scans from one clinic session, so an image-level random split places correlated scans on both sides of the train/test boundary and rewards the model for recognising the patient rather than the disease. In an earlier iteration of this work, exactly that mistake produced an apparently strong 89.6% test accuracy under a leaky image-level split, a number that collapsed as soon as the split was redone at the patient level. We therefore commit, for every experiment reported here, to a patient-stratified protocol in which no patient identifier appears in more than one partition, following the general formulation of data leakage given by Kaufman et al. (2012) and the medical-imaging-specific guidance of Willemink et al. (2020).

This paper presents the first patient-level deep-learning benchmark for disease classification from AFM tear-film images. Three diagnostic granularities are trained and evaluated in parallel: binary screening for any pathology, a three-class grouping that isolates dry eye, and the full five-class differential. The binary and three-class tasks prove tractable; the five-class task is hard and exposes the limits of what 240 images from 42 patients can support.

The contributions of this work are:

- A patient-stratified benchmark for AFM tear-film classification with zero patient overlap between training and held-out sets, together with a quantified leakage gap (89.6% image-level accuracy under leakage $\rightarrow$ 51.0% test macro accuracy / 0.542 macro-F1 at the patient-stratified five-class task).

- A performance hierarchy across three levels of diagnostic granularity (binary, three-class, five-class) that identifies binary screening as the most clinically tractable framing at the current cohort size.

- A comparative analysis spanning handcrafted features, standard pretrained CNNs, and a regularised modern CNN, showing that the three families converge in a narrow band at five-class, evidence that the bottleneck is signal, not capacity.

2 Related Work

A research lineage originating from the Košice group has, over roughly a decade, built up a multimodal account of the tear residue. Early work established the "structural fingerprint" of systemic disease (diabetes in the first instance) by combining synchronous fluorescence spectroscopy (SFS), Fourier-transform infrared spectroscopy (FTIR), and AFM (Glinská et al., 2019). Later studies moved from description toward mechanism: AFM-IR traced the loss of dendritic patterns in diabetic tears to lysozyme damage and glycation (Kondrakhova et al., 2025). The same toolkit was extended to psychiatric and neurological conditions, positioning tear fluid as a candidate biomarker platform for major depressive disorder (Krajčíková et al., 2021) and multiple sclerosis (Tomečková et al., 2023).

Classical tear-film morphology predates the AFM work by decades. Mucus fern ing at the millimetre scale (crystalline fronds in dried tear films) has been used as a dry-eye indicator since Rolando's 1984 observation and is still graded clinically (Masmali et al., 2014; Traipe-Castro et al., 2014). AFM extends the same idea to the nanometre scale, trading visual grading for quantitative height maps.

In parallel, deep learning has been applied to tear-film analysis along a different axis: stability and break-up dynamics. Su et al. (2018) first applied CNNs to automated tear film break-up time (TBUT) estimation; follow-up work has covered classification of interference-colour images (Kikukawa et al., 2023) and temporal modelling of break-up patterns (Qian et al., 2025). Portable capture has enabled clinical translation, with the Smart Eye Camera pairing a phone-grade device with a DL backend for dry-eye diagnosis (Shimizu et al., 2023). Adjacent modalities, including OCT en-face lipid-layer maps, have seen similar treatment (Stegmann et al., 2023). More broadly, small-cohort transfer learning is the established template for medical image classification; Esteva's dermatology benchmark (Esteva et al., 2017) and the survey by Litjens et al. (2017) are the canonical references for that setting.

What is still missing is rigorous image-based benchmarking on AFM tear-film images themselves. Image-level metrics in this space are routinely inflated by patient-level dependencies that a standard random split does not break. Pre-analytical choices (capillary collection versus flushing, for example) are known to shift the proteomic and morphological signal materially (Krajčíková et al., 2022), so any model that accidentally learns acquisition artefacts rather than disease can achieve strong image-level numbers. Our contribution is to bridge the Košice fingerprinting lineage with modern deep learning while enforcing patient-level evaluation, so that the reported numbers are tied to diagnosis rather than to patient identity.

3 Data: Acquisition and Preparation

The dataset is a set of AFM scans of human tear-film residues, acquired at a partner ophthalmology clinic on a Bruker Dimension Icon NanoScope (scanner head type SG). The underlying signal is a nanometre-scale height map of the dried residue surface, which is qualitatively different from the two-dimensional intensity images produced by light microscopy. Each scan is stored both as a 24-bit BMP rendered by the acquisition software and as a binary AFM container with the raw 16-bit height samples and an ASCII metadata header of roughly 34 KB.

The task is a single-label classification over five clinically defined groups (healthy controls, multiple sclerosis, glaucoma, diabetes, and dry eye disease) for a total of 240 scans. The class distribution is strongly imbalanced: multiple sclerosis contributes 95 scans and dry eye only 13, a ratio that reflects the prevalence of each condition in the

referring clinic's population and that we preserve through to the final evaluation.

Scans were not collected under a machine-learning-oriented protocol: regions, magnifications, and line counts vary between sessions, and a portion of the corpus carries the instrument artefacts typical of contact-mode AFM: scan-line streaks, tip-contamination bands, local piezo drift. No scan was excluded on quality grounds; these variations are part of what a downstream model in this setting would actually encounter.

The only preprocessing applied before modelling is a lossless conversion of each BMP to PNG at its native resolution (approximately 704×575 px). We do not resample, normalise, or edit content at this stage. A curated subset of the AFM header fields (scan size, scan rate, Z-sensitivity, among others) is also extracted into plain text; it is intended as an auxiliary text modality alongside the image.

4 Evaluation Integrity: Patient-Level Splitting

The corpus contains 240 images but only 42 unique patients: many clinic sessions produced several scans of the same residue, so the scans are correlated at the patient level. A random image-level split breaks that correlation in the wrong place (scans from the same patient land in both partitions) and rewards the model for recognising which patient the scan belongs to. This is the classic leakage failure mode documented by Kaufman et al. (2012) and a known source of inflated metrics in medical imaging (Oakden-Rayner et al., 2020). In an earlier pass we observed precisely this failure mode: an 89.6% accuracy under an image-level split, a figure that collapsed as soon as the split was redone at the patient level (Figure 1).

