دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده برق

پایان‌نامه کارشناسی
رشته و گرایش الکترونیک

تشخیص سرطان لوسمی در نمونه‌های میکروسکوپیک خون با استفاده از شبکه عصبی عمیق

نگارش
پارسا پیله‌ور

استاد راهنما
دکتر سیدین

مرداد ۱۴۰۳

صفحه فرم ارزیابی و تصویب پایان‌نامه- فرم تأیید اعضای کمیته دفاع

در این صفحه (هر سه مقطع تحصیلی) باید تصویر فرم ارزیابی یا تأیید و تصویب پایان‌نامه/رساله موسوم به فرم کمیته دفاع برای مقاطع کارشناسی‌ارشد و دکتری و تصویر فرم تصویب برای مقطع کارشناسی، موجود در پرونده آموزشی را قرار دهند.

نکات مهم:

- ✓ نگارش پایان‌نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان‌نامه‌های دانشگاه صنعتی امیرکبیر باشد. (دستورالعمل و راهنمای حاضر)؛
- ✓ تحویل پایان‌نامه به زبان انگلیسی، برای دانشجویان بین‌الملل با شرایط دستورالعمل حاضر بلامانع است و داشتن صفحه عنوان فارسی به همراه چکیده مبسوط فارسی، ۳۰ صفحه برای پایان‌نامه کارشناسی‌ارشد و ۵۰ صفحه برای رساله دکتری در ابتدای آن الزامی است؛
- ✓ دریافت پایان‌نامه کارشناسی،کارشناسی‌ارشد و رساله دکتری، بصورت نسخه دیجیتال مطابق راهنمای وبسایت و دستورالعمل حاضر می‌باشد؛
- ✓ درصورتی که یک عنوان پایان‌نامه دارای دو نویسنده است، فقط یکبار فایل و فرم اطلاعات آن با ذکر هر دو نویسنده بارگذاری و تکمیل گردد؛
- ✓ با توجه به اینکه در ورود۲۰۱۶یا بالاتر، احتمال تغییر ترتیب ذکر زیر فصل‌ها وجود دارد لطفا در انتها به شماره‌دهی زیر فصل‌ها توجه نمایید که بصورت صحیح باشد. از راست به چپ: شماره فصل-زیرفصل۱-زیرفصل۲-زیرفصل۳ و ... .

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ: مرداد ۱۴۰۳

اینجانب پارسا پیله‌ور متعهد می‌شوم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است.
نقل مطالب با ذکر مآخذ بلامانع است.

در صفحه تعهدنامه اصالت اثر، در قسمت بالا سمت چپ، تاریخ دفاع خود را جایگزین تاریخ نوشته شده کنید.

پارسا پیله‌ور

امضا

سپاس گزاری

بدینوسیله مراتب قدردانی و امتنان خود را خدمت،

جناب دکتر سیدین، استاد راهنمای گرانقدر،

ابراز و از تمامی زحمات ایشان تشکر می نمایم.

پارسا پیله‌ور

مرداد ۱۴۰۳

چکیده

لوسمی‌ها یک دسته کشنده از بیماری‌های سرطانی هستند که افراد در تمام سنین، از جمله کودکان و بزرگسالان را تحت تأثیر قرار می‌دهند و یکی از علل مهم مرگ و میر در سراسر جهان هستند. به طور خاص، این بیماری با افزایش تعداد لنفوسیت‌های نابالغ همراه است و باعث آسیب به مغز استخوان و یا خون می‌شود. در حال حاضر، تجزیه و تحلیل دستی نمونه‌های خون به دست آمده از طریق تصاویر میکروسکوپی برای تشخیص این بیماری انجام می‌شود که اغلب بسیار کند، زمان‌بر و کم دقت است.

تشخیص خودکار لوسمی یا سرطان خون کاری چالش‌برانگیز و بسیار مورد نیاز در مراکز درمانی است. در دهه‌های گذشته، یادگیری عمیق با استفاده از شبکه‌های عصبی عمیق رویکردهای پیشرفته‌ای برای مسائل طبقه‌بندی تصویر ارائه داده‌اند. با این حال، هنوز فاصله‌ای برای بهبود کارایی، روند یادگیری و عملکرد آنها وجود دارد. بنابراین، در این مطالعه تحقیقاتی، ما یک نسخه جدید از الگوریتم یادگیری عمیق را برای تشخیص بیماری لوسمی لنفوبلاستیک حاد (ALL) مبتنی بر یادگیری عمیق و ماشینی با تجزیه و تحلیل تصاویر میکروسکوپی نمونه‌های خون پیشنهاد کرده‌ایم.

ما مدل ترکیبی ViT-CNN را برای طبقه‌بندی تصاویر سلول‌های سرطانی و سلول‌های طبیعی جهت کمک به تشخیص لوسمی لنفوبلاستیک حاد پیشنهاد می‌کنیم. مدل ViT-CNN یک مدل ترکیبی است که مدل ترانسفورمر بینایی و مدل شبکه عصبی کانولوشنی را با هم ترکیب می‌کند. مدل ترکیبی ViT-CNN می‌تواند ویژگی‌های تصاویر سلول‌ها را به دو روش کاملاً متفاوت استخراج کند تا به نتایج طبقه‌بندی بهتری دست یابد. دقت طبقه‌بندی مدل ترکیبی ViT-CNN بر روی مجموعه تست به ۹۵/۷۴٪ رسیده است که قابل مقایسه با روش‌های دیگر است. روش پیشنهادی می‌تواند سلول‌های سرطانی را به‌طور دقیق از سلول‌های عادی تمییز دهد و می‌تواند به‌عنوان یک روش مؤثر برای تشخیص به کمک کامپیوتر در لوسمی لنفوبلاستیک حاد استفاده شود.

واژه‌های کلیدی:

لوسمی لنفوبلاستیک حاد، پردازش تصویر در پزشکی، یادگیری عمیق، ترانسفورمر بینایی.

