دانشگاه فردوسی مشهد

گروه مهندسی کامپیوتر

گزارش پروژه درس

مبانی هوش محاسباتی

Blood Cell Classification with CNNs

دسته‌بندی تصاویر گلبول‌های سفید خون با شبکه‌های عصبی کانولوشنی

نام و نام خانوادگی	شماره دانشجویی
امیرحسین ابوالفضلی	۴۰۲۲۲۶۲۰۳۵
آرمان بیجاری	۴۰۲۱۲۶۲۱۳۱

استاد: **دکتر فضل ارثی**

نیمسال دوم ۱۴۰۵–۱۴۰۴

فهرست مطالب

۱. مقدمه

تشخیص نوع گلبول‌های سفید خون (White Blood Cells) یکی از مراحل مهم در آزمایشگاه‌های پزشکی است و به پزشکان کمک می‌کند تا بیماری‌هایی مانند عفونت، آلرژی و سرطان خون را شناسایی کنند. روش سنتی این کار توسط متخصص انسانی و زیر میکروسکوپ انجام می‌شود که وقت‌گیر و خطاپذیر است.

در این پروژه، از شبکه‌های عصبی کانولوشنی (CNN) برای دسته‌بندی خودکار تصاویر میکروسکوپی گلبول‌های سفید به چهار نوع اصلی استفاده می‌کنیم. دو رویکرد اصلی بررسی می‌شود: آموزش یک CNN از پایه، و استفاده از یادگیری انتقالی با مدل‌های از پیش آموزش‌دیده روی ImageNet.

۲. مجموعه داده

۱.۲. معرفی

مجموعه داده مورد استفاده Blood Cell Images نام دارد که توسط Paul Timothy Mooney در پلتفرم Kaggle منتشر شده و از تصاویر اصلی موسسه ملی بهداشت آمریکا (NIH) گرفته شده است. این مجموعه دارای مجوز CC0 (Public Domain) است. تصاویر به‌صورت JPEG رنگی با ابعاد 320×240 پیکسل ذخیره شده‌اند و در پوشه‌هایی با نام کلاس مربوطه سازماندهی شده‌اند (سازگار با flow_from_directory).

۲.۲. ساختار و توزیع داده

بخش	تعداد کل	تعداد به ازای هر کلاس	کاربرد
TRAIN	۹,۹۵۸	۲,۴۹۰ ≈	آموزش و اعتبارسنجی
TEST	۲,۴۸۸	۶۲۲ ≈	ارزیابی نهایی
TEST_SIMPLE	۷۲	متغیر	نمایش و دمو

جدول ۱: توزیع مجموعه داده در سه بخش

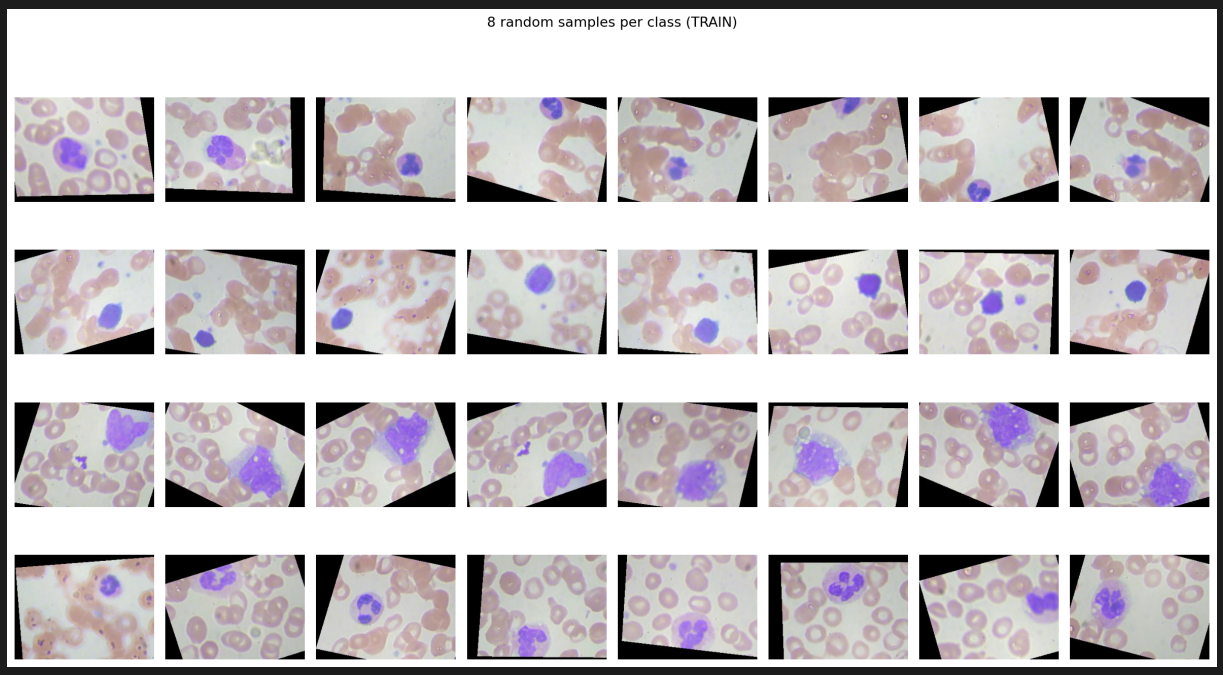
چهار کلاس هدف عبارتند از:

- **Eosinophil**: گلبول سفید با دانه‌های قرمز-نارنجی مشخص، نقش در واکنش‌های آلرژیک
- **Lymphocyte**: سلول گرد با هسته بزرگ تیره، مرتبط با سیستم ایمنی
- **Monocyte**: بزرگ‌ترین گلبول سفید با هسته کلیوی شکل
- **Neutrophil**: شایع‌ترین گلبول سفید با هسته چندلوبه بنفش

۳.۲. تحلیل اکتشافی داده (EDA)

۱.۳.۲. نمونه تصاویر از هر کلاس

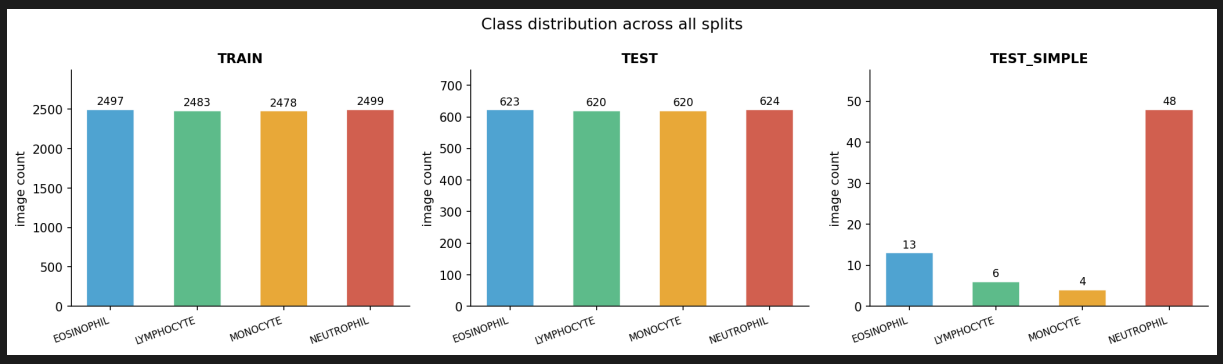
شکل ۱ نمونه‌هایی از هر چهار کلاس را نشان می‌دهد. تفاوت بصری بین کلاس‌ها قابل مشاهده است؛ اما تشخیص Monocyte از Lymphocyte برای مدل دشوارتر خواهد بود.



شکل ۱: ۸ نمونه تصادفی از هر کلاس در مجموعه آموزشی

۲.۳.۲. توزیع کلاس‌ها

همان‌طور که در شکل ۲ مشاهده می‌شود، توزیع کلاس‌ها در هر سه بخش تقریباً متوازن است. این ویژگی مهم است زیرا نیاز به وزن‌دهی کلاس یا روش‌های oversampling را برطرف می‌کند.



شکل ۲: تعداد تصاویر هر کلاس در سه بخش TRAIN، TEST و TEST_SIMPLE

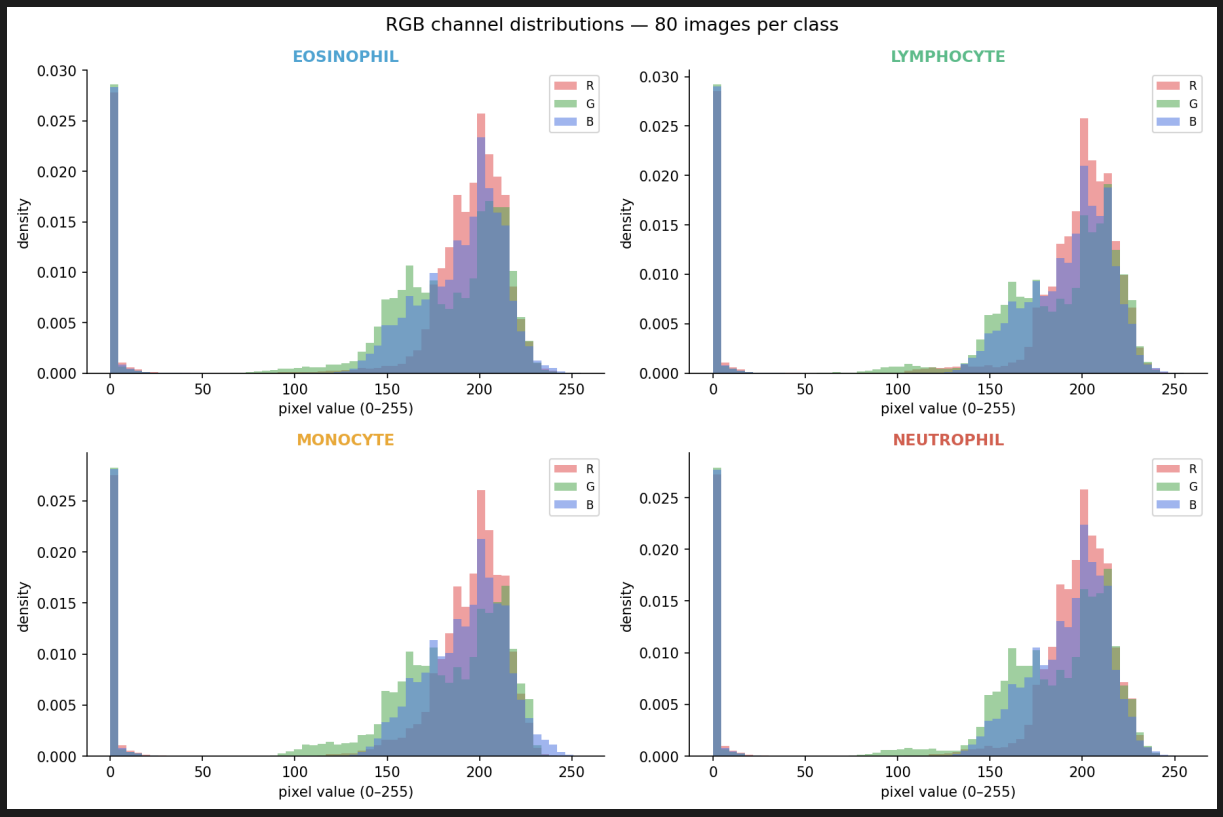
۳.۳.۲. ابعاد تصاویر

بررسی ۳۰۰ نمونه تصادفی از مجموعه آموزشی نشان داد که تمامی تصاویر دارای ابعاد یکسان 320×240 پیکسل و فرمت RGB هستند. در مرحله آموزش، این تصاویر به 224×224 پیکسل تغییر اندازه داده می‌شوند (برای EfficientNetB3 به 300×300).