We therefore commit to a patient-stratified split throughout. Patient identity is recovered from the filename stems by cutting on the vendor's scan-index separator; everything that reduces to the same stem is treated as one session of one patient. Splitting is then performed independently per class so that the 80/20 ratio holds within each diagnostic group, with the constraint that no patient code appears in more than one partition. The resulting split has 189 images from 33 patients in training and 51 images from 9 patients held out.

Because several classes contain only a handful of unique patients, a three-way split would either

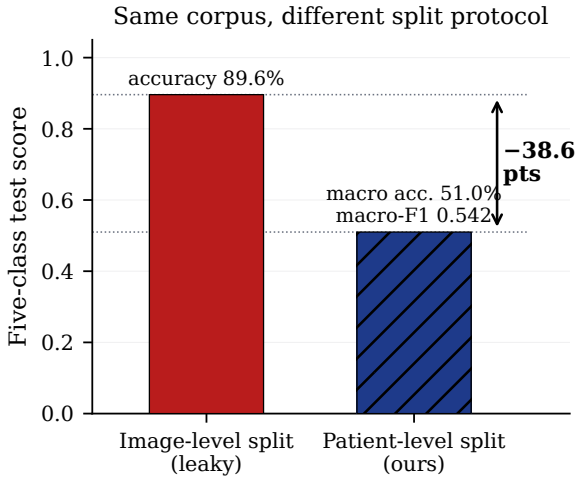

Figure 1: Five-class performance on the same underlying corpus under an image-level (leaky) split versus the patient-stratified split we commit to in the rest of the paper. The left bar reports plain accuracy as originally recorded; the right reports macro accuracy, because the patient-level held-out set is class-imbalanced. The -38.6 point drop is the central methodological observation of the paper.

empty the held-out side or force leakage, so validation and test are identical by construction and every held-out row is emitted under both labels. This makes the lack of an independent test cohort explicit rather than hidden, and keeps downstream code that expects both labels working unchanged. For the same reason we do not use k-fold cross-validation: with so few patients per class, fold-to-fold variation would be dominated by which patients happened to land in which fold rather than by model differences.

The splitting script verifies two invariants at the end of every run and aborts otherwise: no patient code appears in both training and held-out, and every held-out image appears exactly twice in the split CSV (once as val, once as test). All training, hyperparameter selection, and reported metrics below are produced against this single fixed split.

5 Methods

With 240 images from 33 training patients, the scarce resource is not capacity but supervision. A transformer trained from scratch would memorise the 33 patients and fail on the held-out 9; a hand-tuned feature pipeline with enough prior knowledge could plausibly match whatever a large network learns end-to-end. Our experimental design therefore spans three families of deliberately different inductive bias so that the reported ceiling reflects

the task, not the architecture. If the three families converge on the same held-out number, the signal itself is the bottleneck; if they separate, the gap tells us which inductive bias pays off at this sample size.

5.1 Diagnostic Task Framings

We train three framings side by side, each with its own checkpoint per architecture. The *binary* task separates healthy controls from any pathological sample and tests whether the modality carries a disease signal at all. The *three-class* task keeps healthy and dry eye separate and groups multiple sclerosis, glaucoma, and diabetes into a single *disease* label; dry eye is isolated because it is the condition with the most direct mechanical effect on tear-film composition and volume, so the framing asks whether its signature is separable from the remaining pathologies without yet demanding fine-grained discrimination among them. The *five-class* task is the full differential and the hardest of the three; it exposes the ceiling most clearly.

5.2 Model Families and Inductive Biases

We compare three families covering a wide span of inductive bias.

Classical handcrafted baselines. A deliberately low-capacity reference built to answer one question: how much of the task can be solved without any learned representation. For each cropped image we compute a 68-dimensional feature vector (grayscale moments and percentiles, a 32-bin intensity histogram, Sobel-gradient statistics, a Laplacian-variance focus proxy, per-channel RGB statistics, and 16 radial bins of the Fourier magnitude spectrum) and fit four `scikit-learn` (Pedregosa et al., 2011) classifiers in parallel (logistic regression, SVM with RBF, a 400-tree random forest, and a 300-estimator gradient-boosting ensemble), keeping the one with the best validation macro-F1. A small gap between this family and the deep models is evidence that the discriminative information at this sample size is largely low-dimensional.

Standard pretrained CNNs. ImageNet-pretrained (Deng et al., 2009) ResNet-18 (He et al., 2016), ResNet-34, and MobileNet-V3-Small (Howard et al., 2019), resized to 224×224 after the AFM crop. The only head change is replacing the final linear layer with one of width `num_classes`; training uses AdamW (Loshchilov and Hutter, 2019) at 3×10^{-4}, weight decay 10^{-4}, a cosine schedule (Loshchilov and Hutter, 2017) without warmup, plain cross-entropy (no label smoothing or class weighting), and no MixUp or CutMix. Fifteen epochs at batch size 32 are enough: best validation macro-F1 is reached inside the first few epochs on all three backbones. ResNet-18 (11 M parameters) and MobileNet-V3-Small (1.5 M) bracket a $7\times$ difference in capacity; if either meaningfully beats the other, capacity is a limiting factor. In practice their held-out numbers lie within noise of one another.

Alternatives considered and rejected. Larger EfficientNets (V2-M and B3) were trained under the same recipe as V2-S and did not improve held-out macro-F1 despite two- to three-times the parameter count, which we take as evidence that capacity is not the binding constraint at 240 images. Pure Vision Transformers (Dosovitskiy et al., 2021) were not trained: matching CNNs on small downstream tasks typically requires very large pretraining corpora or strong distillation, neither of which was within the scope of this study. A CLIP-style contrastive backbone (Radford et al., 2021) (ViT-L/14 features with logistic regression) was explored in an earlier iteration and reached roughly 57% five-class test accuracy, within the same band as the CNNs we report, but with no natural path to exploit the AFM metadata header. A vision-language direction combining Qwen2.5-VL-7B (Bai et al., 2025) with QLoRA (Dettmers et al., 2023) fine-tuning on image–header pairs is under way; its zero-shot accuracy is near chance, consistent with the absence of AFM priors in the pretraining mix, so we defer its reporting until the fine-tune stabilises.