فهرست مطالب

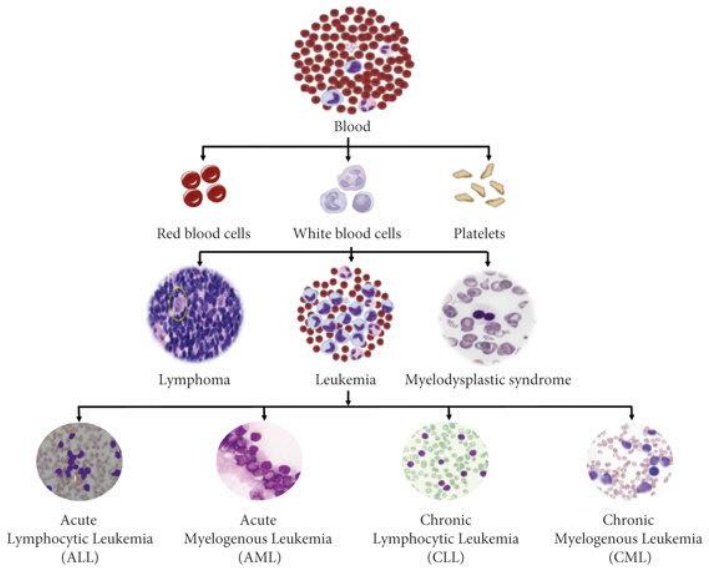
فهرست اشکال

فهرست جداول

فصل اول: مقدمه

لوسمی، نوعی سرطان خون است که از مغز استخوان منشأ می‌گیرد و مقدار زیادی سلول‌های خونی غیرطبیعی، که معمولاً به عنوان بلاست‌ها یا سلول‌های لوسمی شناخته می‌شوند، تولید می‌کند. این سلول‌های خونی نابالغ باعث علائم مختلفی، از جمله خونریزی، کبودی، درد استخوان، خستگی، تب و افزایش حساسیت به عفونت‌ها به دلیل کمبود سلول‌های خونی طبیعی می‌شوند. معمولاً برای تأیید تشخیص لوسمی، نمونه مغز استخوان یا آزمایش‌های خون لازم است [۱].

تمایز و تشخیص انواع مختلف لوسمی می‌تواند توسط هماتولوژیست‌ها در مراکز پیوند سلول بر اساس تصاویر میکروسکوپی انجام شود. اسلایدهای رنگ‌آمیزی شده مناسب می‌توانند در تشخیص برخی انواع لوسمی کمک کنند [۷]. شایع‌ترین انواع لوسمی را می‌توان در اسلایدهای رنگ‌آمیزی شده شناسایی کرد، همانطور که در شکل ۱-۱ نشان داده شده است. لوسمی به چهار دسته اصلی طبقه‌بندی می‌شود، از جمله لوسمی میلوئیدی حاد (AML)، لوسمی لنفوبلاستیک حاد (ALL)، لوسمی میلوئیدی مزمن (CML)، و لوسمی لنفوسیتی مزمن (CLL)، علاوه بر چندین نوع کمتر شایع [۲].



شکل ۱-۱ چهار دسته متداول لوسمی

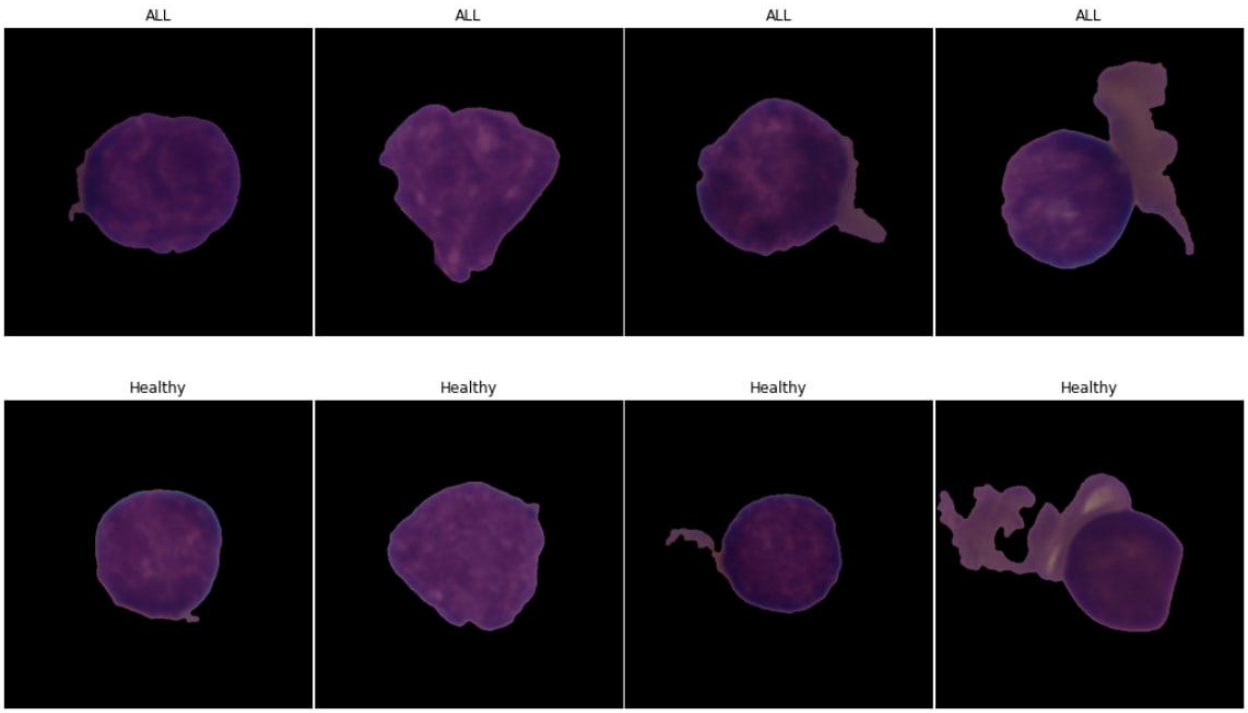
اگرچه علل دقیق لوسمی هنوز ناشناخته است، اعتقاد بر این است که نتیجه ترکیبی از عوامل ژنتیکی و محیطی اثرگذار هستند [۱]. در لوسمی لنفوبلاستیک حاد، مغز استخوان مقدار زیادی گلبول سفید غیرطبیعی و نابالغ