۴.۳.۲. توزیع شدت پیکسل

شکل ۳ توزیع سه کانال رنگی R، G، B را برای هر کلاس نشان می‌دهد. تفاوت‌های قابل توجهی بین کلاس‌ها وجود دارد؛ به‌ویژه Eosinophil به دلیل رنگ‌آمیزی eosin غلبه قرمز-صورتی نشان می‌دهد، در حالی که Neutrophil

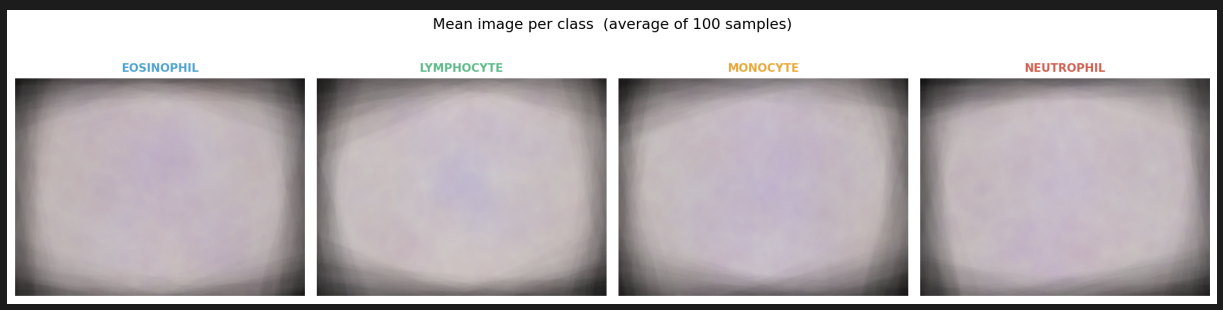
دارای توزیع آبی بیشتری است.



شکل ۳: توزیع شدت کانال‌های RGB برای ۸۰ نمونه از هر کلاس

۵.۳.۲. تصویر میانگین هر کلاس

شکل ۴ میانگین ۱۰۰ تصویر از هر کلاس را نشان می‌دهد. تصاویر میانگین تیز و قابل تشخیص هستند که نشان می‌دهد هر کلاس از نظر شکل ظاهری نسبتاً همگن است.



شکل ۴: تصویر میانگین هر کلاس (میانگین ۱۰۰ نمونه)

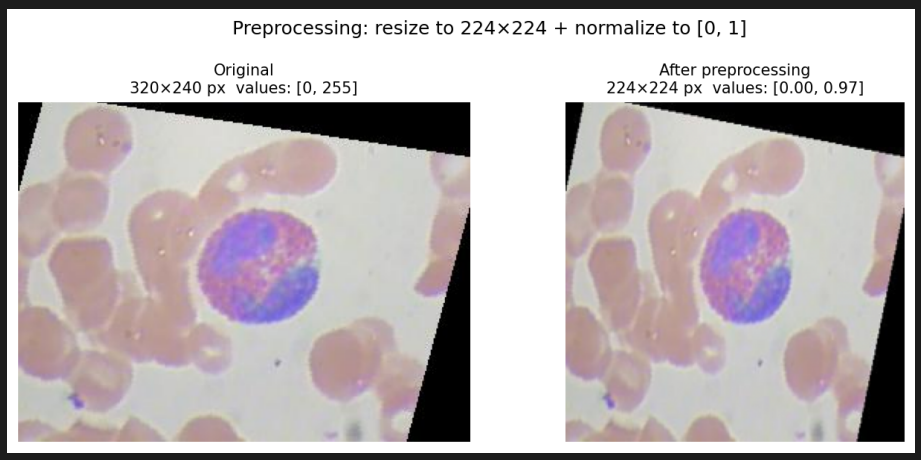
۳. پیش‌پردازش و افزایش داده

۱.۳. تغییر اندازه و نرمال‌سازی

همه تصاویر در مرحله بارگذاری به اندازه ثابت 224×224 پیکسل تغییر اندازه داده می‌شوند (برای EfficientNetB3 به 300×300). سپس مقادیر پیکسل از بازه $[0, 255]$ به $[0, 1]$ نرمال می‌شوند:

$$x_{\text{norm}} = \frac{x}{255}$$

برای مدل‌های از پیش آموزش‌دیده، به‌جای تقسیم ساده، از تابع preprocess_input مربوط به هر مدل استفاده می‌شود که نرمال‌سازی متناسب با وزن‌های ImageNet را اعمال می‌کند. شکل ۵ اثر این پیش‌پردازش را نشان می‌دهد.



شکل ۵: تصویر اصلی در مقابل تصویر پیش‌پردازش‌شده (تغییر اندازه به 224×224 و نرمال‌سازی)

۲.۳. تقسیم داده: آموزش و اعتبارسنجی

بخش TRAIN به نسبت **۸۰ درصد آموزش** و **۲۰ درصد اعتبارسنجی** تقسیم می‌شود. این تقسیم به‌صورت تصادفی و با seed=42 برای بازتولیدپذیری انجام می‌شود:

بخش	تعداد تصاویر	درصد	کاربرد
آموزش (train)	۷,۹۶۶ ≈	۸۰%	به‌روزرسانی وزن‌های مدل
اعتبارسنجی (validation)	۱,۹۹۲ ≈	۲۰%	انتخاب بهترین checkpoint
آزمون (test)	۲,۴۸۸	—	ارزیابی نهایی مستقل

جدول ۲: تقسیم داده‌ها برای آموزش، اعتبارسنجی و آزمون

۳.۳. ساخت Data Generator

برای هر سه بخش یک generator مناسب با ImageDataGenerator کرس ساخته می‌شود. generator آموزش شامل افزایش داده و shuffle است؛ در حالی که generatorهای اعتبارسنجی و آزمون فقط نرمال‌سازی دارند.

چرا train باید shuffle شود ولی validation و test نه؟

- train: اگر shuffle نشود، هر batch ممکن است فقط از یک کلاس باشد. این باعث می‌شود گرادیان‌های به‌روزرسانی وزن‌ها غیرقابل اطمینان شوند و مدل آموزش ضعیفی ببیند.
- validation و test: shuffle نمی‌شوند تا نتایج ارزیابی در هر اجرا یکسان و قابل مقایسه باشند، و بتوان پیش‌بینی‌ها را دقیقاً به تصاویر متناظر نسبت داد.

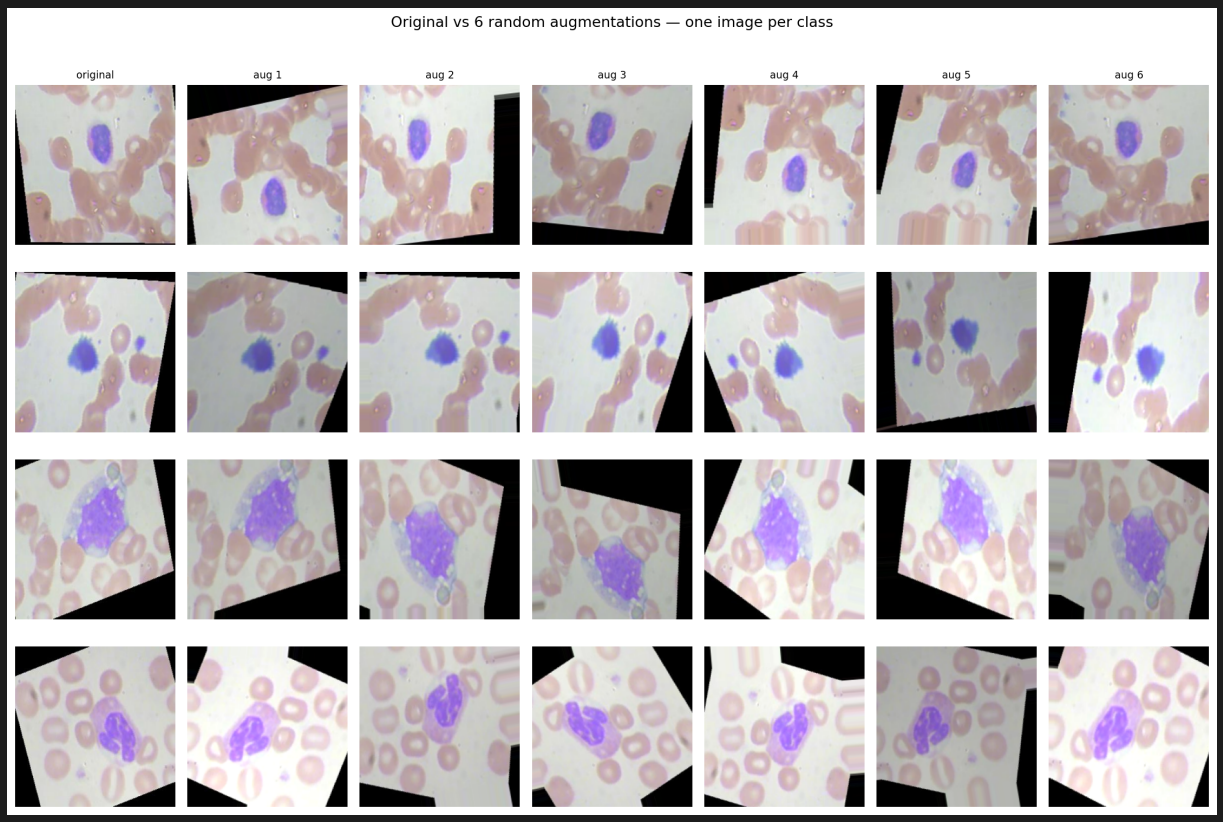
۴.۳. افزایش داده

افزایش داده (Data Augmentation) فقط روی مجموعه آموزشی اعمال می‌شود. توجیه هر تبدیل از ماهیت تصاویر میکروسکوپی گلبول‌های خون می‌آید:

تبدیل	مقدار	دلیل
چرخش	± 15	سلول‌ها در تمام جهات ظاهر می‌شوند
انعکاس افقی	فعال	تصویر آینه‌ای در میکروسکوپ معتبر است
انعکاس عمودی	فعال	جهت بالا/پایین در میکروسکوپ اهمیتی ندارد
جابجایی افقی/عمودی	$\pm 10\%$	سلول همیشه دقیقاً در مرکز نیست
زوم	$\pm 10\%$	تغییر بزرگنمایی میکروسکوپ
روشنایی	$\pm 20\%$	شدت رنگ‌آمیزی بین اسلایدها متفاوت است
حالت پر کردن	nearest	از ایجاد لبه‌های سیاه جلوگیری می‌کند

جدول ۳: تبدیل‌های افزایش داده و توجیه آن‌ها

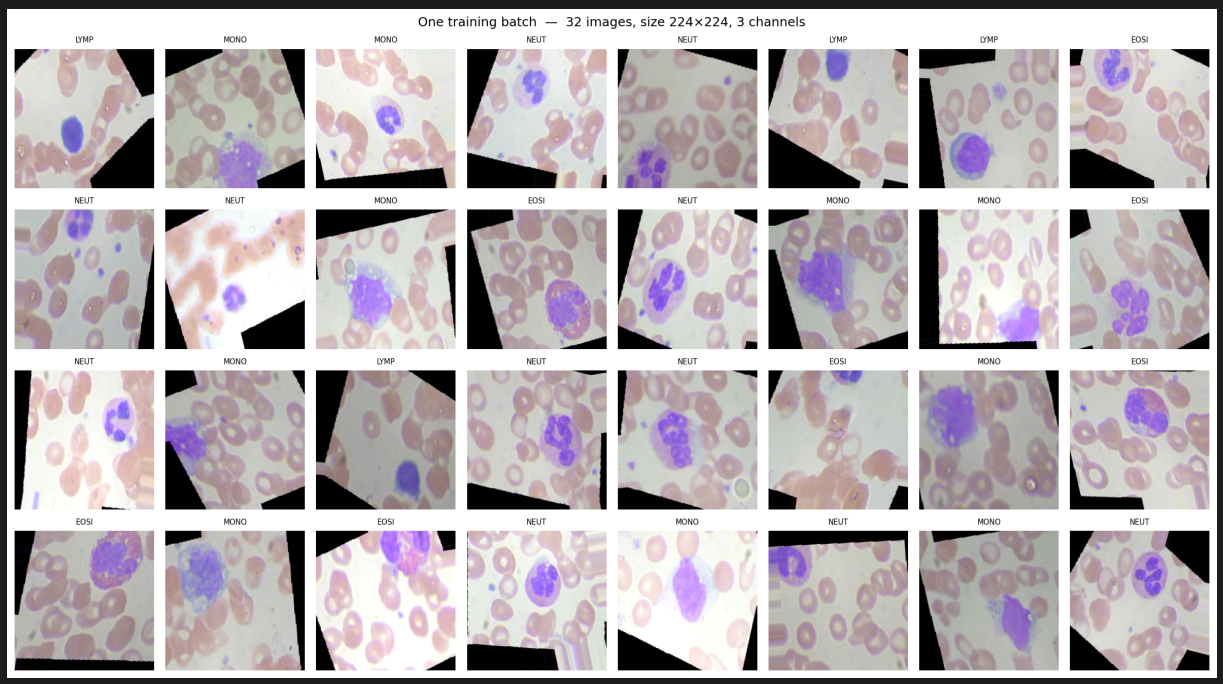
شکل ۶ نمونه‌ای از تصویر اصلی در کنار ۶ نسخه افزوده‌شده را برای هر کلاس نشان می‌دهد.