Reference model: EfficientNet-V2-S. V2-S (Tan and Le, 2021) (21 M parameters, ImageNet1K-V1 weights, 384×384) is our reference. Among the candidates above, it reached the best validation macro-F1 on the five-class framing and fits comfortably on the single NVIDIA RTX 5090 (32 GB) used for all training runs in this paper. Its fine-tuning recipe differs from the simple-CNN family in six ways, each motivated by the data regime: (i) the backbone is frozen for the first two epochs so the freshly initialised head cannot push large gradients into the pretrained features; skipping this stage destabilises the first epoch of head updates on every run. (ii) The backbone learns at 5×10^{-5} and the head at $10\times$ that rate under a single AdamW optimiser. (iii) The schedule is cosine decay with

5% linear warmup. (iv) MixUp (Zhang et al., 2018) ($\alpha = 0.2$) and CutMix (Yun et al., 2019) ($\alpha = 1.0$) are sampled per batch as regularisers against memorisation of specific scans; on 189 training images they provide meaningful regularisation. (v) Cross-entropy is class-weighted by inverse frequency with mean 1 (Johnson and Khoshgoftaar, 2019) (a simpler alternative to focal loss (Lin et al., 2017) or effective-number weighting (Cui et al., 2019), which we did not find necessary at this corpus size), and label smoothing (Szegedy et al., 2016) is set to $\varepsilon = 0.1$; without class weighting the 7:1 imbalance pushes minority-class recall to zero. (vi) Stochastic depth (Huang et al., 2016) rates are multiplied by 1.5 and capped at 0.3, and head dropout is 0.4, both motivated by the tendency of V2-S to overfit within the first five epochs on this corpus.

5.3 Alternative Pipeline: Physics-Aware Tabular Features

Alongside the image pipeline above, we run a non-image tabular pipeline that feeds the tree ensembles reported in §7. Rather than the rendered BMP, it operates on the raw 4-channel AFM container (`height_sensor`, `amplitude_error`, `phase`, `height`) after a physics-aware preprocessing chain applied per channel: least-squares plane leveling on the two height channels to remove sample tilt, per-row median subtraction to correct line-by-line piezo hysteresis, clipping to the 1st–99th percentile against tip-crash spikes, and robust scaling with median/IQR. Phase and amplitude-error channels skip the plane-leveling step because they carry no tilt artefact.

Nine feature extractors are then run on each cleaned channel: 16 basic statistics (moments, percentiles, RMS, R_a, R_q, skewness, kurtosis), 4 spatial-gradient statistics, 4 GLCM texture descriptors (contrast, homogeneity, energy, entropy at 16 quantisation levels), 4 radial-band powers of the 2D FFT plus spectral entropy, 12 Local Binary Pattern histogram summaries across 3 radii, 16 Gabor responses at 2 frequencies and 4 orientations (mean and std), 7 Daubechies-4 wavelet energies across 2 decomposition levels, 1 box-counting fractal dimension on the median-thresholded binarisation, and 3 peak/valley topology counts. Cross-channel dependency is captured by 6 pairwise Pearson correlations between the 4 channels, and a further 15 bookkeeping features (7 quality-control flags and 8 channel-presence masks) are appended, giving $(16{+}4{+}4{+}4{+}12{+}16{+}7{+}1{+}3){\times}4{+}6{+}7{+}8 = 289$ features per scan. Scans larger than 1024 px on a side are stride-downsampled before extraction so the per-scan extractor budget stays bounded.

This pipeline is deliberately not a CV pipeline: no geometric or photometric augmentation applies, because the features are per-scan summary statistics rather than pixels; the patient-stratified split is the same one used for the image experiments; and the tabular classifiers (hist-gradient-boosting, random forest, XGBoost, LightGBM) are fit directly on the 289-dimensional vectors. The strongest single feature is the height-sensor kurtosis: multiple-sclerosis scans fall in the 10–28 range while other classes sit below 5, which we take as a sanity check that the physics-aware features are tracking a real disease-linked signal rather than acquisition artefact.

6 Augmentation and Training

With 240 images and a 7:1 class imbalance, augmentation and training recipe do more work here than architecture choice. This section first sets out the AFM-specific constraints that rule out a large fraction of the standard photometric toolbox, then describes the geometric and photometric blocks that survive those constraints and the leave-one-out ablation we run over them. We then compare offline and online sampling of the same pipeline, check whether the resulting ranking holds on a smaller backbone, and close with the shared optimisation protocol used for every deep model reported in the paper.

6.1 AFM-Specific Constraints

AFM scans are not ordinary photographs. Each pixel's colour is a deterministic function of a scalar height measurement, mapped through the vendor's lookup table, and two constraints follow directly. Any augmentation that breaks the monotone mapping between height and colour (hue shifts, channel swaps, saturation boosts, inversion, solarisation) also breaks the signal the model is supposed to learn, and is excluded outright. Second, the features that distinguish the classes are fine dendrite edges and thin-film granularity that occupy only a handful of pixels, so even ostensibly benign photometric operations have to be tightly bounded, or they erase the informative structure before the network sees it.

Within those bounds, we still want to expose the model to the variations that occur between real

AFM sessions: arbitrary in-plane rotation of the tip–sample coordinate frame, session-to-session differences in operator-chosen contrast and gain, scanline streaks from momentary tip–surface decoupling, mild focus drift, and the possibility that a future dataset is rendered under a different colormap. Before any augmentation and also at evaluation time, each image is cropped to the AFM data region (`left=93, top=0, right=616, bottom=531`), a 523×531 patch that removes the white side margins and the scale-bar strip that carries burned-in acquisition text. Without this crop the network learns to read the scale bar instead of looking at the scan.

6.2 Pipeline Layout and Leave-One-Out Ablation

The geometric block applies five operations in fixed order: a uniform $\pm 180^{\circ}$ rotation with reflective padding, independent horizontal and vertical flips, a small translation-only affine transform, a mild `RandomResizedCrop` to 384×384, and a low-amplitude elastic deformation. The photometric block, applied after geometry, contains seven operations: a colormap remapping under perceptually uniform palettes (viridis, plasma, inferno, magma, cividis, or grayscale) that preserves the height ordering, random brightness/contrast, random gamma, additive Gaussian noise, a custom scanline operator that injects 1–4 horizontal bands to mimic tip–surface decoupling, Gaussian blur, and CLAHE. Operations are implemented as an Albumentations (Buslaev et al., 2020) `Compose` whose members can be toggled through named configurations, which supports a leave-one-out ablation over the full pipeline (`full`), a preprocessing-crop-only baseline (`none`), geometric-only and photometric-only subsets, and seven single-operation removals.