تولید می‌کند. این سلول‌های غیرطبیعی، تعداد گلبول‌های قرمز و سفید سالم و پلاکت‌ها را در خون و مغز استخوان کاهش می‌دهند [۲]. و مبارزه بدن با عفونت و بیماری را دشوارتر می‌کنند. این بیماری حاد قلمداد می‌شود زیرا به سرعت و با شدت پیشرفت می‌کند. این گلبول‌های سفید می‌توانند وارد جریان خون شوند و به قسمت‌های مختلف بدن مانند طحال، مغز، کلیه و کبد آسیب برسانند که می‌تواند منجر به انواع خطرناک دیگری از سرطان شود [۲، ۳]. از آنجا که لوسمی لنفوبلاستیک حاد می‌تواند به سرعت در سراسر بدن گسترش یابد، گاهی اوقات اگر در مراحل اولیه درمان یا تشخیص داده نشود، می‌تواند منجر به مرگ شود. اگر لوسمی، و به ویژه لوسمی لنفوبلاستیک حاد، در مراحل اولیه تشخیص داده شود، به راحتی قابل درمان است. علائم لوسمی شبیه به بیماری‌های دیگر مانند کم‌خونی، درد مفاصل، تب، ضعف و درد استخوان است، به همین دلیل تشخیص لوسمی گاهی اوقات دشوار می‌شود [۷]. لوسمی لنفوبلاستیک حاد شایع‌ترین بدخیمی و لوسمی دوران کودکی است. لوسمی لنفوبلاستیک حاد در کودکان معمولاً قابل درمان است. این بیماری در بزرگسالان نادرتر و درمان آن دشوارتر است، اما با درمان مناسب، بزرگسالان نیز می‌توانند بهبود یابند [۱]. در ایالات متحده، لوسمی بیش از ۳.۵ درصد از تمام تشخیص‌های جدید سرطان را تشکیل می‌دهد، به طوری که در سال ۲۰۱۸ بیش از ۶۰,۰۰۰ مورد جدید ثبت شده است [۳].

چندین روش تهاجمی وجود دارد که معمولاً توسط هماتولوژیست‌ها برای تشخیص بیماری استفاده می‌شود. اگرچه بررسی میکروسکوپی خون رایج‌ترین روش تشخیصی است [۶-۹]، اما نمونه‌برداری و آنالیز مغز استخوان نیز بسیار مهم محسوب می‌شود. به طور کلی، بیوپسی انجام می‌شود که یک روش تهاجمی است و با انجام آزمایش‌هایی روی خون، مغز استخوان یا مایع نخاعی صورت می‌گیرد. این روش‌ها دردناک، پرهزینه و زمان‌بر هستند. در این معاینات، متخصص پزشکی بررسی می‌کند که آیا تعداد سلول‌های سفید به طور غیر طبیعی بالاست و اگر مشاهدات فیزیکی مرتبط دیگری وجود داشته باشد، احتمال وجود لوسمی لنفوبلاستیک حاد بسیار زیاد است. این روش‌های دستی و وابسته به متخصص نیز مستعد خطا هستند زیرا نتایج چنین روش‌هایی به شدت به دانش و مهارت متخصصی که تحلیل را انجام می‌دهد، بستگی دارد [۷]. برای اجتناب از پیچیدگی‌های مرتبط با چنین روش‌های تهاجمی و ارائه راه‌حل‌های سریع‌تر، ایمن‌تر و مقرون به صرفه‌تر، از تکنیک‌های مبتنی بر تحلیل تصاویر پزشکی استفاده می‌شود. روش‌های مبتنی بر پردازش تصویر و بینایی ماشین به راحتی قابل تعمیم هستند و خطای عامل انسانی را حذف می‌کنند.

تحلیل مبتنی بر تصویر به راحتی توسط رادیولوژیست‌ها قابل انجام است. چنین روش‌هایی می‌توانند به این صورت مفید باشند که غیرتهاجمی هستند، اما از همان مشکلات روش‌های تهاجمی رنج می‌برند. تحلیل دستی توسط

رادیولوژیست‌ها زمانی که متخصصان انسانی باید مجموعه داده‌های بزرگ شامل صدها و هزاران تصویر پزشکی را تحلیل کنند، پرزحمت، مستعد خطا و به شدت خسته‌کننده می‌شود. ویژگی‌های ذاتی همپوشان مانند مورفولوژی و بافت در تصاویر پزشکی، این وظیفه را دشوار می‌کند. این واقعیت در شکل ۱-۲، که تصاویر سلول‌های سرطانی و سلول‌های سالم را نشان می‌دهد، به وضوح نمایان است. می‌توان مشاهده کرد که تشخیص و طبقه‌بندی این لکوسیت‌ها به دلیل همگنی درون‌کلاسی بالا و جدایی‌پذیری بین‌کلاسی پایین دشوار است.



شکل ۱-۲ تعدادی از نمونه‌های سالم و لوسمی

در نتیجه، توسعه روش‌شناسی‌های دقیق و قابل اعتماد برای تشخیص لوسمی برای تشخیص به موقع و درمان زودهنگام مهم است. با این حال، هرچند به طور جزئی، سلول‌های سرطانی لوسمی لنفوبلاستیک حاد از سلول‌های سالم بر اساس عوامل مختلفی از جمله مورفولوژی، اندازه سلول، شکل و بافت متفاوت هستند [۷]. یک طبقه‌بندی‌کننده محاسباتی می‌تواند از هر یک از این ویژگی‌های متمایزکننده برای تشخیص تصاویر سلول‌های سرطانی و سلول‌های سالم استفاده کند.