شکل ۶: تصویر اصلی در مقابل ۶ نسخه افزوده‌شده برای یک نمونه از هر کلاس

۵.۳. نمایش یک batch از داده‌های آموزشی

شکل ۷ یک batch نمونه از generator آموزشی را نشان می‌دهد. هر batch شامل ۳۲ تصویر به شکل $(32, 224, 224, 3)$ و برچسب‌های one-hot به شکل $(32, 4)$ است.



شکل ۷: یک batch از داده‌های آموزشی — ۳۲ تصویر 224×224 با برچسب کلاس

۴. مدل CNN از پایه

۱.۴ معماری

در این فاز یک شبکه‌ی CNN ساده از پایه طراحی و آموزش داده شد، بدون هیچ‌گونه وزن از پیش آموزش‌دیده، Dropout، BatchNorm یا افزایش داده. هدف اصلی داشتن یک خط پایه (baseline) برای مقایسه‌ی فازهای بعدی است.

معماری شبکه شامل چهار بلوک کانولوشنی است که هر کدام از یک لایه‌ی Conv2D با فعال‌سازی ReLU و یک MaxPooling تشکیل شده‌اند. در انتها دو لایه‌ی متصل کامل قرار دارد که خروجی نهایی چهار کلاس را با Softmax پیش‌بینی می‌کند:

$$\text{ورودی}\ (128 \times 128 \times 3) \xrightarrow{\text{Conv۳۲+Pool}} \xrightarrow{\text{Conv۶۴+Pool}} \xrightarrow{\text{Conv۱۲۸+Pool}} \xrightarrow{\text{Conv۱۲۸+Pool}} \xrightarrow{\text{Flatten}} \xrightarrow{\text{Dense۶۴}} \xrightarrow{\text{Dense۴}}$$

لایه	نوع	شکل خروجی	تعداد پارامتر
ورودی	Input	$128 \times 128 \times 3$	—
بلوک ۱	Conv2D(32)	$128 \times 128 \times 32$	۸۹۶
	MaxPool	$64 \times 64 \times 32$	—
بلوک ۲	Conv2D(64)	$64 \times 64 \times 64$	۱۸,۴۹۶
	MaxPool	$32 \times 32 \times 64$	—
بلوک ۳	Conv2D(128)	$32 \times 32 \times 128$	۷۳,۸۵۶
	MaxPool	$16 \times 16 \times 128$	—
بلوک ۴	Conv2D(128)	$16 \times 16 \times 128$	۱۴۷,۵۸۴
	MaxPool	$8 \times 8 \times 128$	—
Flatten	—	8192	—
Dense(64)	ReLU	64	۵۲۴,۳۵۲
Dense(4)	Softmax	4	۲۶۰
جمع پارامترها			**۷۶۵,۴۴۴**

جدول ۴: معماری Baseline CNN — کمتر از یک میلیون پارامتر

تعداد کل پارامترهای قابل آموزش ۷۶۵,۴۴۴ است که زیر حد یک میلیون پارامتر مجاز می‌ماند.

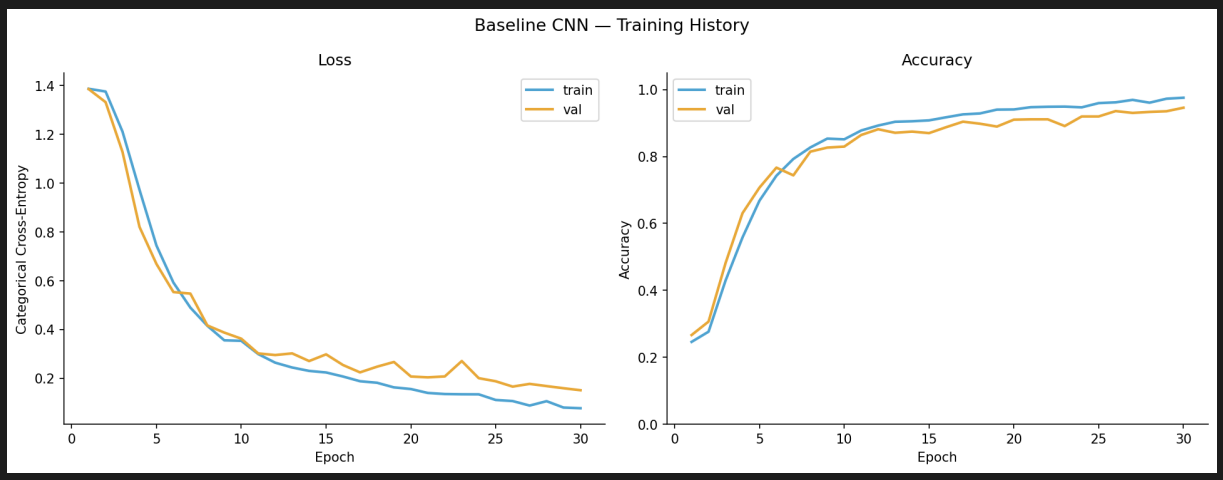
۲.۴. تنظیمات آموزش

تنظیم	مقدار
بهینه‌ساز	Adam
نرخ یادگیری	0.0001
تابع هزینه	Categorical Cross-Entropy
اندازه batch	32
تعداد epoch	30
اندازه تصویر	128×128 پیکسل
نرمال‌سازی	تقسیم بر 255
ذخیره بهترین مدل	ModelCheckpoint بر اساس val_accuracy

جدول ۵: تنظیمات آموزش Baseline CNN

۳.۴. نمودار آموزش

شکل ۸ روند تغییرات loss و دقت در طول ۳۰ epoch را برای داده‌های آموزشی و اعتبارسنجی نشان می‌دهد.



شکل ۸: نمودار loss و دقت در طول آموزش — Baseline CNN

هر دو منحنی آموزش و اعتبارسنجی با هم پیشرفت کردند و تا epoch آخر نیز val accuracy در حال بهتر شدن بود. در پایان آموزش، دقت آموزشی به ۹۷.۵۵% و دقت اعتبارسنجی به ۹۴.۵۷% رسید — اختلاف تنها ۲.۹۸% است که نشان می‌دهد overfitting شدیدی رخ نداده است.

۴.۴. ارزیابی روی مجموعه آزمون

بهترین مدل ذخیره‌شده (بر اساس val_accuracy) روی مجموعه‌ی آزمون مستقل ارزیابی شد:

معیار	مقدار
دقت (Accuracy)	۷۹.۵۳%
دقت ماکرو (Precision)	۸۱.۷۸%
فراخوانی ماکرو (Recall)	۷۹.۵۵%
F1-score ماکرو	۸۰.۱۹%
AUC ماکرو	۹۵.۵۲%

جدول ۶: نتایج ارزیابی Baseline CNN روی مجموعه آزمون

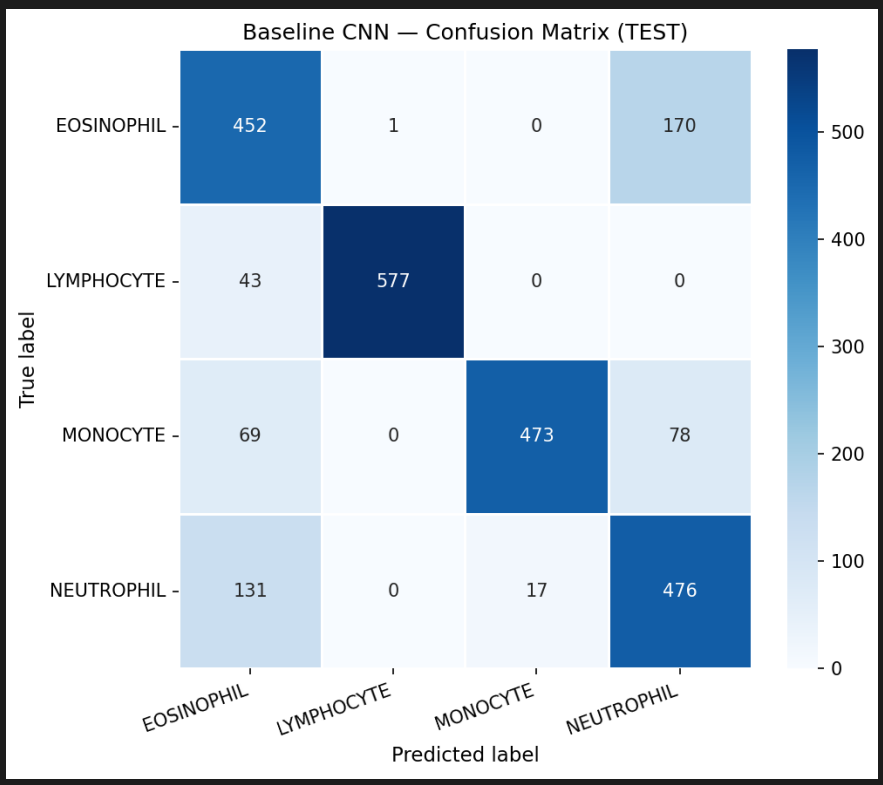
عملکرد مدل به تفکیک هر کلاس در جدول زیر آمده است:

کلاس	Precision	Recall	F1	AUC
Eosinophil	۰.۶۵	۰.۷۳	۰.۶۹	۰.۹۱۵۰
Lymphocyte	۱.۰۰	۰.۹۳	۰.۹۶	۰.۹۹۹۶
Monocyte	۰.۹۷	۰.۷۶	۰.۸۵	۰.۹۸۴۸
Neutrophil	۰.۶۶	۰.۷۶	۰.۷۱	۰.۹۲۱۵

جدول ۷: عملکرد به تفکیک کلاس — Baseline CNN

۵.۴. ماتریس درهم‌ریختگی

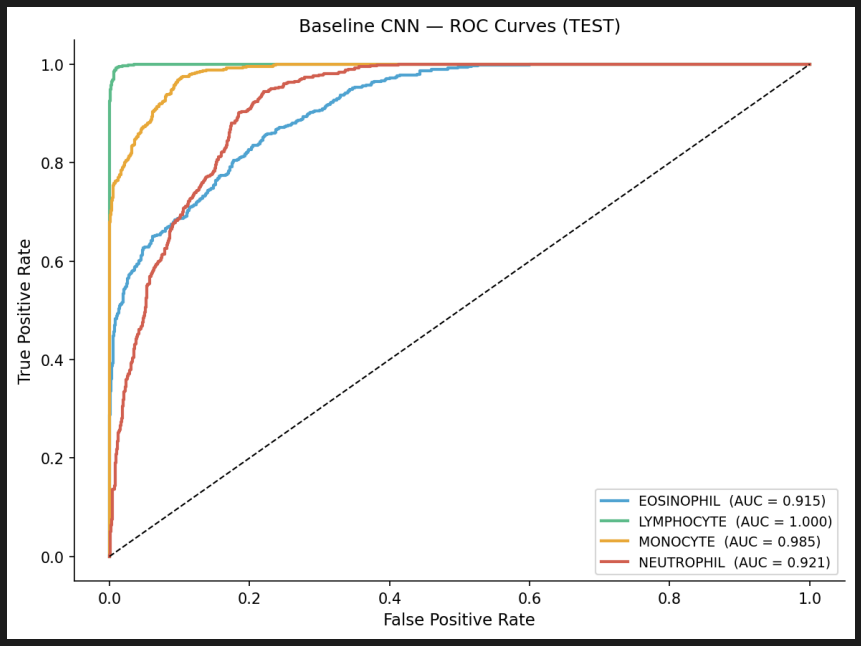
شکل ۹ ماتریس درهم‌ریختگی مدل روی مجموعه آزمون را نشان می‌دهد.