6.3 Offline vs. Online Augmentation

We initially ran augmentation in an *offline* mode: ten augmented variants per training image were materialised to disk once, and the loader treated them as additional fixed samples. The effect was closer to a one-shot dataset expansion than to regularisation, since every epoch saw the same augmented copies. Re-running the same ablation grid in an *online* mode (augmentations resampled stochastically inside `__getitem__` on every fetch) improved macro-F1 on every single configuration at seed 3407, with a mean gap of $+0.097$. Under the online pipeline, the geometric-only subset becomes the best configuration (0.583), narrowly above the no-augmentation baseline (0.559) and well above the full pipeline with photometric operations included (0.534). The photometric block functions as a net regulariser only when paired with strong geometric priors; on its own it is the weakest configuration in the grid. For the main-text headline we retain the no-augmentation run, for continuity with the offline comparison and because it is the conservative choice; the online geometric-only pipeline is flagged as the strongest single-seed configuration, pending the seed-17 and seed-2026 confirmations that would be needed to state it as a multi-seed result.

6.4 Backbone Sensitivity

To check whether the geometric-only win is specific to V2-S or is a property of the task, we repeated a five-configuration subset on the smaller EfficientNet-B0 (5.3 M parameters, 224×224 input) at seed 3407. Two findings carry across both backbones: `online_geom_only` is the best configuration on both, and `none` beats `full` in both the offline and online pipelines on both models, so the photometric block is net-negative at this corpus size regardless of capacity. One finding does not carry. On B0, `online_none` and `offline_none` are numerically identical, which is consistent with B0 having too little capacity to overfit the frozen offline copies in the first place, the very failure mode that the online pipeline was designed to break. More importantly, the largest single improvement in the whole grid is not an augmentation effect at all but a capacity effect: the best B0 configuration (0.642) exceeds the best V2-S configuration (0.583) by $+0.059$ at a single seed (Figure 2). A multi-seed confirmation of this finding, together with an intermediate B1 comparison, is therefore the highest-leverage follow-up experiment.

6.5 Training Protocol

The deep models are trained with AdamW under a cosine schedule with linear warmup; given the 7:1 imbalance between the largest and smallest classes, the loss is class-weighted cross-entropy with inverse-frequency weights. Training runs on a single NVIDIA RTX 5090 (32 GB), and the checkpoint with the best validation macro-F1 is restored before test evaluation. All random seeds are fixed at 3407 throughout, matching the seed used for the patient split.

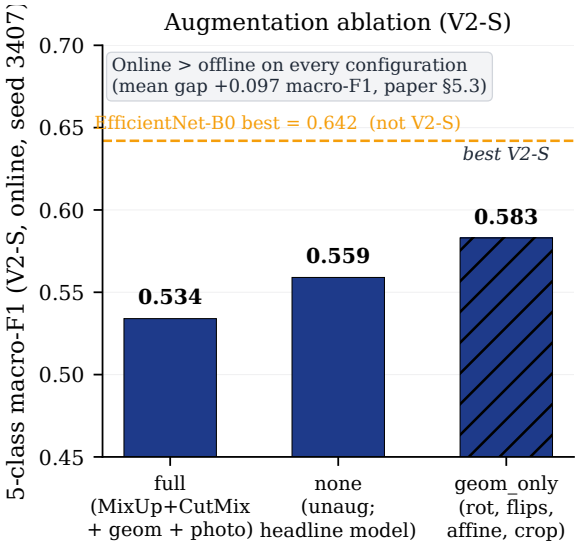

Figure 2: Five-class macro-F1 for three augmentation configurations on EfficientNet-V2-S under the online resampling pipeline at seed 3407. The dashed line marks the single-seed EfficientNet-B0 best (0.642, different backbone), which exceeds every V2-S configuration. Paper §5.3 reports online > offline on every configuration in the leave-one-out grid, with a mean macro-F1 gap of +0.097.

7 Results

Performance falls monotonically with granularity. Table 1 gives the headline numbers; Table 2 lists every configuration.

7.1 Binary Screening

EfficientNet-V2-S reaches 94.1% test macro accuracy and 0.894 macro-F1. The held-out split is imbalanced at 42 disease against 9 healthy, so we report macro-averaged metrics throughout rather than plain accuracy, which would be inflated by the majority class. The 0.09 macro-F1 gap over the classical baseline (0.804) is modest, but the deep model picks up structure that the 68-dimensional feature vector misses.

Two additional baseline families sit alongside the image classifiers on the binary task. Tabular classifiers fit on the 289-dimensional physics-aware feature vector extracted from the raw 4-channel AFM container (§5.3), with no image input at all, do most of the work: histogram gradient boosting reaches 0.912 / 0.874 (macro accuracy / macro-F1), approaching V2-S at 0.941 / 0.894, with random forest at 0.891 / 0.838, LightGBM at 0.881 / 0.821, and XGBoost at 0.866 / 0.832. Two alternative deep-learning supervision regimes, a hierarchical multiple-instance-learning model applied to patches (0.808 / 0.712) and a direct four-channel architecture (0.717 / 0.690), underperform the pretrained simple CNNs, with training-to-test macro-F1 gaps of 0.216 and 0.203 respectively. The practical reading is that physics-aware summary statistics over the raw AFM channels already explain most of the binary-screening signal, which suggests the rendered image is not strictly required for screening at this cohort size.

7.2 Three-Class Separability

Grouping multiple sclerosis (MS), glaucoma, and diabetes into a single disease label while keeping healthy and dry eye separate, V2-S reaches 82.4% test macro accuracy and 0.664 macro-F1. The drop from binary is real but not catastrophic, which is consistent with dry eye retaining a partially separable signature once the four-way pathology distinction is relaxed. The V2-S macro-F1 advantage over ResNet-18 is larger on three-class (0.664 vs 0.533) than on binary (0.894 vs 0.826), suggesting that the regularised recipe pays off once per-class support rises above the tiny-sample regime of `dry_eye` and `diabetes`; at a single seed this remains a hypothesis.

7.3 Five-Class Benchmark

Five-class differential is where the models stall. The best configuration is the unaugmented V2-S at 51.0% test macro accuracy and 0.542 macro-F1; the full pipeline with MixUp and CutMix scores 49.0% / 0.461, most likely because those regularisers erase the fine dendrite structure the minority classes depend on. The four deep configurations sit within about ten percentage points of one another on macro accuracy, and the classical baseline (macro-F1 0.455) is within roughly the same band, with no run clearing 60%. We read the convergence as a signal bottleneck, not a capacity one: at 240 images and 42 patients, within-class morphological variance is comparable to between-class variance for the four pathologies, and no backbone or recipe change in the range we tried moves the number meaningfully.