الگوریتم‌های یادگیری ماشین (ML) ابزاری ایده‌آل برای مقابله با حجم زیادی از داده‌های پیچیده نشان داده‌اند و در نتیجه در درک بیماری‌ها مفید هستند. پزشکان به طور سنتی آزمایش‌های تشخیصی و داده‌های بیمار را بر اساس سال‌ها آموزش و تجربه پزشکی خود ارزیابی می‌کنند. با این حال، مطالعات اخیر نشان داده‌اند که الگوریتم‌های یادگیری ماشین در چندین وظیفه، از جمله تشخیص اولیه و پیش‌بینی مشکلات درمان و پیگیری عود در بدخیمی‌های خون‌شناسی، قابل مقایسه با تشخیص متخصصان هستند. در سال‌های اخیر، از ابزارهای یادگیری ماشینی (ML) برای تحلیل تصاویر آزمایشگاهی خون جهت تشخیص، تمایز و شمارش سلول‌ها در انواع مختلف لوسمی استفاده شده است. این مطالعات با هدف غلبه بر محدودیت‌های تشخیص دیرهنگام و بهبود شناسایی زیرگروه‌های لوسمی انجام می‌شوند [۵].

یادگیری عمیق شاخه‌ای شناخته شده از یادگیری ماشینی است که شامل الگوریتم‌ها و روابط ریاضی است. این فناوری به سرعت در تحقیقات بالینی ادغام شده و به کامپیوترها امکان می‌دهد بدون برنامه‌نویسی صریح از داده‌ها یاد بگیرند. ادغام فناوری‌های یادگیری عمیق در پردازش داده‌های پزشکی نتایج قابل توجهی به همراه داشته و در تشخیص بیماری مؤثر بوده است. تحقیقات نشان می‌دهد که رویکردهای یادگیری عمیق به طور قابل توجهی فرآیندهای پیچیده تصمیم‌گیری پزشکی را در پردازش تصاویر پزشکی با استخراج و ارزیابی ویژگی‌های تصویر بهبود می‌بخشند [۱].

دشواری در طبقه‌بندی تصاویر عاری از لوسمی و تصاویر مبتلا به لوسمی در شناسایی تفاوت‌های بصری ظریف بین این دو نوع تصویر نهفته است. تصاویر مبتلا به لوسمی ممکن است حاوی سلول‌های غیرطبیعی باشند که ممکن است شبیه سلول‌های سالم به نظر برسند و تمایز بین این دو را دشوار سازند. علاوه بر این، ظاهر سلول‌های غیرطبیعی ممکن است بسته به مرحله و نوع لوسمی بسیار متفاوت باشد. بنابراین، طبقه‌بندی دقیق این تصاویر نیازمند ترکیبی از تکنیک‌های پیشرفته تصویربرداری و الگوریتم‌های یادگیری ماشین است که بتوانند به طور مؤثر ویژگی‌های بصری این تصاویر را تجزیه و تحلیل و طبقه‌بندی کنند.

روش استفاده شده در این پژوهش شامل دو مدل می‌باشد که به صورت میانگین وزن دار باهم ترکیب می‌شوند. مدل اول مبتنی بر شبکه‌های کانوولوشنی و مدل دوم بر اساس ترانسفورمر بینایی است. از آنجا که استخراج ویژگی

به نحو متفاوتی در این دو مدل صورت می‌پذیرد، ترکیب دو مدل به نتیجه‌ای بهتر از هر یک از دو مدل پایه خواهد رسید.

این مطالعه در چند فصل شرح داده شده است که در ادامه به هر یک پرداخته می‌شود؛

- کارهای پیشین، بررسی مدل‌ها و شبکه‌هایی که مطالعات سالیان اخیر ساخته شده‌اند
- مقدمه‌ای بر شبکه‌های کانوولوشنی به ویژه شبکه EfficientNet، و ترانسفورمرهای بینایی که در نهایت در مورد روش پیشنهادی به طور گسترده بحث خواهد شد
- طراحی و به کار گیری مدل به کمک زبان برنامه نویسی پایتون
- نتایج مدل بر داده‌های از پیش دیده نشده
- جمع بندی

فصل دوم: کارهای پیشین

در این فصل تعدادی از مطالعات در زمینه استفاده از یادگیری ماشینی برای تشخیص لوسمی لنفوبلاستیک حاد طی سال‌های اخیر مورد بررسی قرار گرفته‌اند. مطالعه این پژوهش‌ها در تعیین قدم بعدی حائز اهمیت خواهد بود.

در این مطالعه [۶]، مولفین تغییراتی در وزن‌ها و پارامترهای مدل‌های ResNet50 و VGG16 برای آموزش روی مجموعه داده لوسمی لنفوبلاستیک حاد ایجاد کردند. علاوه بر این، آنها شش رویکرد متمایز یادگیری ماشین و یک شبکه کانولوشنی با ده لایه کانولوشن و یک لایه طبقه‌بندی پیشنهاد دادند. شبکه کانولوشنی دقتی ۸۲.۱ درصدی به دست آورد، در حالی که شبکه VGG16 به دقت ۸۴.۶ درصدی رسید. تکنیک یادگیری ماشین RF بهترین میزان دقت را با ۸۱.۷۲ درصد ارائه داد.

در [۷]، نویسندگان یک شبکه عصبی کانولوشنی مبتنی بر توجه را برای طبقه‌بندی سلول‌های لوسمی لنفوبلاستیک حاد پیشنهاد دادند. مدل آنها یک ساختار VGG16 را با یک ماژول Efficient Channel Attention ترکیب می‌کند تا نمایش ویژگی‌ها و دقت طبقه‌بندی را بهبود بخشد. نویسندگان مدل خود را از ابتدا روی مجموعه داده C-NMC 2019 که شامل بیش از ۱۰,۰۰۰ تصویر تک سلولی است، آموزش دادند. قابل توجه است که آنها مجموعه داده را بر اساس تغییرپذیری در سطح صحت سنجی به ۷ بخش تقسیم کردند، عاملی که اغلب در مطالعات قبلی نادیده گرفته شده بود. با استفاده از رویکرد cross-validation شش بخشی، روش آنها به میانگین دقت ۹۱.۱٪ دست یافت که از مدل ساده VGG16 بهتر عمل کرد و قابل مقایسه با سایر روش‌های پیشرفته بود.