شکل ۹: ماتریس درهم‌ریختگی Baseline CNN روی مجموعه آزمون

۶.۴. منحنی ROC

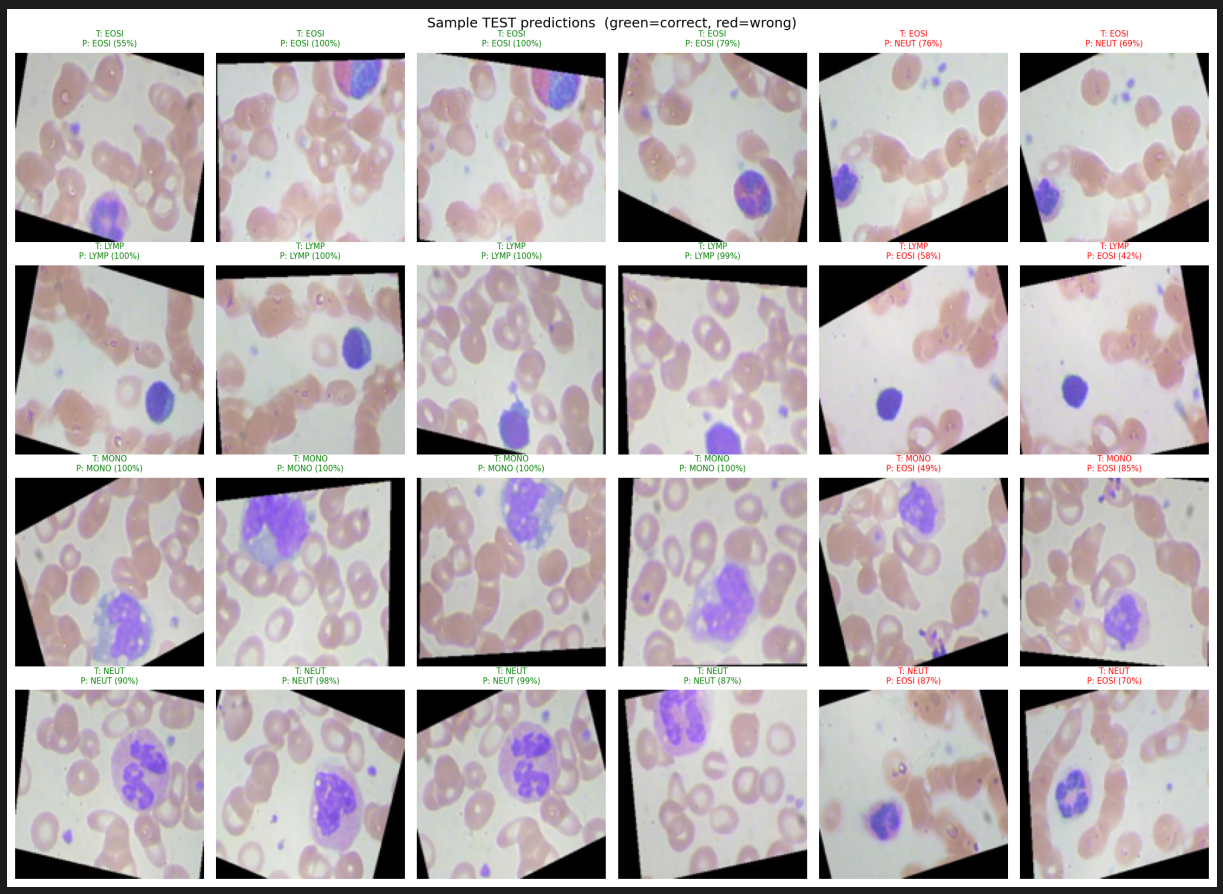
شکل ۱۰ منحنی‌های ROC به تفکیک هر کلاس را نشان می‌دهد. سطح زیر منحنی برای Lymphocyte تقریباً ۱.۰۰ است که نشان می‌دهد مدل این کلاس را با اطمینان بالا از بقیه جدا می‌کند.



شکل ۱۰: منحنی ROC برای چهار کلاس — Baseline CNN

۷.۴. نمونه پیش‌بینی‌ها

شکل ۱۱ چند نمونه از پیش‌بینی‌های مدل را نشان می‌دهد. عنوان سبز نشانه‌ی پیش‌بینی درست و عنوان قرمز نشانه‌ی اشتباه است.



شکل ۱۱: نمونه‌هایی از پیش‌بینی‌های Baseline CNN روی مجموعه آزمون

۸.۴. تحلیل نتایج

شکاف اعتبارسنجی–آزمون: اختلاف دقت اعتبارسنجی (۹۴.۵۷%) با دقت آزمون (۷۹.۵۳%) حدود ۱۵ درصد است. این شکاف ناشی از overfitting نیست چون منحنی‌های آموزش و اعتبارسنجی با هم پیش رفتند. دلیل احتمالی این است که ImageDataGenerator هنگام validation_split، آخرین ۲۰% فایل‌ها را بر اساس ترتیب الفبایی نام فایل انتخاب می‌کند نه به‌صورت تصادفی — بنابراین مجموعه اعتبارسنجی ممکن است به‌خوبی نماینده‌ی توزیع واقعی آزمون نباشد.

بهترین و بدترین کلاس‌ها: Lymphocyte با F1 برابر ۰.۹۶ و AUC برابر ۰.۹۹۹۶ بهترین عملکرد را داشت. این کلاس دارای هسته‌ای بزرگ و گرد و تیره است که از نظر بصری کاملاً متمایز است. در مقابل، Eosinophil و Neutrophil با F1 حدود ۰.۶۹ و ۰.۷۱ بدترین عملکرد را داشتند. هر دوی این‌ها گرانولوسیت هستند و هسته‌ای چندلوبه‌ای دارند؛ تنها تفاوت اصلی آن‌ها در نوع و رنگ دانه‌های داخل سلول است که در تصاویر با اندازه کوچک‌شده‌ی 128×128 بخشی از اطلاعات آن‌ها از دست می‌رود.

نتیجه‌گیری: این مدل پایه با ۷۶۵,۴۴۴ پارامتر و بدون هیچ تکنیک تنظیم‌سازی، به دقت آزمون ۷۹.۵۳% و AUC ماکرو ۹۵.۵۲% رسید. این اعداد مبنای مقایسه برای فازهای بعدی هستند که در آن‌ها از افزایش داده، BatchNorm، Dropout و یادگیری انتقالی استفاده خواهد شد.

۵. مدل CNN بهبودیافته

۱.۵. تغییرات نسبت به مدل پایه

در این فاز، معماری مدل پایه‌ی فاز دوم ثابت نگه داشته شد و سه دسته تغییر روی آن اعمال گردید:

۱. **افزایش داده (Data Augmentation):** پنج تبدیل تصادفی تنها در مرحله‌ی آموزش اعمال شد: چرخش افقی و عمودی (RandomFlip)، چرخش تصادفی تا ۱۵ درجه (RandomRotation)، زوم تصادفی تا ۱۰% (RandomZoom)، جابجایی تا ۱۰% در هر جهت (RandomTranslation)، و تغییر روشنایی ±۲۰% (RandomBrightness).

۲. **Batch Normalization:** لایه‌های BN پس از هر لایه‌ی کانولوشن و پیش از تابع فعال‌سازی درج شدند تا آمار ورودی را تنظیم کرده و همگرایی پایدارتری فراهم شود.

۳. **Dropout:** لایه‌ی Dropout با نرخ‌های مختلف پس از لایه‌های متراکم افزوده شد تا از بیش‌برازش جلوگیری شود.

معماری پایه شامل ۳ بلوک کانولوشن با 32، 64، و 128 فیلتر، به همراه یک لایه‌ی متراکم 256 نورونی، بدون تغییر باقی ماند.

۲.۵. تنظیم فراپارامترها (Hyperparameter Tuning)

پنج آزمایش طراحی و اجرا شد. در تمامی آزمایش‌ها از بهینه‌ساز Adam با نرخ یادگیری 1e-4 و EarlyStopping بر اساس val_loss (با صبر ۱۰ دوره) استفاده شد. جدول ۸ تنظیمات و نتایج بهترین دوره‌ی هر آزمایش را نشان می‌دهد.

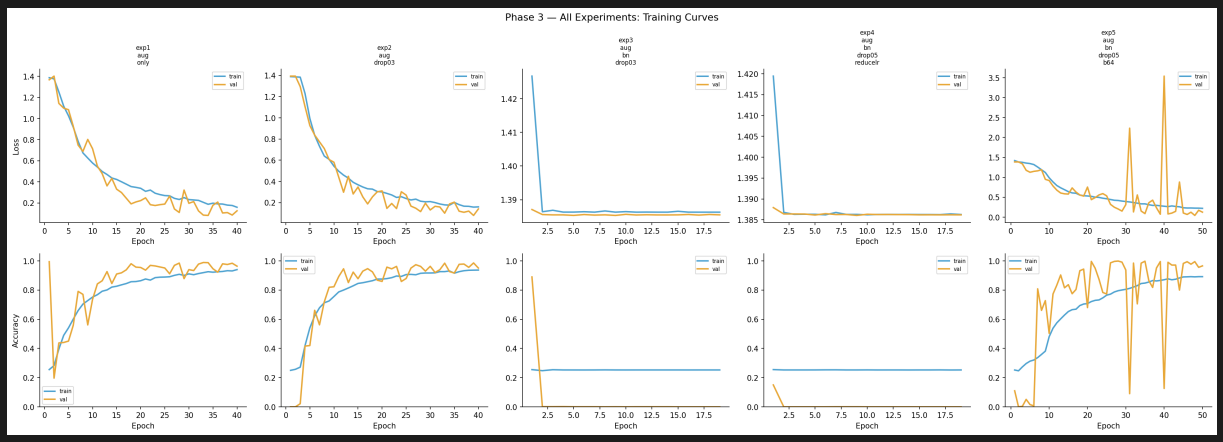
جدول ۸: آزمایش‌های فاز سوم و نتایج بهترین دوره

آزمایش	BN	Dropout	Batch	ReduceLR	دوره	Val Loss	دقت اعتبارسنجی
exp1 (تنها افزایش داده)	خیر	0.0	32	خیر	34	0.0816	%۹۸.۷۹
exp2 (افزایش داده + Dropout 0.3)	خیر	0.3	32	خیر	39	0.0788	%۹۸.۵۹
exp3 (Dropout 0.3 + BN)	بله	0.3	32	خیر	19	1.3853	%۰.۰۵
exp (ReduceLR + Dropout 0.5 + BN)	بله	0.5	32	بله	19	1.3860	%۰.۰۵
exp5 (batch=64 + Dropout 0.5 + BN)	بله	0.5	64	بله	48	0.0428	%۹۹.۵۰

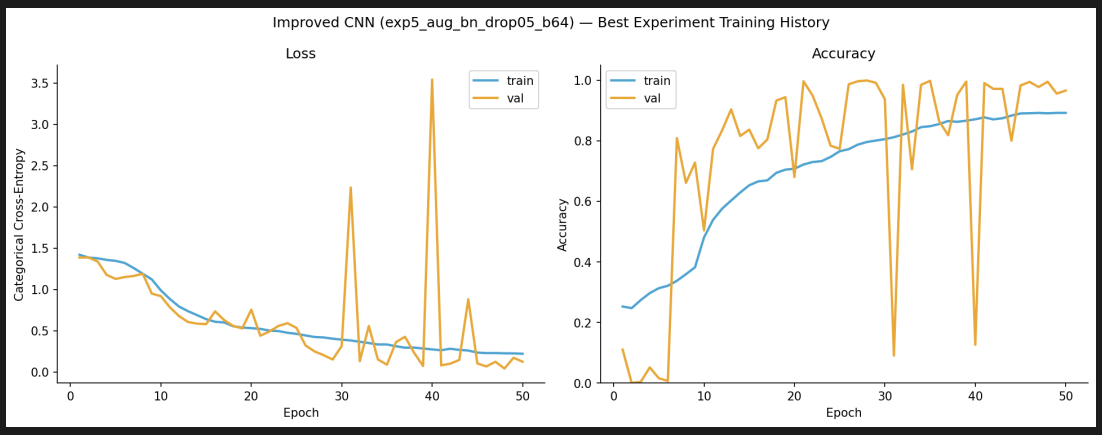
ناهنجاری exp3 و exp4 — شبکه‌ی مرده (Dead Network): آزمایش‌های سوم و چهارم از ابتدا در حالت «شبکه‌ی مرده» گیر کردند: خطا ثابت روی 1.386 ≈ ln(4) (معادل حدس تصادفی برای ۴ کلاس) ماند و دقت اعتبارسنجی صفر بود. علت اصلی ناسازگاری BatchNorm با اندازه‌ی دسته‌ی کوچک (batch=32) در TF 2.21 است: آمارهای میانگین و واریانس که BN هر دسته محاسبه می‌کند با 32 نمونه بسیار پرنویز هستند و شبکه از همان ابتدا به حالت تباهیده‌ای می‌افتد که از آن خارج نمی‌شود. افزایش اندازه‌ی دسته به 64 در exp5 این مشکل را برطرف کرد.

۳.۵. نمودارهای آموزش

شکل ۱۲ منحنی‌های خطا و دقت همه‌ی آزمایش‌ها را با هم نشان می‌دهد. خطوط تخت exp3 و exp4 مشکل شبکه‌ی مرده را به‌وضوح آشکار می‌کنند. شکل ۱۳ منحنی‌های بهترین مدل (exp5) را به‌تنهایی نمایش می‌دهد.