8 Discussion

The binary-to-five-class gradient is the paper's main observation. At 94.1% binary test macro accuracy (macro-F1 0.894), there is enough coarse structural signal in the residue to reliably flag disease, which is consistent with the Košice group's earlier finding that healthy dendritic patterns give way to

Framing	Setting	Test Macro Accuracy	Macro-F1
Binary (Screening)	EfficientNet-V2-S	0.941	0.894
Three-Class (Intermediate)	EfficientNet-V2-S	0.824	0.664
Five-Class (Diagnostic)	EfficientNet-V2-S (No-Aug)	0.510	0.542

Table 1: Headline performance across diagnostic granularities under patient-level evaluation. Plain accuracy is not reported because the held-out set is class-imbalanced; the test-macro-accuracy column reports the per-class average and macro-F1 is shown alongside.

disordered or fern-like ones under systemic stress (Glinská et al., 2019; Krajčíková et al., 2021). Five-class is the harder question: it asks the model to resolve branching angles and local crystal densities across four pathologies, and at 240 images it cannot separate those features from within-class variance. Part of that variance is biological, but part is pre-analytical: Krajčíková et al. (2022) showed that collection method alone materially shifts the proteomic and morphological profile, so a fine-grained classifier without a standardised protocol is partly reading how the tear was collected rather than what the patient has.

Recent AFM-IR work attributes the loss of dendritic structure in diabetic tears to lysozyme damage and glycation (Kondrakhova et al., 2025), which puts the visual signal downstream of biochemistry. A purely visual model can learn the pattern but not the cause. Clinical-grade differential diagnosis is therefore likely to need either substantially more patients, a fixed sampling protocol, or multimodal input (AFM-IR, FTIR, Raman, or the header digest we extracted but do not yet use). Patient-level evaluation is the precondition for any of those next steps: without it, improvements cannot be told apart from better patient memorisation.

9 Limitations

The cohort is small: 42 unique patients and 240 images, with only 33 patients in training and 9 held out. Patient-level stratification keeps the held-out estimate honest, but it does not make the estimate large, and the numbers reported here should be read as single-seed point estimates on a 51-image test set, where one misclassified image is worth roughly two percentage points of macro accuracy and 0.02 of macro-F1. Multi-seed confirmations and a patient-level majority-vote evaluation are queued but not included in the current reporting.

The class distribution is imbalanced at roughly 7:1 between the largest and smallest class. Class-weighted cross-entropy and macro-F1 reporting mitigate this but do not remove it: minority-class F1 remains the noisiest part of the grid, and the specific set of 9 held-out patients materially shapes the reported numbers.

Validation and test are identical by construction. We chose this over a three-way split because carving a third disjoint patient group at this corpus size would either leave evaluation partitions empty or force leakage, but the consequence is that hyperparameters selected on the validation fold are implicitly tuned on what we report as the test fold. The right fix is an external cohort, which we do not yet have.

The corpus comes from a single clinical site, a single AFM instrument, and a single operator workflow. External validity on other clinics or hardware is untested, which is the point that Kaushal et al. (2020) and Futoma et al. (2020) make about medical-ML cohorts more generally. We also lack the demographic metadata (age, sex, ethnicity) that would enable a subgroup fairness analysis. Calibration (ECE, reliability diagrams) is not reported either; in a screening setting the argmax is not sufficient, because downstream decisions depend on the threshold being meaningful.

10 Conclusion

We report a patient-level benchmark for disease classification from AFM images of dried tear-film residues across three diagnostic granularities. Under a strict patient-stratified protocol, EfficientNet-V2-S reaches 94.1% test macro accuracy (macro-F1 0.894) on binary screening, 82.4% (0.664) on the three-class task, and 51.0% (0.542) on the five-class differential, a monotone drop that exposes where the modality does and does not currently pay off. Three model families with very different inductive biases converge at the five-class level, which we read as evidence that signal, not capacity, is the binding constraint at 240 images from 42 patients.

The practical reading is narrow: AFM tear-film imaging is currently a plausible screening adjunct, not a standalone differential-diagnosis tool. The methodological reading is wider. The 89.6% image-level accuracy that dropped to 51.0% test macro accuracy / 0.542 macro-F1 at five-class once patient identities were respected is a concrete illustration of how much of this modality's apparent signal can be patient memorisation, and any future work in this space should be reported against a patient-stratified split by default. Standardising acquisition protocols, collecting external cohorts, and bringing

Framing	Family	Model	Augmentation	Test Macro Accuracy	Macro-F1
binary	Classical	LogReg	None	—	0.804*
binary	CNN (Simple)	ResNet-18	Standard	0.882	0.826
binary	CNN (Simple)	MobileNet-V3-Small	Standard	0.882	0.837
binary	CNN (Ref)	EfficientNet-V2-S	Standard	0.941	0.894
binary	Tabular (AFM)	HistGradientBoosting	—	0.912	0.874
binary	Tabular (AFM)	RandomForest	—	0.891	0.838
binary	Tabular (AFM)	LightGBM	—	0.881	0.821
binary	Tabular (AFM)	XGBoost	—	0.866	0.832
binary	DL (MIL)	HierarchicalMIL	—	0.808	0.712
binary	DL (4-channel)	Direct4Ch	—	0.717	0.690
three	Classical	SVM-RBF	None	—	0.491*
three	CNN (Simple)	ResNet-18	Standard	0.784	0.533
three	CNN (Ref)	EfficientNet-V2-S	Standard	0.824	0.664
multi (five-class)	Classical	SVM-RBF	None	—	0.455*
multi (five-class)	CNN (Simple)	MobileNet-V3-Small	Standard	0.569	0.467
multi (five-class)	CNN (Simple)	ResNet-18	Standard	0.588	0.451
multi (five-class)	CNN (Ref)	EfficientNet-V2-S	Full	0.490	0.461
multi (five-class)	CNN (Ref)	EfficientNet-V2-S	No-Aug	0.510	0.542

Table 2: Full inventory of classification experiments across the three framings and all model families. Plain accuracy is not reported because the held-out set is strongly class-imbalanced; test macro accuracy (per-class average) is shown alongside macro-F1. *Indicates validation macro-F1 used as proxy for test macro-F1 under the patient-stratified val ≡ test construction.

in the multimodal information (header metadata, FTIR, Raman) are the obvious next steps.