این مقاله [۸]، یک رویکرد deep learning برای تشخیص لوسمی لنفوبلاستیک حاد (ALL) با استفاده از تصاویر میکروسکوپی اسمیر خون ارائه می‌دهد. نویسندگان دو مدل پیشنهاد می‌کنند: Multi-Attention EfficientNetV2S و Multi-Attention EfficientNetB3، که بر اساس معماری‌های از پیش آموزش دیده EfficientNet با مکانیسم‌های multi-attention اضافه شده هستند. آنها از مجموعه داده ISBI-2019 استفاده می‌کنند و تکنیک‌های data augmentation را برای رفع عدم تعادل کلاس‌ها به کار می‌گیرند. مدل‌ها به ترتیب دقت طبقه‌بندی بالای ۹۹.۷۳٪ و ۹۹.۲۵٪ را در تمایز بین سلول‌های نرمال و ALL به دست می‌آورند. نویسندگان نتایج خود را با سایر مدل‌های deep learning و ensemble در ادبیات موجود مقایسه می‌کنند و عملکرد برتر را نشان می‌دهند. آنها همچنین تحلیل Grad-CAM را برای نمایش مناطق تمرکز مدل انجام

می‌دهند. به طور کلی، این مطالعه پتانسیل transfer learning و مکانیسم‌های attention را در بهبود تشخیص ALL از تصاویر میکروسکوپی نشان می‌دهد.

در [۹] روشی مبتنی بر deep learning برای شناسایی لوسمی لنفوبلاستیک حاد و زیرگونه‌های آن با استفاده از تصاویر میکروسکوپی اسمیر خون پیشنهاد کردند. مجموعه داده مورد استفاده در این مطالعه شامل ۳۲۵۶ تصویر میکروسکوپی خون از ۸۹ فرد مشکوک به ALL است. این روش یک رویکرد کم هزینه برای segmentation سلول‌های لوسمی را شامل می‌شود و از جفت‌های تصاویر segmented و اصلی استفاده می‌کند. مدل شامل یک بلوک استخراج ویژگی مبتنی بر DenseNet-201 و یک بلوک طبقه‌بندی است. ویژگی‌های استخراج شده از تصاویر segmented و اصلی برای آموزش مدل DenseNet-201 جهت طبقه‌بندی به خوش‌خیم/بدخیم و زیرگونه‌های بدخیم ترکیب شدند. اگرچه معماری چند مرحله‌ای پیشنهادی، تحلیل جامع تصاویر میکروسکوپی اسمیر خون را برای تشخیص لوسمی لنفوبلاستیک حاد و طبقه‌بندی زیرگونه‌ها بهبود می‌بخشد، اما هزینه محاسباتی بالایی دارد و نیازمند دانش تخصصی است. این مدل دقت ۹۹.۸۵٪ را بدست می‌آورد.

در [۱۱] یک شبکه عصبی کانوولوشنی بهینه‌سازی شده بر اساس بیزین (Bayesian) برای تشخیص لوسمی لنفوبلاستیک حاد از تصاویر میکروسکوپی اسمیر پیشنهاد کردند. معماری شبکه کانوولوشنی و هایپرپارامترهای آن از طریق بهینه‌سازی Bayesian با داده‌های ورودی تطبیق داده می‌شود، تکنیکی که به صورت تکراری فضای هایپرپارامترها را برای به حداقل رساندن یک تابع خطای هدف اصلاح می‌کند. یک مجموعه داده هیبریدی با ترکیب دو مجموعه داده عمومی در دسترس برای آموزش و آزمایش شبکه عصبی کانوولوشنی بهینه‌سازی شده بیزین ایجاد شد نتایج تجربی نشان داد که مدل شبکه کانوولوشنی بهینه‌سازی شده بیزین عملکرد برتری در طبقه‌بندی لوسمی لنفوبلاستیک حاد از مجموعه آزمایش تصاویر اسمیر خون نشان می‌دهد و از سایر مدل‌های طبقه‌بندی لوسمی لنفوبلاستیک حاد با یادگیری عمیق بهینه‌سازی شده پیشی می‌گیرد. در حالی که این تحقیق اثربخشی مدل شبکه کانوولوشنی بهینه‌سازی شده مبتنی بر بیزین را در افزایش دقت تشخیص لوسمی لنفوبلاستیک حاد از تصاویر میکروسکوپی نشان داد، اما در دسته‌بندی زیرگونه‌های خاص لوسمی لنفوبلاستیک حاد کمکی نمی‌کند. از آنجا که این یک نشانه‌گذاری پذیرفته شده است و درمان بهینه و موثر به تشخیص دقیق نوع بیماری و میزان گسترش آن در بدن بستگی دارد.

مطالعه [۱۰] بر کاربرد تکنیک‌های یادگیری عمیق همراه با روش جمعی (ensemble) برای پیش‌بینی لوسمی لنفوبلاستیک حاد و شناسایی زیرگونه‌های آن با استفاده از تصاویر میکروسکوپی اسمیر خونی تمرکز دارد. در این مطالعه از مجموعه داده C-NMC-2019 برای ساخت مدل جمعی استفاده شده است. تکنیک oversampling برای حل مشکل عدم تعادل کلاس‌ها به کار گرفته شد که منجر به ایجاد یک مجموعه آموزشی با ۱۱۶۴۴ تصویر شد. شبکه‌های از پیش آموزش دیده به نام‌های VGG-16، Xception، MobileNetV2، InceptionResNet-V2 و DenseNet-121 با استفاده از یادگیری انتقالی مورد استفاده قرار گرفتند. مدل‌های جمعی عملکرد نسبتاً بالایی در شناسایی ALL نشان می‌دهد. نتایج تجربی نشان داد که روش پیشنهادی به ترتیب به دقت ۸۹.۷۲٪ دست یافت. یافته‌های کلی این تحقیق نشان داد که ensemble learning، با ترکیب قابلیت‌های شبکه‌های متنوع، اثربخشی کلی مدل را برای شناسایی لوسمی لنفوبلاستیک حاد در تصاویر پزشکی افزایش می‌دهد.