شکل ۱۲: منحنی‌های آموزش و اعتبارسنجی همه‌ی آزمایش‌های فاز سوم



شکل ۱۳: منحنی‌های آموزش بهترین مدل (exp5) — کاهش یکنواخت val_loss تا دوره‌ی 48

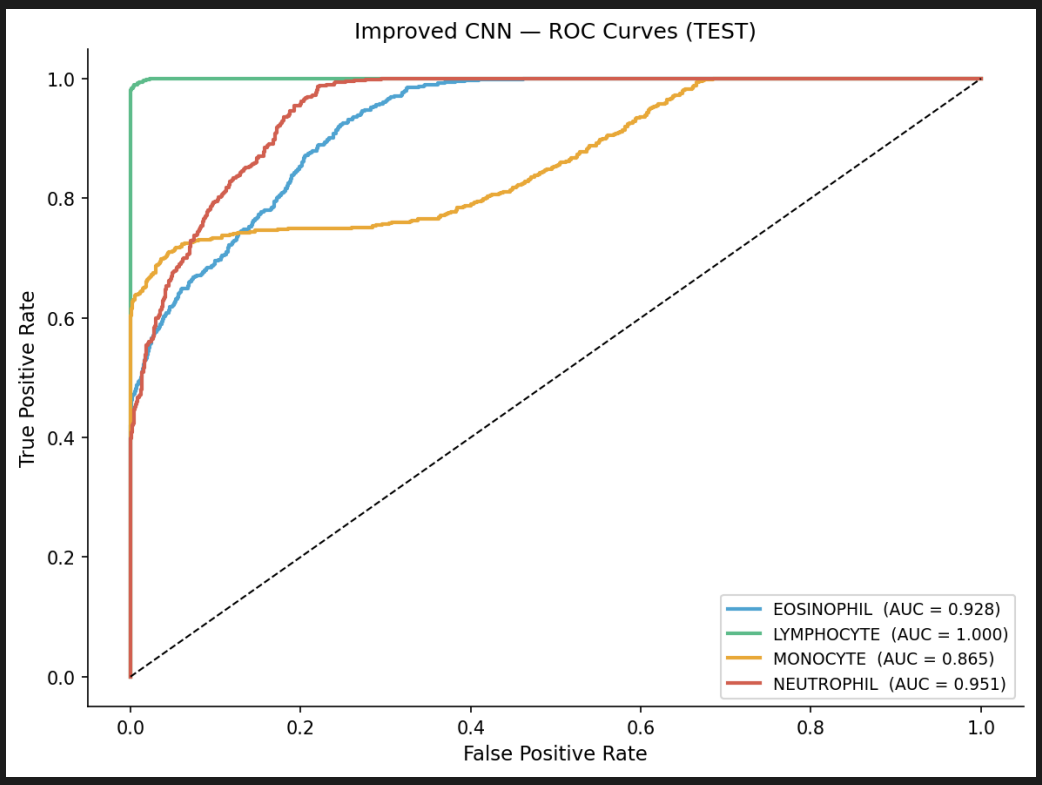
۴.۵. نتایج کامل مدل نهایی روی مجموعه‌ی آزمون

بهترین مدل (exp5_aug_bn_drop05_b64) روی مجموعه‌ی آزمون مستقل (2487 تصویر) ارزیابی شد. جدول ۹ نتایج کامل را نشان می‌دهد.

جدول ۹: نتایج کامل مدل نهایی فاز سوم (exp5) روی مجموعه‌ی آزمون

کلاس	دقت (Precision)	بازخوانی (Recall)	F1	AUC
EOSINOPHIL	0.74	0.66	0.69	0.9277
LYMPHOCYTE	1.00	0.55	0.71	0.9998
MONOCYTE	0.79	0.72	0.76	0.8650
NEUTROPHIL	0.60	0.97	0.74	0.9509
میانگین ماکرو	0.78	0.73	0.73	0.9359

دقت کل روی مجموعه‌ی آزمون: ۷۲.۳۸%، خطای آزمون: 1.4701.



شکل ۱۴: منحنی‌های ROC مدل نهایی فاز سوم

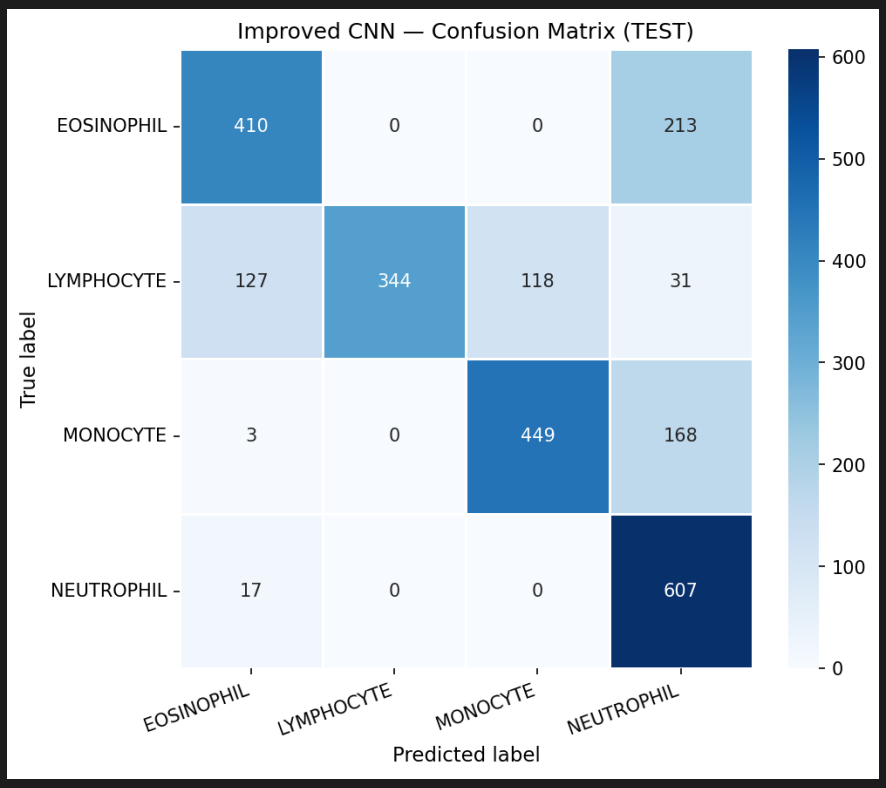
۵.۵. مقایسه‌ی عددی با فاز دوم

جدول ۱۰ مقایسه‌ی مستقیم بهترین مدل فاز سوم با مدل پایه‌ی فاز دوم را نشان می‌دهد.

جدول ۱۰: مقایسه‌ی مدل پایه (فاز دوم) با بهترین مدل فاز سوم

معیار	فاز دوم (مدل پایه)	فاز سوم (exp5)	تغییر
دقت آزمون	%۷۹.۵۳	%۷۲.۳۸	pp ۷.۱۵−
دقت اعتبارسنجی	%۹۴.۵۷	%۹۹.۵۰	pp ۴.۹۳+
F1 ماکرو	0.80	0.73	۰.۰۷−
AUC ماکرو	0.9552	0.9359	۰.۰۱۹۳−

۶.۵. تحلیل ماتریس درهم‌ریختگی



شکل ۱۵: ماتریس درهم‌ریختگی مدل نهایی فاز سوم روی مجموعه‌ی آزمون

چند نکته‌ی کلیدی از ماتریس درهم‌ریختگی:

- **NEUTROPHIL بیش‌تشخیص:** با Recall=0.97 اما Precision=0.60، مدل نمونه‌های زیادی از سایر کلاس‌ها — خصوصاً LYMPHOCYTE و EOSINOPHIL — را به‌اشتباه NEUTROPHIL تشخیص می‌دهد.
- **LYMPHOCYTE کم‌بازخوانی:** با Precision=1.00 اما Recall=0.55، هر نمونه‌ای که مدل LYMPHOCYTE می‌خواند درست است، اما حدود ۴۵% از LYMPHOCYTEهای واقعی از دست می‌روند و احتمالاً به NEUTROPHIL نسبت داده می‌شوند.
- **EOSINOPHIL:** با Recall=0.66، حدود ۳۴% از نمونه‌ها درست تشخیص داده نشدند. از آنجا که Eosinophil و Neutrophil هر دو گرانولوسیت با هسته‌ی چندلوبه‌ای هستند، تشابه ظاهری آن‌ها دلیل اصلی این اشتباهات است.

- **MONOCYTE:** با F1=0.76، بهترین عملکرد را در میان کلاس‌های مشکل‌دار داشت.

۷.۵. علت کاهش عملکرد نسبت به فاز دوم

علی‌رغم دقت اعتبارسنجی بالاتر (۹۹.۵۰% در برابر ۹۴.۵۷%)، دقت آزمون از ۷۹.۵۳% به ۷۲.۳۸% کاهش یافت. دو دلیل اصلی این پدیده عبارتند از:

۱. **جابجایی توزیع (Distribution Shift):** پوشه‌های TRAIN و TEST دارای ویژگی‌های تصویری متفاوتی هستند (روشنایی، کنتراست، رنگ‌بندی پس‌زمینه). مدل فاز سوم با افزایش داده‌ی قوی‌تر به توزیع آموزشی بهتر تطبیق یافت، اما همین امر تعمیم‌پذیری روی توزیع متفاوت آزمون را کاهش داد. افزایش داده طراحی‌شده برای تنوع درون‌توزیعی است، نه پوشش‌دادن به تفاوت‌های بین‌توزیعی.

۲. **معیار اعتبارسنجی گمراه‌کننده:** دقت اعتبارسنجی بالا (99.5%) تصویر خوش‌بینانه‌ای داد، چون زیرمجموعه‌ی اعتبارسنجی از همان TRAIN جدا شده و با توزیع آزمون فرق دارد. پس از رفع باگ validation_split در فاز قبل، دقت اعتبارسنجی بسیار بالا شد، اما این به‌تنهایی تضمین‌کننده‌ی دقت آزمون نیست.

در مجموع، اگرچه تکنیک‌های افزایش داده، BN، و Dropout به بهبود آموزش کمک کردند، مشکل اصلی — جابجایی توزیع میان مجموعه‌های TRAIN و TEST — با این روش‌ها برطرف نشد. انتظار می‌رود فازهای بعدی که از مدل‌های ازپیش‌آموزش‌دیده‌ی بزرگ‌تر بهره می‌برند، این شکاف را کاهش دهند.

۶. تفسیر مدل

برای اینکه بفهمیم مدل واقعاً به چه چیزی در تصویر توجه می‌کند، دو کار انجام دادیم: اول وزن فیلترهای کانولوشن را بررسی کردیم، و بعد از Grad-CAM استفاده کردیم تا ببینیم هنگام تشخیص هر سلول، مدل کجای تصویر را نگاه می‌کند. مدل مورد استفاده همان exp5 از فاز سوم است (BN + Dropout 0.5 + batch=64).

۱.۶. فیلترهای کانولوشن

شکل ۱۶ وزن فیلترهای سه لایه را نشان می‌دهد: اولین لایه (conv2d، 32 فیلتر)، دومین لایه (conv2d_1، 64 فیلتر)، و چهارمین لایه (conv2d_3، 128 فیلتر).

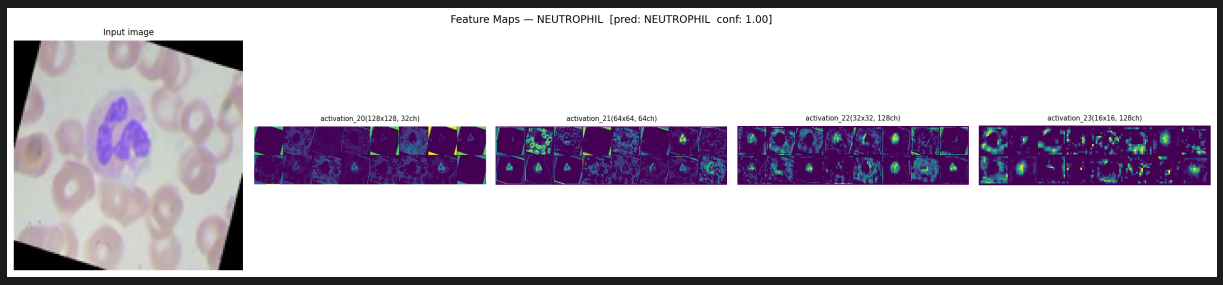
شکل ۱۶: وزن فیلترهای سه لایه‌ی کانولوشن — از چپ: لایه‌ی اول، دوم، چهارم

در لایه‌ی اول هر فیلتر مستقیماً روی پیکسل‌های RGB کار می‌کند؛ به همین دلیل وزن‌ها به شکل پچ‌های رنگی کوچک دیده می‌شوند. تنوع رنگی زیاد نشان می‌دهد مدل از همان ابتدا به طیف رنگی متفاوتی — از بنفش

تا صورتی تا سبز — حساس شده که در رنگ‌آمیزی H&E تصاویر خون معنادار است. در لایه‌های عمیق‌تر دیگر نمی‌توان فیلترها را با رنگ تفسیر کرد چون هر فیلتر روی ترکیبی از ده‌ها کانال قبلی عمل می‌کند؛ اینجا بزرگی وزن‌ها را به صورت خاکستری نشان می‌دهیم. الگوهای تیره‌وروشن در این لایه‌ها با فیلترهای لبه‌یاب و بافت‌یاب تطابق دارند.