A TL;DR: Summary of the CV Pipeline

This appendix summarises the computer-vision branch of the work (everything that operates on the rendered AFM image). A parallel tabular branch over the raw 4-channel AFM container is reported separately.

Problem. First patient-level deep-learning benchmark for disease classification from rendered AFM tear-film images, motivated by the leakage-prone image-level random split that is standard in this modality.

Dataset. 240 AFM scans from 42 unique patients across five classes (healthy controls, multiple sclerosis, glaucoma, diabetes, dry eye disease), acquired on a single Bruker Dimension Icon NanoScope. 7:1 class imbalance. BMPs losslessly converted to PNG at native $\sim 704 \times 575$ px; no resampling, normalisation, or content editing at the data-prep stage.

Evaluation integrity. 80/20 patient-stratified split (189 training images / 33 patients vs 51 held-out / 9 patients) with hard invariants: no patient code in more than one partition, and every held-out image emitted twice (val and test). Val ≡ test by construction, because a three-way split at this cohort size either empties partitions or forces leakage.

Leakage quantification. The same underlying corpus yielded 89.6% accuracy under a prior image-level random split and 51.0% macro accuracy / 0.542 macro-F1 under the patient-stratified five-class split: a $\sim$38-point drop that is the central methodological finding.

Task framings. Three granularities trained in parallel, each with its own checkpoint per architecture: binary (healthy vs pathological), three-class (healthy, dry eye, grouped disease), five-class (full differential).

Preprocessing. Every image is cropped to the AFM data region (`left=93, top=0, right=616, bottom=531`) at train and eval time to remove white side margins and the scale-bar strip that carries burned-in acquisition text.

Model families.

- **Classical handcrafted.** 68-dimensional feature vector (grayscale moments and percentiles, 32-bin intensity histogram, Sobel gradient statistics, Laplacian-variance focus proxy, per-channel RGB statistics, 16 radial FFT bins) fed in parallel to LogReg, SVM-RBF, 400-tree random forest, and 300-

estimator gradient boosting; best-validation model kept.

- **Simple CNNs.** ImageNet-pretrained ResNet-18, ResNet-34, MobileNet-V3-Small at 224×224. AdamW at 3×10^{-4}, weight decay 10^{-4}, cosine schedule without warmup, plain cross-entropy (no label smoothing, no class weighting, no MixUp/CutMix), 15 epochs at batch size 32. ResNet-18 (11 M) vs MobileNet-V3-Small (1.5 M) brackets a $7\times$ capacity gap; held-out numbers lie within noise.

- **Reference CNN.** EfficientNet-V2-S (21 M params, ImageNet1K-V1 weights, 384×384). Six recipe changes motivated by the data regime: 2-epoch backbone freeze, $10\times$ discriminative head LR over backbone LR of 5×10^{-5}, cosine schedule with 5% linear warmup, MixUp (α=0.2) + CutMix (α=1.0) sampled per batch, inverse-frequency class-weighted CE with label smoothing ε=0.1, stochastic-depth rates multiplied by 1.5 and capped at 0.3 with head dropout 0.4.

- **Alternatives rejected.** Larger EfficientNets (V2-M, B3) did not improve held-out macro-F1 despite 2–3$\times$ the parameter count. Pure ViTs were not trained (small-corpus transfer gap). CLIP ViT-L/14 linear-probe reached $\sim$57% five-class test accuracy, in the same band as the CNNs. Qwen2.5-VL + QLoRA on image–header pairs is deferred until the fine-tune stabilises.

Augmentation study. AFM-specific constraint: each pixel's colour is a deterministic function of scalar height, so hue shifts, channel swaps, saturation boosts, inversion, and solarisation are excluded outright.

- **Geometric block (5 ops).** $\pm 180^\circ$ rotation with reflective padding, independent H/V flips, small translation-only affine, mild `RandomResizedCrop` to 384×384, low-amplitude elastic deformation.

- **Photometric block (7 ops).** Perceptual-colormap remap among viridis/plasma/inferno/magma/cividis/grayscale, brightness/contrast, gamma, additive Gaussian noise, custom scanline operator (1–4 horizontal bands), Gaussian blur, CLAHE.

- **Ablation.** Leave-one-out over `full`, `none`, geometric-only, photometric-only, and seven single-op removals, implemented as an Albumentations `Compose` with named toggles.

- **Offline vs online.** Online resampling inside `__getitem__` beats offline materialisation on every configuration (mean $+0.097$ macro-F1 at seed 3407).

- **Best configuration.** `online_geom_only` at 0.583 macro-F1, narrowly above `none` (0.559) and well above `full` (0.534). Photometric block is a net regulariser only when paired with strong geometric priors.

- **Backbone sensitivity.** On EfficientNet-B0 (5.3 M, 224×224), the `online_geom_only` win and the `none` $>$ `full` ordering hold, but B0 reaches 0.642 macro-F1, exceeding the best V2-S configuration by $+0.059$ at a single seed.

Training protocol. AdamW, cosine schedule with linear warmup, inverse-frequency class-weighted cross-entropy, single NVIDIA RTX 5090 (32 GB), all random seeds fixed at 3407 matching the patient-split seed, checkpoint with best validation macro-F1 restored before test evaluation.

Headline results (EfficientNet-V2-S on images). Binary: 94.1% test macro accuracy / 0.894 macro-F1. Three-class: 82.4% / 0.664. Five-class: 51.0% / 0.542. Performance drops monotonically with granularity.

Main finding. Classical, simple-CNN, and regularised-CNN families converge within $\sim$0.10 macro-F1 at five-class, which we read as evidence that the bottleneck is signal, not capacity, at 240 images / 42 patients.

CV-pipeline limitations. Single-site cohort, single AFM instrument, single operator workflow; 42 patients total; val $\equiv$ test by construction; single-seed reporting; no calibration metrics; no demographic metadata for subgroup analysis.

Next steps. Multi-seed confirmation at seeds 17 and 2026; patient-level majority voting; external cohort collection; standardised acquisition protocol. Complementary non-CV data branches (raw 4-channel AFM container, header metadata digest) are reported separately.