با توجه به پتانسیل ترانسفورمرهای بینایی و تکنیک‌های یادگیری جمعی در [۱۰]، ایده انتخابی در این پژوهش ترکیب مدل‌های مبتنی بر ترانسفورمرهای بینایی و شبکه‌های کانوولوشنی به روش جمعی (ensemble) انتخاب شده است. مزیت این روش نسبت به مطالعه [۱۰] این است که مدل‌های ترانسفورمر بینایی ویژگی‌های تصاویر را از روشی متفاوت از شبکه‌های کانوولوشنی استخراج می‌کند و ترکیب این دو، بدلیل نوع نگاه متفاوت آنها به تصاویر می‌تواند بسیار امیدوارکننده باشد.

فصل سوم: روش پیشنهادی

روش استفاده شده در این پژوهش برای تشخیص لوسمی لنفوبلاستیک حاد، به این صورت است که بعد از پردازش داده اولیه، شامل تغییر سایز و افزایش تعداد داده‌های آموزش، دو مدل آموزش داده می‌شود؛

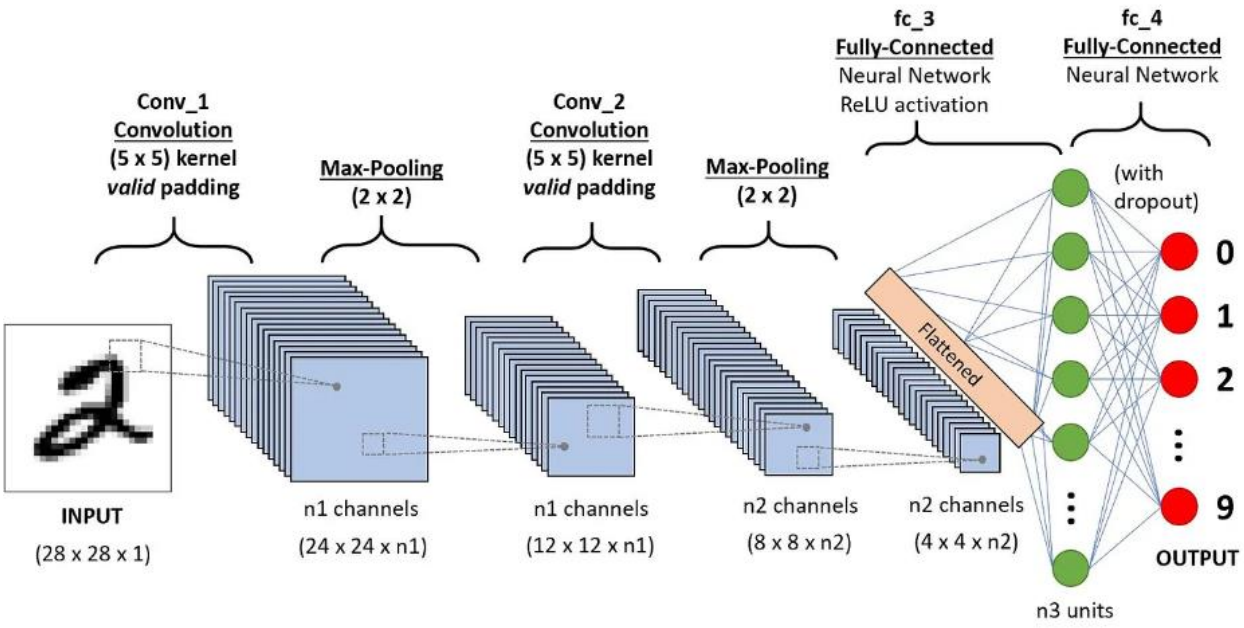
۱. مدل اول با استفاده از معماری EfficientNet و با استفاده از یادگیری انتقالی
۲. مدل دوم بر مبنای ترانسفورمر بینایی و با استفاده از یادگیری انتقالی

پس از آموزش این دو مدل ترکیب آنها را به عنوان مدل نهایی ViT-CNN در نظر خواهیم گرفت این مجموع به گونه‌ای میانگین وزن دار از هر دو مدل خواهد بود. با این ایده که به مدل دارای عملکرد و دقت بهتر وزن بیشتری نسبت به مدل دیگر اختصاص دهیم.

در ادامه به توضیح هر یک از دو مدل مطرح شده می‌پردازیم.

۳-۱ مروری بر شبکه‌های کانوولوشنی

شبکه‌های کانوولوشنی دسته‌ای از شبکه‌های عصبی عمیق برای پردازش داده‌هایی با الگوی شبکه‌ای مانند تصاویر هستند که از سازماندهی قشر بینایی حیوانات الهام گرفته شده‌اند و برای یادگیری خودکار و تطبیقی سلسله مراتب فضایی ویژگی‌ها، از الگوهای سطح پایین تا سطح بالا طراحی شده‌اند. شبکه کانوولوشنی یک ساختار ریاضی است که معمولاً از سه نوع لایه (یا بلوک‌های ساختمانی) تشکیل شده است: لایه‌های کانولوشن، پولینگ و کاملاً متصل. دو لایه اول، کانولوشن و پولینگ، استخراج ویژگی را انجام می‌دهند، در حالی که سومی، لایه کاملاً متصل، ویژگی‌های استخراج شده را به خروجی نهایی، مانند طبقه‌بندی، نگاشت می‌کند. شکل ۳-۱ نمایی کلی از یک شبکه کانوولوشنی را نشان می‌دهد.