۲.۶. نقشه‌های فعال‌سازی

برای هر کلاس یک نمونه‌ی درست با اطمینان بالا انتخاب کردیم و خروجی چهار لایه‌ی فعال‌سازی را با هم مقایسه کردیم. شکل ۱۷ نتیجه را برای کلاس NEUTROPHIL نشان می‌دهد.



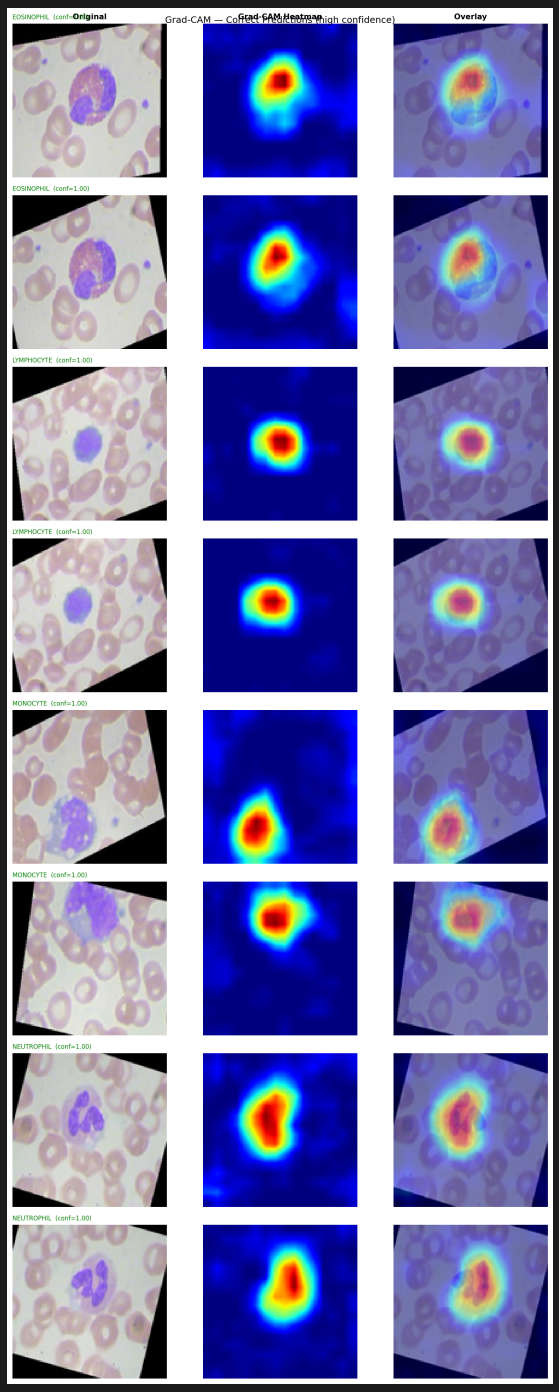
شکل ۱۷: نقشه‌های فعال‌سازی چهار مرحله برای یک نمونه‌ی NEUTROPHIL

در لایه‌ی اول (128×128) تقریباً همه‌ی نقشه‌ها فعال‌اند و ساختار کلی تصویر — مرز سلول‌ها و رنگ پس‌زمینه — خوانا است. با رفتن به لایه‌های بعدی، نقشه‌ها خلوت‌تر می‌شوند و فقط نواحی خاص‌تری روشن می‌مانند. در لایه‌ی آخر (16×16) فقط چند نورون فعال هستند که به نواحی بسیار مشخصی مثل هسته‌ی سلول اشاره دارند. این رفتار کاملاً طبیعی است: هرچه عمق بیشتر، ویژگی انتزاعی‌تر و مکان‌یابی دقیق‌تر.

۳.۶. Grad-CAM

Grad-CAM از گرادیان امتیاز کلاس هدف نسبت به خروجی آخرین لایه‌ی کانولوشن استفاده می‌کند تا نشان دهد کدام نواحی تصویر بیشترین سهم را در آن تصمیم داشتند. نواحی قرمز و زرد در نقشه‌ی حرارتی جاهایی هستند که مدل به آن‌ها توجه کرده.

شکل ۱۸ نتیجه را برای هشت نمونه‌ی درست از مجموعه‌ی آزمون نشان می‌دهد — دو نمونه از هر کلاس.



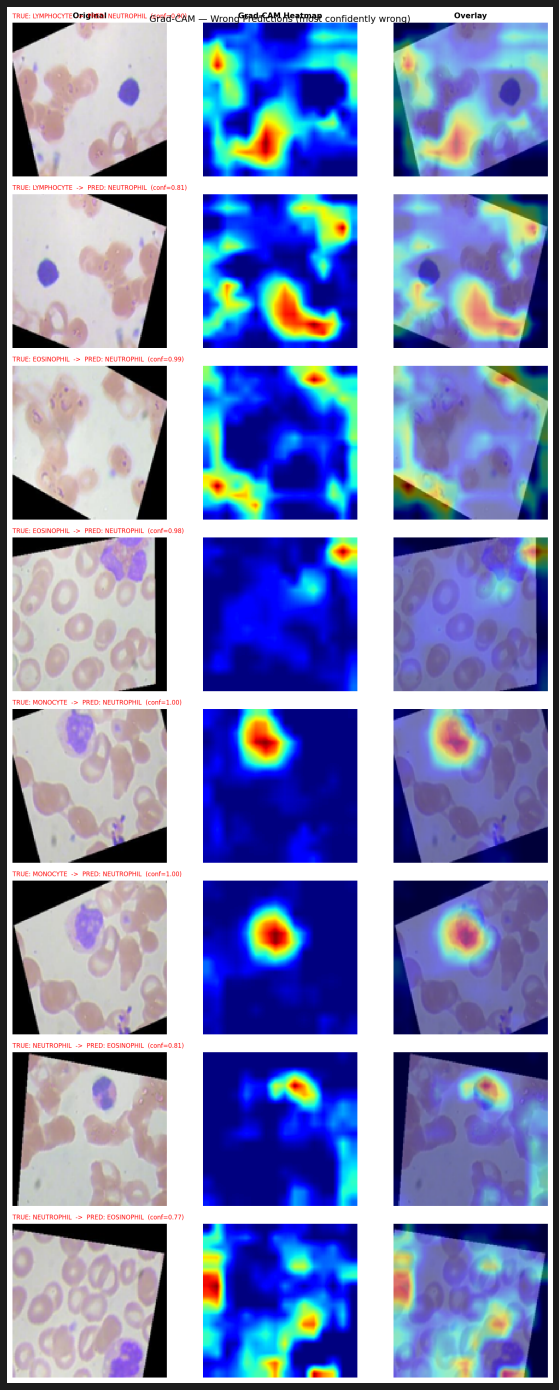
شکل ۱۸: Grad-CAM برای پیش‌بینی‌های درست — دو نمونه از هر کلاس

در همه‌ی این موارد نقطه‌ی داغ روی سلول اصلی قرار گرفته و نسبتاً فشرده است. میانگین مساحت فعال (نواحی با heatmap > 0.5) برای همه‌ی کلاس‌ها بین 4.5% تا 5.8% بود؛ یعنی مدل تنها روی یک ناحیه‌ی کوچک

و مشخص از تصویر تمرکز می‌کند نه کل صحنه.

۴.۶. مقایسه‌ی پیش‌بینی‌های درست و اشتباه

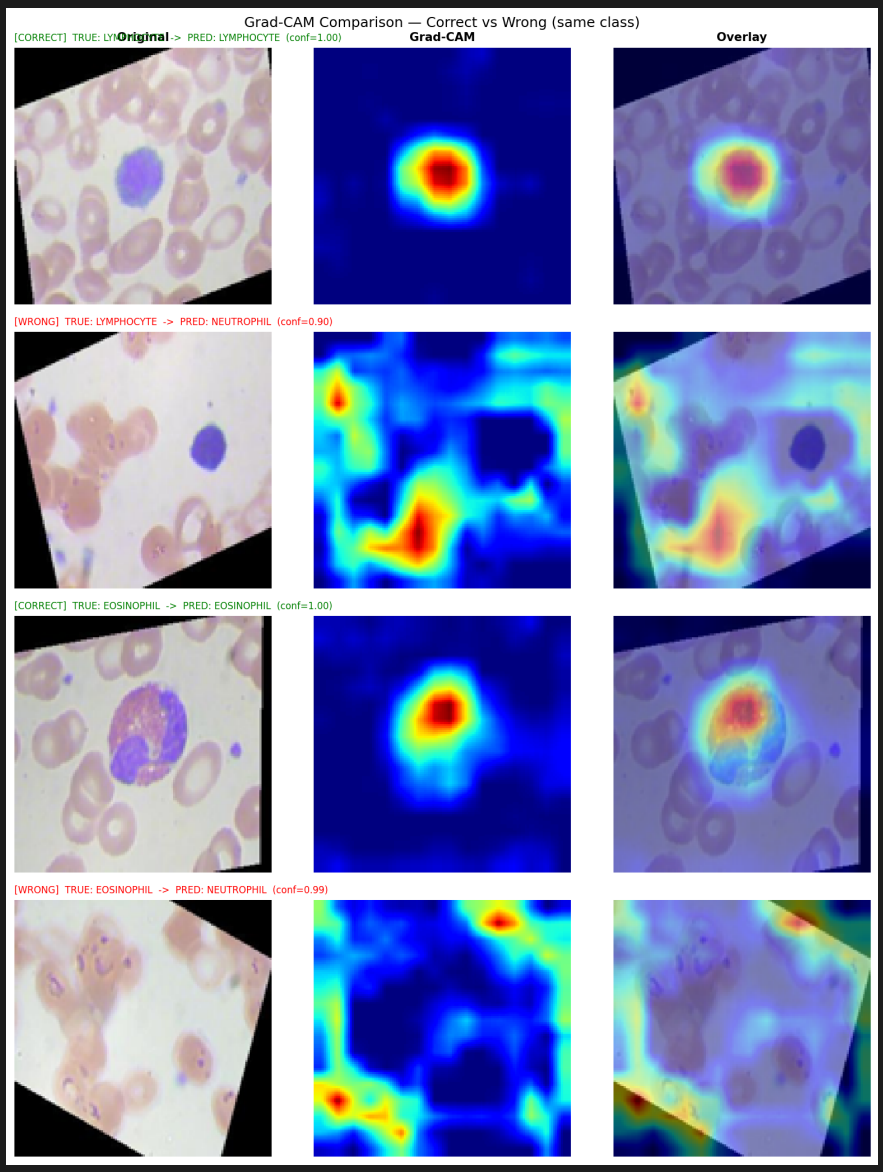
شکل ۱۹ همین تحلیل را برای هشت نمونه‌ی اشتباه نشان می‌دهد که مدل آن‌ها را با اطمینان بالا اشتباه تشخیص داده.



شکل ۱۹: Grad-CAM برای پیش‌بینی‌های اشتباه

تفاوت با نمونه‌های درست کاملاً مشهود است. نقشه‌ی حرارتی پراکنده است و روی ناحیه‌ی مشخصی تمرکز ندارد؛ در بعضی موارد مدل به پس‌زمینه یا سلول‌های قرمز اطراف توجه کرده نه به سلول اصلی. این رفتار نشان

می‌دهد وقتی مدل گیج است، به جای پیدا کردن ویژگی تمایزدهنده، به اطلاعات زمینه‌ای تکیه می‌کند. شکل ۲۰ این تفاوت را برای دو کلاس به صورت مقایسه‌ای نشان می‌دهد.