B TL;DR: Summary of the Raw-AFM Pipeline

This appendix covers the non-CV branch: two complementary approaches that operate on the raw 4-channel AFM container (`height`, `height_sensor`, `amplitude_error`, `phase`) rather than on the vendor-rendered BMP. Both approaches use the same patient-stratified split as the CV pipeline.

B.1 Tabular vs Deep Learning

Two fundamentally different approaches to the same classification task on the 240-scan, 42-patient corpus.

Tabular (tree-based ML).

- **Input.** $\sim$150 handcrafted features per scan (surface statistics, GLCM texture, FFT frequency bands, LBP, Gabor, wavelets, fractal dimension, cross-channel correlations); a superset of $\sim$289 features is described in §5.3.
- **Models.** XGBoost, HistGradientBoosting, RandomForest (ensemble of three).
- **Training.** StratifiedGroupKFold CV on `patient_id`, optional Optuna hyperparameter tuning.
- **Aggregation.** Patient-level majority voting or confidence-weighted mean.
- **Strengths.** Fast iteration, interpretable (SHAP), no GPU needed, works well on small datasets, captures domain-specific physics (roughness R_a/R_q, kurtosis, spectral entropy).
- **Weaknesses.** Features are hand-designed (may miss patterns), no spatial reasoning, cannot learn new representations.

Deep learning (CNN-based).

- **Input.** Raw 4-channel AFM images (`Height`, `HeightSensor`, `AmplitudeError`, `Phase`) at 160×160.
- **Models.** Three architectures — Direct4Ch (concatenated channels), MIL (gated attention over channels), Hybrid (CNN + tabular fusion).
- **Backbone.** EfficientNet-B0 (ImageNet pretrained), discriminative LR (backbone $0.1\times$ head).
- **Training.** AdamW, cosine LR with linear warmup, geometry-only augmentation (preserves AFM physics), 4-way flip TTA, label smoothing 0.1, early stopping on balanced accuracy.
- **Aggregation.** Patient-level mean probability.
- **Strengths.** Learns spatial patterns directly from the raw channels, can discover features invisible to global summary statistics, attention mechanisms provide channel-level interpretability.
- **Weaknesses.** Needs GPU, prone to overfitting at this corpus size, slower iteration, harder to debug.

Key differences. See Table 3.

Aspect	Tabular	Deep Learning
Feature source	Hand-engineered	Learned from pixels
Data efficiency	Better (small N)	Worse (needs more)
Compute	CPU, seconds	GPU, min–hr
Spatial awareness	None (global)	Full (conv filters)
Missing channels	QC flags + impute	Zero-fill or attention mask
Evaluation	GroupKFold CV (5-fold)	Fixed split
Patient grouping	StratifiedGroupKFold	Patient-stratified fixed
Best single feature	`height_sensor_kurtosis` (MS 10–28 vs others < 5)	Learned (not inspectable)

Table 3: Tabular vs deep-learning approaches on the raw 4-channel AFM container.

Why both? The small AFM corpus sits at the boundary where tabular ML is competitive with deep learning. Tabular captures known discriminative physics (kurtosis, phase statistics). DL may find spatial patterns invisible to global summary statistics. The hybrid fusion experiment explicitly combines both signal sources; a final submission likely ensembles across both approaches.

B.2 Deep-Learning Shared Configuration

All DL experiments on the raw container use: EfficientNet-B0 backbone, 160×160 images, AdamW optimizer, cosine LR with warmup, geometry-only augmentation, 4-way flip TTA, label smoothing 0.1, patient-level mean-probability aggregation, fixed 80/20 patient-stratified split (`SEED=3407`), training `SEED=42`.

B.3 Enabled Experiments (Binary, Healthy vs Diseased)

- **`direct4ch_binary_fixed`.** Direct 4-channel CNN. Concatenates the four AFM channels into a single 4-channel tensor and

feeds it through a modified EfficientNet-B0 (the 4th conv channel is initialised as the mean of the RGB weights). Simplest deep model — no attention, no handcrafted features. 256-dim projection, batch 16, 25 epochs, patience 6, dropout 0.2. Fastest to train; lowest-complexity baseline.

- **`hierarchical_mil_binary_fixed`.** Scan-level channel-attention MIL. Treats the 4 channels as separate "instances": each is processed independently through a shared backbone and pooled via gated attention, so the model learns per-channel importance weights. 128-dim embed, 64-dim attention, batch 8, 25 epochs, patience 6, dropout 0.3. Most interpretable (channel-attention weights are directly inspectable) and robust to missing channels (mask → zero attention weight).

- **`hybrid_fusion_binary_fixed`.** Image + tabular fusion. Two branches: (i) 4-channel CNN (same as Direct4Ch, 256-dim), and (ii) MLP on ∼150 handcrafted features producing a 64-dim embedding. The concatenated 320-dim vector is passed through a fusion head. Batch 8, 20 epochs, patience 5, dropout 0.3. Richest input signal but highest complexity and most prone to overfitting on 240 scans.

B.4 Disabled Experiments

Multi-class variants (3-class and 5-class) of the three architectures above are present in the codebase but disabled for the current reporting:

- **`*_3class_fixed`** — disabled as "multiclass not trustworthy yet" pending further calibration work.

- **`*_5class_fixed`** — disabled as underpowered given only 14 dry-eye and 26 diabetes samples.

Both sets can be re-enabled once the multi-class calibration issue is resolved and additional minority-class samples are available.

B.5 Codebase Extras

Tabular binary pipeline (`train_binary.py`). Handcrafted-features + tree-ensemble pipeline for the binary task. Extracts the ∼150-feature vector per scan and trains XGBoost, HistGradientBoosting, and RandomForest (optionally with Optuna tuning). StratifiedGroupKFold CV on `patient_id`; patient-level majority voting. No deep learning involved.

B.6 Architecture Comparison

Table 4 compares the four raw-AFM pipelines on input, backbone, aggregation, training budget, and interpretability.

B.7 Interactive 3D Visualisation App

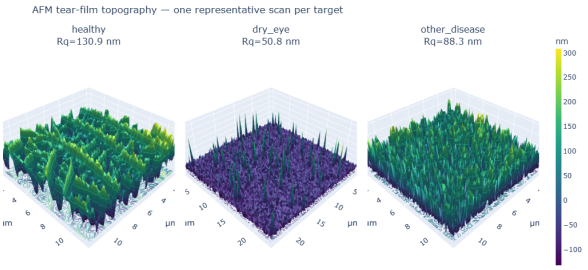

Figure 3: Screenshot from the 3D visualisation app: one representative AFM tear-film scan per target (healthy, dry eye, other disease), rendered from the `height` channel over the native μm grid with annotated R_q roughness. The app is an exploration aid, not part of the classification pipeline.