شکل ۳-۱ شبکه عصبی کانوولوشنی

بیشتر مطالعات اخیر از تکنیک‌های استخراج ویژگی به صورت دستی، مانند تحلیل بافت، و به دنبال آن طبقه‌بندی‌کننده‌های یادگیری ماشین متعارف، مانند جنگل‌های تصادفی و ماشین‌های بردار پشتیبان استفاده می‌کنند. چندین تفاوت قابل توجه بین این روش‌ها و شبکه عصبی کانوولوشنی وجود دارد:

۱. شبکه عصبی کانوولوشنی نیازی به استخراج ویژگی به طور دستی ندارد.
۲. معماری‌های عصبی کانوولوشنی لزوماً به قطعه‌بندی تومورها یا اندام‌ها توسط متخصصان انسانی نیاز ندارند.
۳. شبکه کانوولوشنی بسیار بیشتر به داده نیاز دارد زیرا میلیون‌ها پارامتر قابل یادگیری برای تخمین دارد، و در نتیجه از نظر محاسباتی پرهزینه‌تر است، که منجر به نیاز به واحدهای پردازش گرافیکی (GPU) برای آموزش مدل می‌شود.

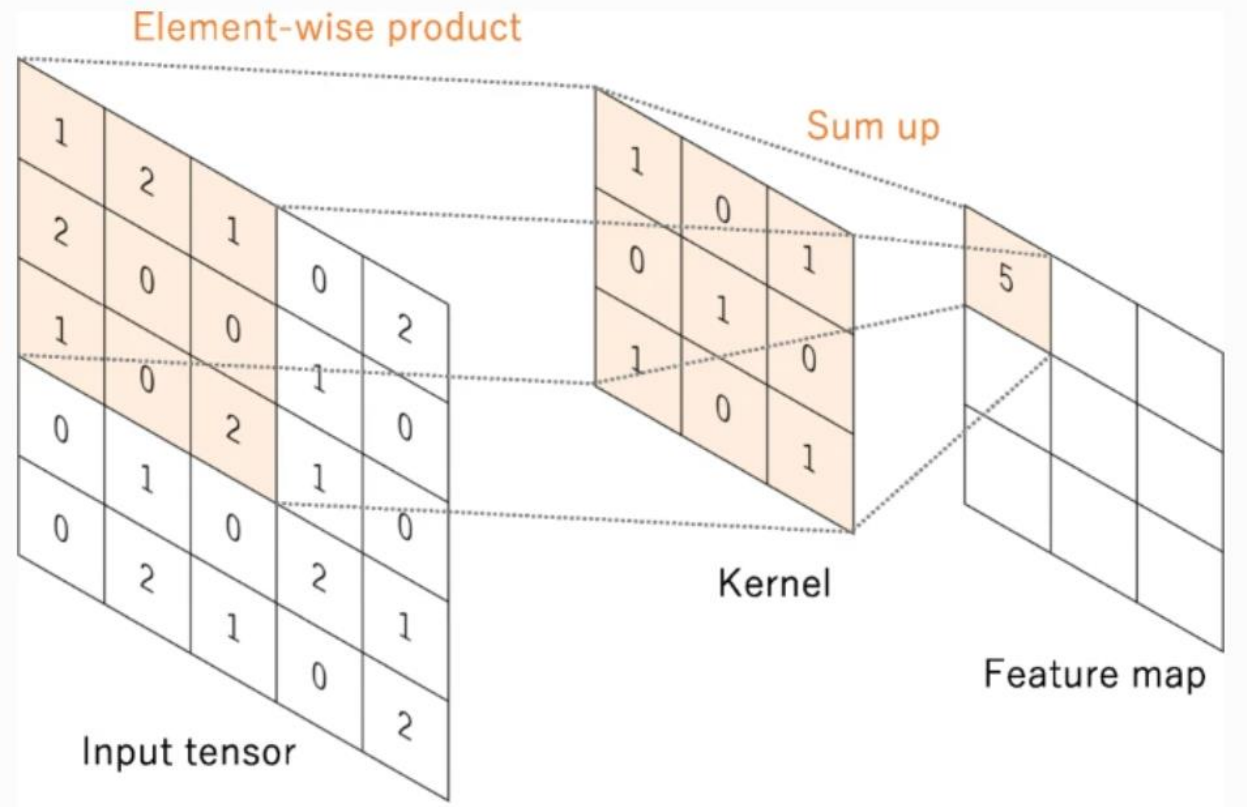
یک شبکه کانوولوشنی شامل چند بلوک اصلی است که در ادامه بررسی شده‌اند:

۳-۱-۱ لایه کانوولوشن

لایه کانولوشن یک جزء اساسی در معماری شبکه کانوولوشنی است که استخراج ویژگی را انجام می‌دهد. این لایه معمولاً از ترکیبی از عملیات خطی و غیرخطی تشکیل شده است، یعنی عملیات کانولوشن و تابع فعال‌سازی. کانولوشن نوعی عملیات خطی تخصصی است که برای استخراج ویژگی استفاده می‌شود. در این عملیات، یک آرایه کوچک از اعداد به نام کرنل، بر روی ورودی که یک آرایه از اعداد به نام تنسور است، اعمال می‌شود. در هر موقعیت تنسور، حاصل‌ضرب عنصر به عنصر بین هر عنصر کرنل و تنسور ورودی محاسبه و جمع می‌شود تا مقدار خروجی در موقعیت متناظر تنسور خروجی، که نقشه ویژگی نامیده می‌شود، به دست آید. شکل ۳-۲ عملیات کانوولوشن را نشان میدهد.

این روند با اعمال چندین کرنل تکرار می‌شود تا تعداد دلخواهی از نقشه‌های ویژگی تشکیل شود که ویژگی‌های مختلف تنسورهای ورودی را نمایش می‌دهند؛ بنابراین، کرنل‌های مختلف را می‌توان به عنوان استخراج‌کننده‌های ویژگی متفاوت در نظر گرفت.