شکل ۲۰: مقایسه‌ی Grad-CAM برای پیش‌بینی درست و اشتباه از یک کلاس

در نمونه‌ی درست LYMPHOCYTE نقطه‌ی داغ دقیقاً روی هسته‌ی گرد و بزرگ قرار گرفته — همان چیزی که این سلول را بصری متمایز می‌کند. در نمونه‌ی اشتباه همان کلاس که به NEUTROPHIL تشخیص داده شده، توجه مدل پراکنده است و به سلول‌های قرمز اطراف هم کشیده شده. برای EOSINOPHIL هم همین الگو تکرار می‌شود: نمونه‌ی درست روی هسته‌ی چندلوبه‌ای متمرکز است، نمونه‌ی اشتباه روی هیچ ناحیه‌ی مشخصی تمرکز ندارد. این تحلیل دلیل Recall پایین LYMPHOCYTE (0.55) و EOSINOPHIL (0.66) در فاز سوم را روشن می‌کند: مدل در نمونه‌های دشوار نمی‌تواند ویژگی تمایزدهنده را پیدا کند.

۷. یادگیری انتقالی با MobileNetV2

در فازهای قبل بهترین دقت آزمون به 72.38% رسید. یکی از دلایل اصلی این عدد، جابجایی توزیع میان مجموعه‌های TRAIN و TEST بود — ویژگی‌هایی که مدل از داده‌های آموزشی یاد گرفته با آنچه در آزمون دیده می‌شود کاملاً منطبق نیست. مدل‌های ازپیش‌آموزش‌دیده روی ImageNet این مشکل را کاهش می‌دهند چون از همان ابتدا ویژگی‌های پایه‌ای قوی‌تری دارند که خوب‌تر تعمیم می‌یابند.

برای این فاز از MobileNetV2 استفاده کردیم. این مدل نسبت به گزینه‌های سنگین‌تر مثل ResNet50 یا EfficientNet پارامتر کمتری دارد و روی GPU های معمولی راحت‌تر آموزش می‌بیند، در حالی که دقت قابل قبولی هم ارائه می‌دهد.

۱.۷. تغییر لایه‌ی خروجی

MobileNetV2 در حالت استاندارد برای طبقه‌بندی 1000 کلاس ImageNet طراحی شده. برای تطبیق آن با مسئله‌ی ما (4 کلاس)، لایه‌ی top اصلی حذف شد (include_top=False) و یک سر طبقه‌بند سبک جایگزین آن شد:

۱. GlobalAveragePooling2D — خروجی سه‌بُعدی پشته‌ی کانولوشنی را به یک بردار یک‌بُعدی تبدیل می‌کند. در مقایسه با Flatten، پارامتر کمتری دارد و به موقعیت مکانی ویژگی‌ها حساسیت کمتری نشان می‌دهد.

۲. Dropout — برای جلوگیری از overfitting سر طبقه‌بند.

۳. Dense(4, softmax) — لایه‌ی خروجی با 4 نورون برای چهار کلاس سلول خونی.

نکته‌ی مهم پیاده‌سازی این است که در تابع forward pass، مدل پایه با آرگومان training=False فراخوانی می‌شود. این کار لایه‌های BatchNormalization درون MobileNetV2 را در حالت استنتاج نگه می‌دارد — یعنی از میانگین و واریانس تجمیعی استفاده می‌کنند نه از آمار دسته‌ی جاری. بدون این تنظیم، آموزش سر طبقه‌بند با پارامترهای BN منجمدشده ناپایدار می‌شود.

۲.۷. پیش‌پردازش ورودی

MobileNetV2 انتظار دارد پیکسل‌ها در بازه‌ی $[-1, +1]$ باشند، نه $[0, 1]$. به همین دلیل به جای تقسیم ساده بر 255، از تابع mobilenet_v2.preprocess_input استفاده شد که پیکسل‌ها را از $[0, 255]$ به $[-1, +1]$ نگاشت می‌کند. این تفاوت ظاهراً جزئی در عمل تأثیر قابل‌توجهی روی کیفیت آموزش دارد، چون وزن‌های ImageNet با همین توزیع ورودی بهینه شده‌اند.

۳.۷. پروتکل آموزش دو مرحله‌ای

۱.۳.۷. مرحله‌ی اول — آموزش سر طبقه‌بند

در مرحله‌ی اول تمام لایه‌های MobileNetV2 منجمد (freeze) شدند و فقط لایه‌های سر طبقه‌بند آموزش دیدند. نرخ یادگیری 10^{-3} انتخاب شد — چون این لایه‌ها از صفر شروع می‌کنند و می‌توان با قدم‌های بزرگ‌تر آموزش داد. هدف این مرحله این است که سر طبقه‌بند ابتدا یاد بگیرد ویژگی‌های MobileNetV2 را تفسیر کند، بدون اینکه وزن‌های ازپیش‌آموزش‌دیده را خراب کند.

۲.۳.۷. مرحله‌ی دوم — Fine-tuning لایه‌های بالایی

پس از همگرایی مرحله‌ی اول، چند لایه‌ی آخر backbone از حالت منجمد خارج (unfreeze) شدند و آموزش با نرخ یادگیری بسیار کوچک‌تر (10^{-5}) ادامه یافت. دلیل انتخاب نرخ یادگیری کوچک این است که وزن‌های ImageNet در این لایه‌ها از قبل خوب هستند؛ فقط نیاز به تطبیق ظریف با دامنه‌ی تصاویر خون داریم. اگر نرخ یادگیری بالا باشد، این وزن‌ها خراب می‌شوند و دقت به سطح تصادفی سقوط می‌کند.

۴.۷. تنظیم ابرپارامترها

در مجموع چهار آزمایش انجام شد. دو آزمایش در مرحله‌ی اول با نرخ Dropout متفاوت، و دو آزمایش در مرحله‌ی دوم با تعداد لایه‌های unfreeze متفاوت.

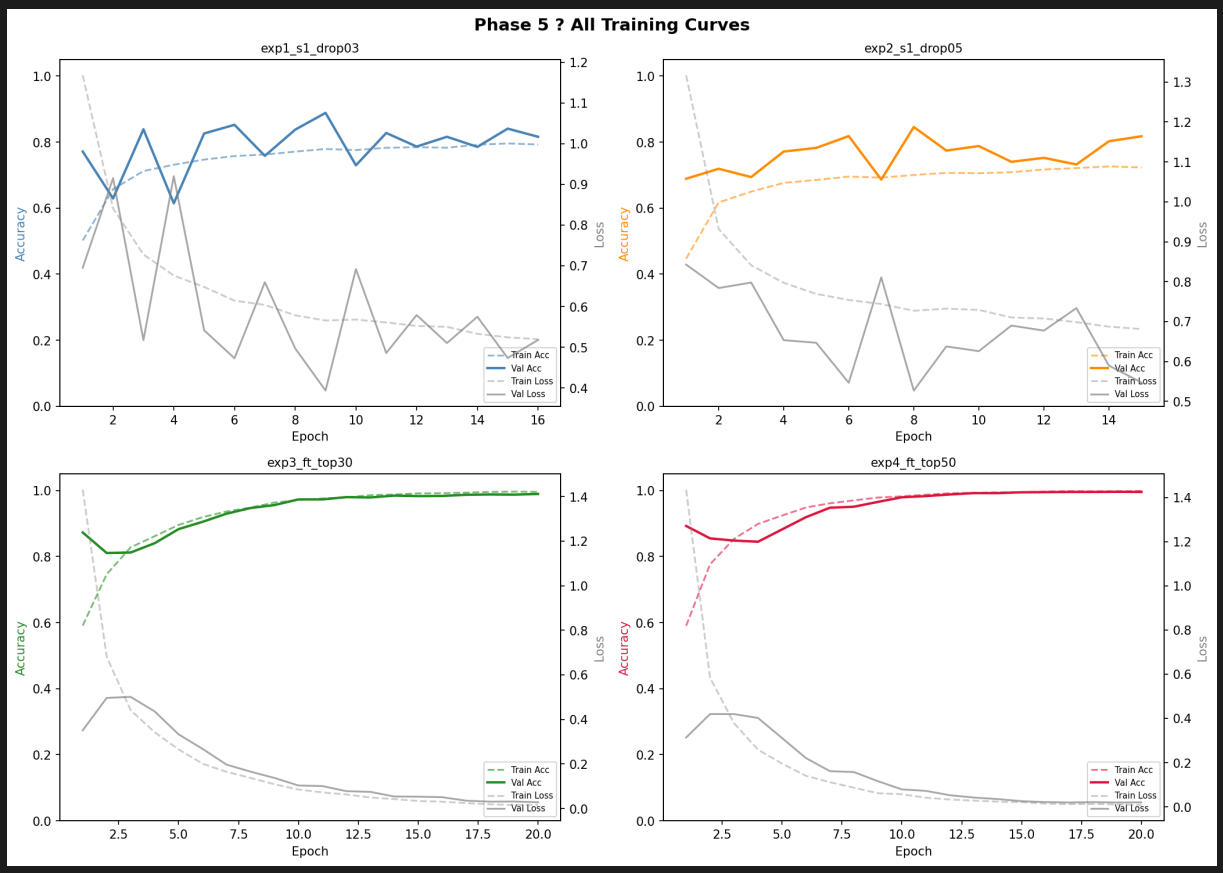
آزمایش	شرح	مرحله	Dropout	لایه‌های آزادشده	بهترین Val Acc
exp1	سر منجمد، Dropout سبک	1	0.3	—	88.85%
exp2	سر منجمد، Dropout قوی	1	0.5	—	84.58%
exp3	Fine-tune کم‌عمق	2	بهترین S1	30 لایه‌ی آخر	98.95%
exp4	Fine-tune عمیق‌تر	2	بهترین S1	50 لایه‌ی آخر	99.60%

جدول ۱۱: آزمایش‌های تنظیم ابرپارامتر — فاز پنجم

در هر دو مرحله از EarlyStopping با صبر 7 و ReduceLROnPlateau با فاکتور 0.5 و صبر 4 استفاده شد تا از آموزش بیش از حد و نوسانات نرخ یادگیری جلوگیری شود.

۵.۷. نمودارهای آموزش

شکل ۲۱ نمودارهای دقت و Loss را برای هر چهار آزمایش نشان می‌دهد.



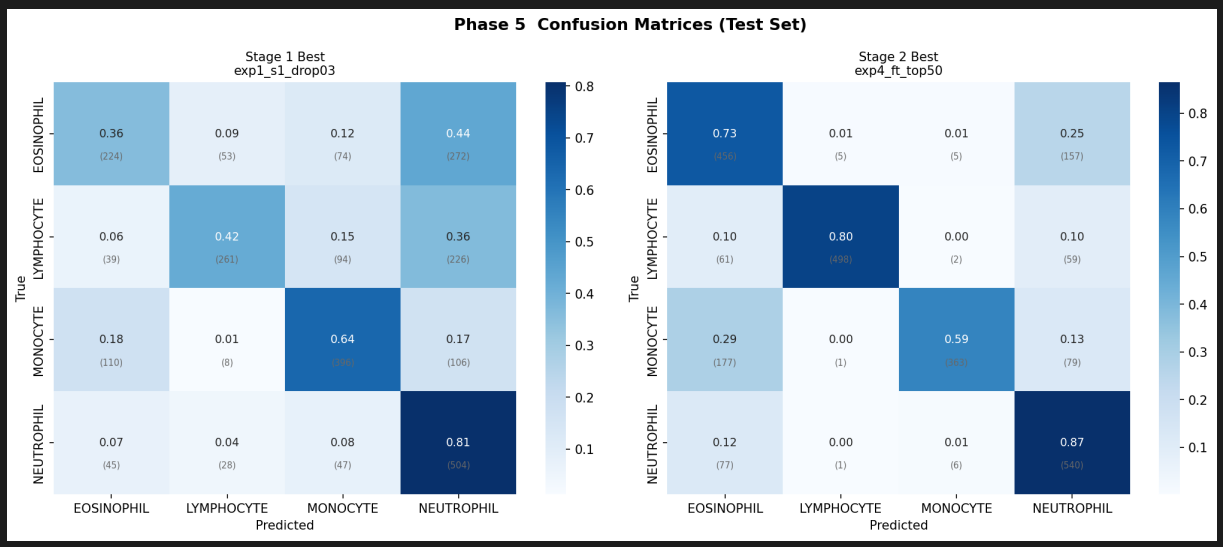
شکل ۲۱: نمودارهای آموزش چهار آزمایش — فاز پنجم

در آزمایش‌های مرحله‌ی اول (exp1 و exp2) دقت اعتبارسنجی نسبتاً سریع به یک سقف می‌رسد چون تمام قدرت ویژگی‌سازی از ImageNet می‌آید و سر طبقه‌بند فقط باید یاد بگیرد این ویژگی‌ها را تفسیر کند. در آزمایش‌های مرحله‌ی دوم (exp3 و exp4) آموزش از جایی که مرحله‌ی اول متوقف شده ادامه می‌یابد و با قدم‌های آرام‌تری بهبود می‌کند.