To support qualitative inspection of the raw AFM container we built a lightweight browser app that renders per-scan 3D topography from the `height` channel and exposes basic controls (target class, patient, scan index, channel). Each surface is plotted over its native μm grid with a viridis colormap and an annotated R_q roughness value (Figure 3), which lets a reviewer eyeball class-level morphology without loading the vendor software. The app is intended purely as an exploration aid — all classification results in this paper are produced by the pipelines described above, not by the visualiser — but it is useful for spot-checking acquisition artefacts, confirming that the patient-stratified split groups visually similar scans, and communicating the modality to collaborators outside the CV/ML team.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and

	Direct4Ch	MIL	Hybrid	Tabular
Input	4-channel tensor	4 separate images	4ch tensor + 150 feats.	150 features
Backbone	EfficientNet-B0	EfficientNet-B0 (shared)	EfficientNet-B0 + MLP	XGB / HGBR / RF
Aggregation	Direct concat	Gated attention	Late-fusion concat	—
Embed dim	256	128	$256 + 64 = 320$	—
Batch	16	8	8	full
Epochs	25	25	20	—
Interpretability	Low	High (channel weights)	Medium (feature imp.)	High (SHAP)
Missing-channel handling	Zero-fill	Attention mask	Zero-fill + QC flags	QC flags
Complexity	Low	Medium	High	Low

Table 4: Architecture comparison across the four raw-AFM pipelines (three deep-learning variants and the tabular ensemble).

Alexandr A. Kalinin. 2020. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492.

Gabriela Glinská, Kristína Krajčíková, Katarína Zakutanská, Oleg Shylenko, Daria Kondrakhova, Natália Tomašovičová, Vladimír Komanický, Jana Mašlanková, and Vladimíra Tomečková. 2019. Non-invasive diagnostic methods for diabetes mellitus from tear fluid. *RSC Advances*, 9(31):18050–18059.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. 2019. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. 2016. Deep networks with stochastic depth. In *European Conference on Computer Vision (ECCV)*, volume 9908 of *Lecture Notes in Computer Science*, pages 646–661.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4):15.

Amit Kaushal, Russ Altman, and Curt Langlotz. 2020. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA*, 324(12):1212–1213.

Yasushi Kikukawa, Shin Tanaka, Takuya Kosugi, and Stephen C. Pflugfelder. 2023. Non-invasive and objective tear film breakup detection on interference color images using convolutional neural networks. *PLOS ONE*, 18(3):e0282973.

Daria Kondrakhova, Michael Unger, Hans Stadler, Katarína Zakutanská, Natália Tomašovičová, Vladimíra Tomečková, Jozef Horák, Tatiana Kimáková, and Vladimír Komanický. 2025. Determination of diabetes mellitus disease markers in tear fluid by photothermal afm-ir analysis. *Nanomedicine: Nanotechnology, Biology and Medicine*, 61:102835.

Kristína Krajčíková, Gabriela Glinská, and Vladimíra Tomečková. 2022. Effect of tear fluid sampling and processing on total protein quantity and electrophoretic pattern. *Taiwan Journal of Ophthalmology*, 12(1):88–92.

Kristína Krajčíková, Erika Semančíková, Katarína Zakutanská, Daria Kondrakhova, Jana Mašlanková, Marek Stupák, Ivan Talian, Natália Tomašovičová, Tatiana Kimáková, Vladimír Komanický, and Vladimíra Tomečková. 2021. Tear fluid biomarkers in major depressive disorder: Potential of spectral methods in biomarker discovery. *Journal of Psychiatric Research*, 138:75–82.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Ali M. Masmali, Christine Purslow, and Paul J. Murphy. 2014. The tear ferning test: a simple clinical technique to evaluate the ocular tear film. *Clinical and Experimental Optometry*, 97(5):399–406.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pages 151–159.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Haochen Qian, Jingyao Chang, Yuling Yan, Zhihong Zhu, Jijia Zheng, Bolei Zhang, and Chunyan Xue. 2025. Deep learning for tear film stability assessment and breakup pattern classification in dry eye diagnosis. *Ophthalmic and Physiological Optics*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Eisuke Shimizu, Toshiki Ishikawa, Makoto Tanji, Naomichi Agata, Shintaro Nakayama, Yo Nakahara, Ryota Yokoiwa, Shinri Sato, Akiko Hanyuda, Yoko Ogawa, Masatoshi Hirayama, Kazuo Tsubota, Yasunori Sato, Jun Shimazaki, and Kazuno Negishi. 2023. Artificial intelligence to estimate the tear film breakup time and diagnose dry eye disease. *Scientific Reports*, 13:5822.

Hannes Stegmann, Valentin Aranha Dos Santos, Doreen Schmidl, Gerhard Garhöfer, Ali Fard, Homayoun Bagherinia, Leopold Schmetterer, and René M. Werkmeister. 2023. Classification of tear film lipid layer en face maps obtained using optical coherence tomography and their correlation with clinical parameters. *Cornea*, 42(4):490–497.

Teng-Yao Su, Zi-Yi Liu, and Ding-Yuan Chen. 2018. Tear film break-up time measurement using deep convolutional neural networks for screening dry eye disease. *IEEE Sensors Journal*, 18(16):6857–6862.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Mingxing Tan and Quoc V. Le. 2021. EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR.

Vladimíra Tomečková, Soňa Tkáčiková, Ivan Talian, Gabriela Fabriciová, Andrej Hovan, Daria Kondrakhova, Katarína Zakutanská, Miriama Skirková, Vladimír Komanický, and Natália Tomašovičová. 2023. Experimental analysis of tear fluid and its processing for the diagnosis of multiple sclerosis. *Sensors*, 23(11):5251.

Leonidas Traipe-Castro, Daniela Salinas-Toro, Daniela López, Marcelo Zanolli, Felipe Srur, Fernando Valenzuela, Alicia Cáceres, and Remigio López-Solís. 2014. Dynamics of tear fluid desiccation on a glass surface: a contribution to tear quality assessment. *Biological Research*, 47:25.

Martin J. Willemink, Wojciech A. Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, Daniel L. Rubin, and Matthew P. Lungren. 2020. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*.