۶.۷. ارزیابی روی مجموعه‌ی آزمون

۱.۶.۷. ماتریس درهم‌ریختگی

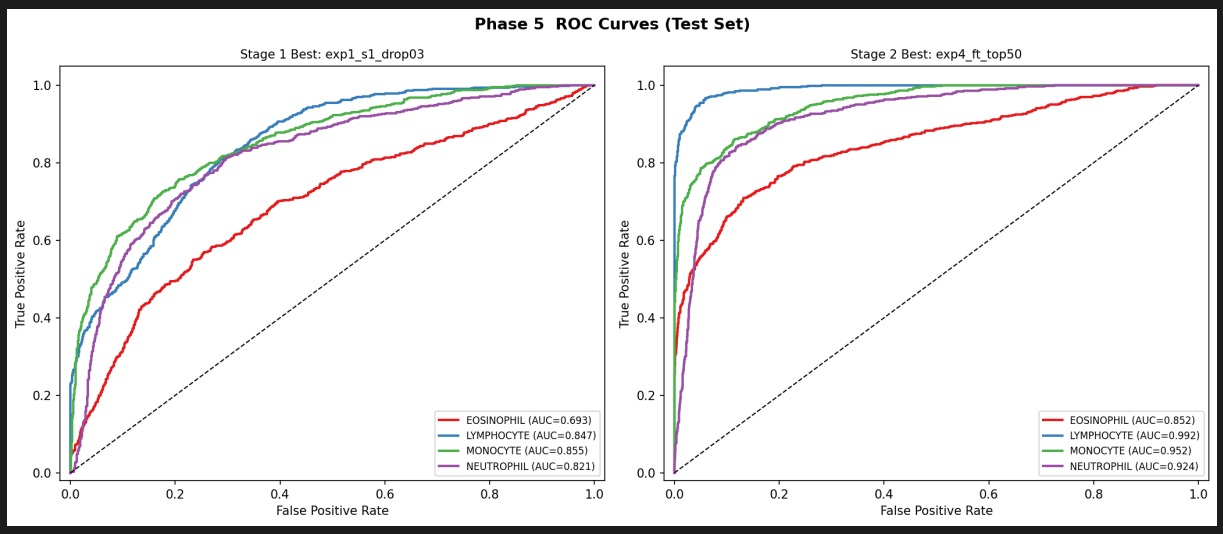
شکل ۲۲ ماتریس درهم‌ریختگی را برای بهترین مدل مرحله‌ی اول و بهترین مدل مرحله‌ی دوم نشان می‌دهد.



شکل ۲۲: ماتریس درهم‌ریختگی روی مجموعه‌ی آزمون — مرحله‌ی اول و دوم

۲.۶.۷. منحنی‌های ROC

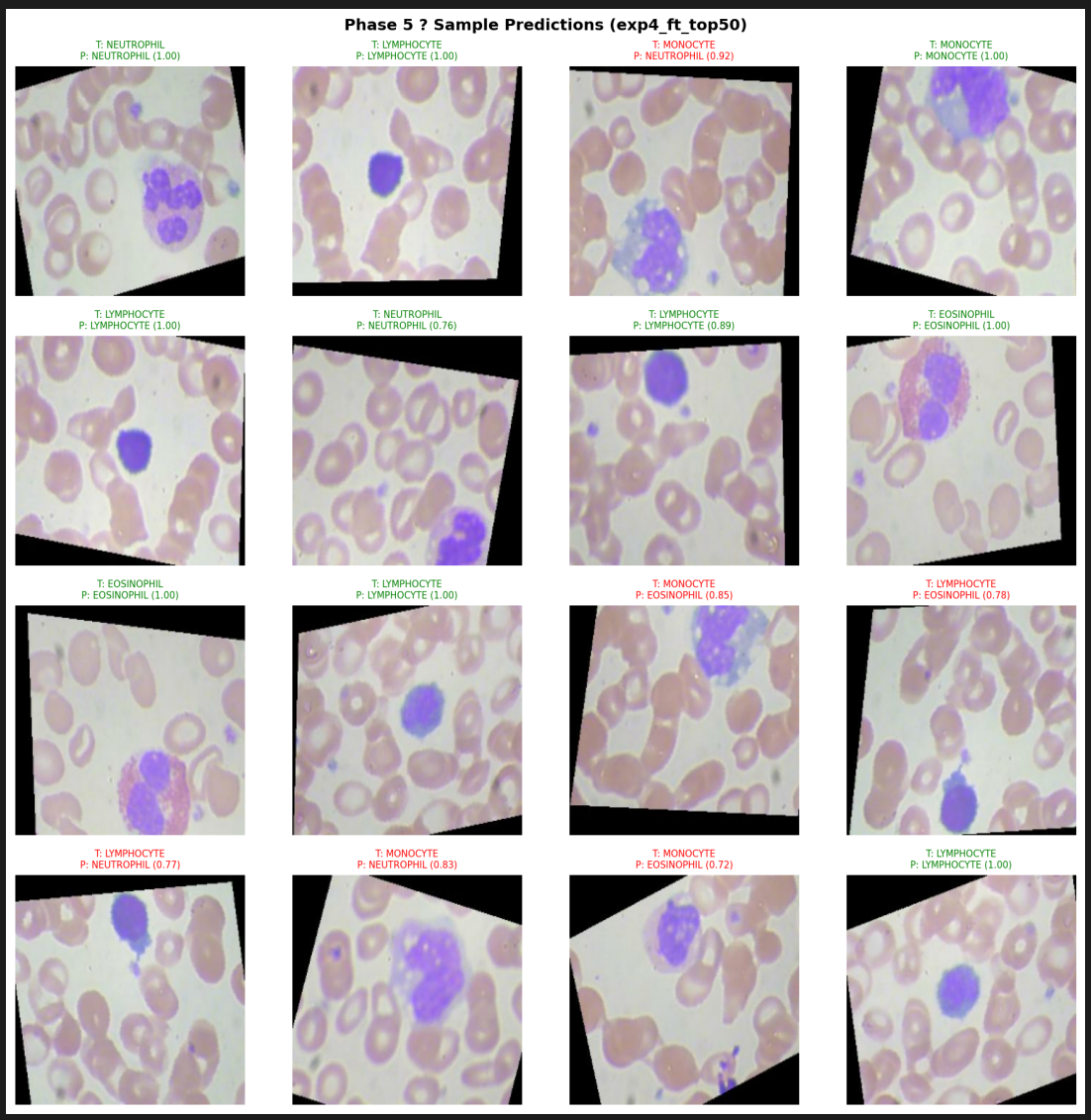
شکل ۲۳ منحنی‌های ROC را به تفکیک کلاس نشان می‌دهد. مساحت زیر منحنی (AUC) برای هر کلاس در داخل نمودار درج شده است.



شکل ۲۳: منحنی‌های ROC روی مجموعه‌ی آزمون — مرحله‌ی اول و دوم

۳.۶.۷. نمونه‌های پیش‌بینی

شکل ۲۴ شانزده نمونه از مجموعه‌ی آزمون را با برچسب واقعی و برچسب پیش‌بینی‌شده نشان می‌دهد. پیش‌بینی‌های درست با کادر سبز و اشتباه‌ها با کادر قرمز مشخص شده‌اند.



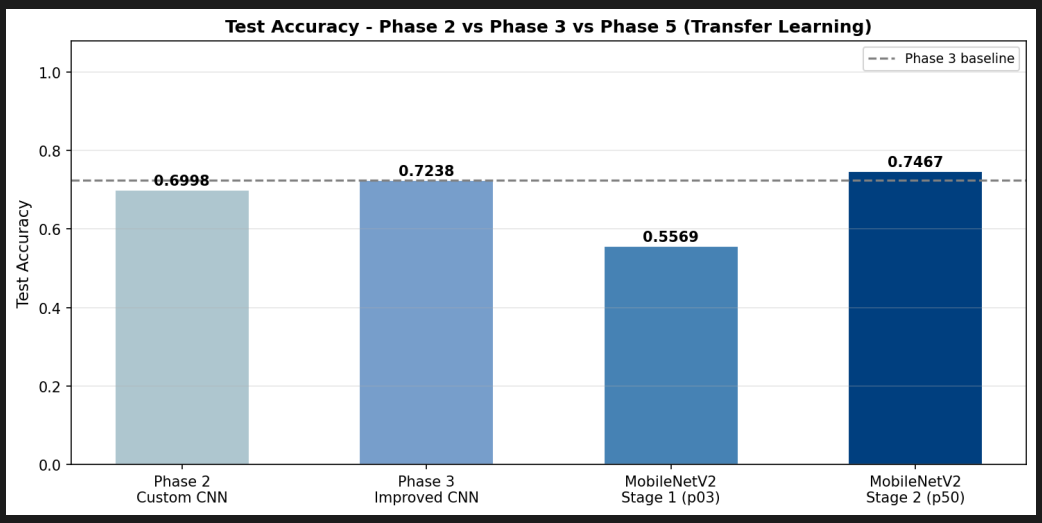
شکل ۲۴: نمونه‌های پیش‌بینی روی مجموعه‌ی آزمون — بهترین مدل فاز پنجم

۷.۷. مقایسه‌ی نهایی با فازهای قبل

جدول ۱۲ دقت آزمون تمام فازها را کنار هم قرار می‌دهد و شکل ۲۵ همین مقایسه را به صورت نمودار میله‌ای نشان می‌دهد.

مدل	فاز	دقت آزمون
بهترین CNN سفارشی	فاز دوم	69.98%
بهترین CNN بهبودیافته	فاز سوم	72.38%
MobileNetV2 مرحله‌ی اول	فاز پنجم	55.69%
MobileNetV2 مرحله‌ی دوم	فاز پنجم	74.67%

جدول ۱۲: مقایسه‌ی دقت آزمون در تمام فازها



شکل ۲۵: مقایسه‌ی دقت آزمون — فاز دوم، سوم، و پنجم

۸.۷ تحلیل نهایی

نتایج این فاز تصویر جالب و در عین حال صادقانه‌ای از محدودیت‌های این دیتاست ارائه می‌دهد.

شکاف اعتبارسنجی و آزمون. بزرگ‌ترین چیزی که این فاز آشکار کرد همان مشکلی بود که از فاز سوم می‌شناختیم: جابجایی توزیع میان TRAIN و TEST. exp1 در اعتبارسنجی به 88.85% رسید اما روی آزمون فقط 55.69% داد — یعنی مدل با وجود وزن‌های ImageNet، به همان دام overfitting به توزیع آموزشی افتاد. exp4 این شکاف را کوچک‌تر کرد: اعتبارسنجی 99.60%، آزمون 74.67%. اما شکاف هنوز وجود دارد و نشان می‌دهد ریشه‌ی مشکل در خود دیتاست است، نه فقط در معماری مدل.

مقایسه با فازهای قبل. بهترین مدل این فاز (exp4، دقت آزمون 74.67%) نسبت به بهترین مدل فاز سوم (72.38%) کمی بهتر شد — حدود 2.3 واحد درصد. این بهبود کوچک‌تر از آن چیزی است که معمولاً از Transfer Learning انتظار می‌رود، اما با توجه به جابجایی توزیع دیتاست قابل توضیح است: وزن‌های ImageNet کمک کردند ویژگی‌های بهتری استخراج شود، اما مشکل اصلی — که تفاوت توزیع است — با هیچ معماری‌ای به‌تنهایی برطرف نمی‌شود.

تحلیل کلاس به کلاس. در بهترین مدل (exp4) کلاس LYMPHOCYTE با Precision بالای 0.99 بهترین عملکرد را داشت — شاید چون این سلول ویژگی بصری مشخص‌تری (هسته‌ی گرد بزرگ) دارد که MobileNetV2 به خوبی آن را یاد گرفت. در مقابل، MONOCYTE با Recall پایین 0.59 مشکل‌دارترین کلاس بود: مدل اکثر MONOCYTE های واقعی را به کلاس‌های دیگر تشخیص داد، که احتمالاً به دلیل شباهت ظاهری این سلول با LYMPHOCYTE در نمونه‌های آزمون است.

نتیجه‌گیری کلی. از میان همه‌ی مدل‌های آزمایش‌شده در این پروژه، MobileNetV2 با Fine-tuning (exp4) بهترین دقت آزمون را داشت. اما بهبود نسبت به فاز سوم اندک بود، چون مشکل اصلی این دیتاست — جابجایی توزیع بین مجموعه‌های آموزش و آزمون — یک مشکل ساختاری است که با انتخاب معماری بهتر به‌تنهایی حل نمی‌شود. برای حل اساسی‌تر این مشکل، روش‌هایی مثل domain adaptation یا جمع‌آوری داده‌ی آموزشی متنوع‌تر لازم است.