دو ابرپارامتر کلیدی که عملیات کانولوشن را تعریف می‌کنند عبارتند از اندازه و تعداد کرنل‌ها. اندازه معمولاً ۳ × ۳ است، اما گاهی ۵ × ۵ یا ۷ × ۷ نیز استفاده می‌شود. تعداد کرنل‌ها اختیاری است و عمق نقشه‌های ویژگی خروجی را تعیین می‌کند.



شکل ۳-۲ نحوه عملکرد عملیات کانوولوشن

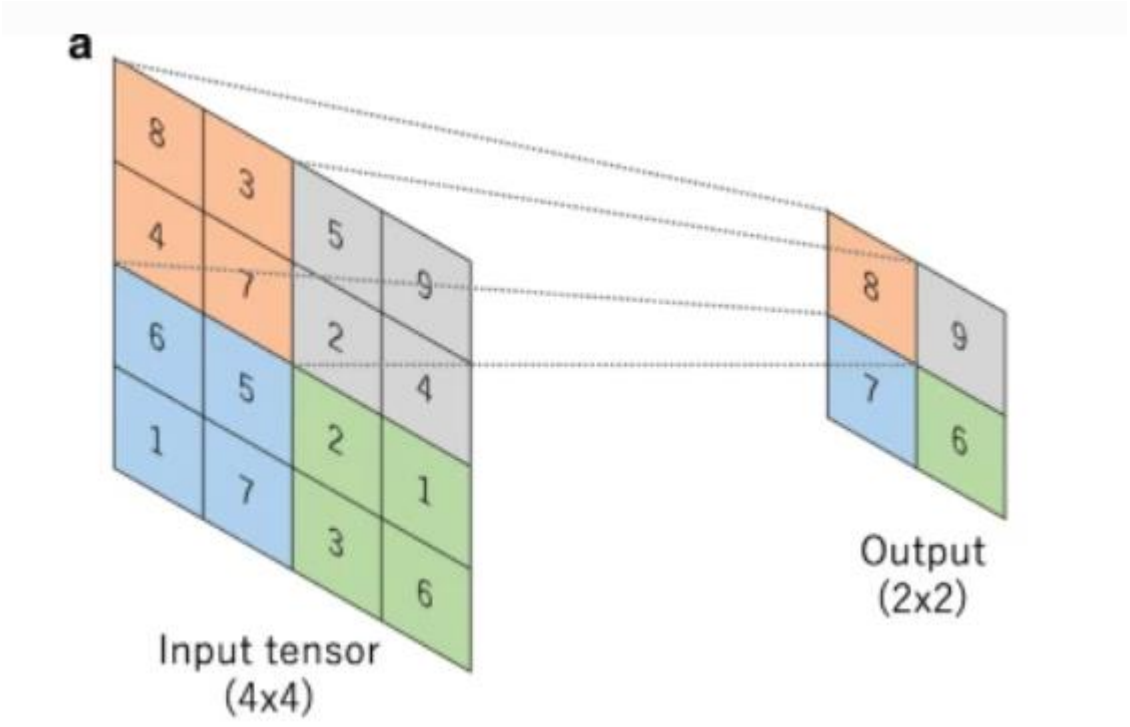
۳-۱-۲ لایه پولینگ

لایه پولینگ یک عملیات نمونه‌برداری کاهشی (downsampling) معمول را ارائه می‌دهد که ابعاد درون صفحه‌ای نقشه‌های ویژگی را کاهش می‌دهد. این کار به منظور ایجاد عدم حساسیت به جابجایی‌ها و اعوجاج‌های کوچک (translation invariance) و کاهش تعداد پارامترهای قابل یادگیری بعدی انجام می‌شود.

قابل توجه است که در هیچ یک از لایه‌های پولینگ، پارامتر قابل یادگیری وجود ندارد. در عوض، اندازه فیلتر، گام (stride) و پدینگ (padding) ابرپارامترهایی هستند که در عملیات پولینگ استفاده می‌شوند، مشابه عملیات کانولوشن.

رایج‌ترین شکل عملیات پولینگ، پولینگ حداکثر (max pooling) است که بخش‌هایی را از نقشه‌های ویژگی ورودی استخراج می‌کند، مقدار حداکثر در هر بخش را به عنوان خروجی می‌دهد و تمام مقادیر دیگر را حذف می‌کند.

در عمل، معمولاً از یک پولینگ حداکثر با فیلتری به اندازه ۲ × ۲ و گام 2 (stride) استفاده می‌شود. این کار ابعاد درون صفحه‌ای نقشه‌های ویژگی را با ضریب ۲ کاهش می‌دهد. برخلاف ارتفاع و عرض، بعد عمق نقشه‌های ویژگی بدون تغییر باقی می‌ماند. شکل ۳-۳ عملیات پولینگ حداکثر را به تصویر می‌کشد.



شکل ۳-۳ عملیات پولینگ

۳-۱-۳ لایه کاملا متصل

نقشه‌های ویژگی خروجی آخرین لایه کانولوشن یا پولینگ معمولاً مسطح می‌شوند، یعنی به یک آرایه یک بعدی از اعداد (یا بردار) تبدیل می‌شوند و به یک یا چند لایه کاملاً متصل، که به عنوان لایه‌های متراکم نیز شناخته می‌شوند، متصل می‌گردند. در این لایه‌ها، هر ورودی به هر خروجی توسط یک وزن قابل یادگیری متصل می‌شود.

پس از اینکه ویژگی‌های استخراج شده توسط لایه‌های کانولوشن و نمونه‌برداری کاهشی شده توسط لایه‌های پولینگ ایجاد شدند، آنها توسط زیرمجموعه‌ای از لایه‌های کاملاً متصل به خروجی‌های نهایی شبکه نگاشت می‌شوند. این خروجی‌ها می‌توانند، برای مثال، احتمالات برای هر کلاس طبقه‌بندی باشند.

آخرین لایه کاملاً متصل معمولاً به تعداد کلاس‌ها نورون خروجی دارد. هر لایه کاملاً متصل به دنبال یک تابع فعال ساز غیرخطی، مانند ReLU قرار می‌گیرد.

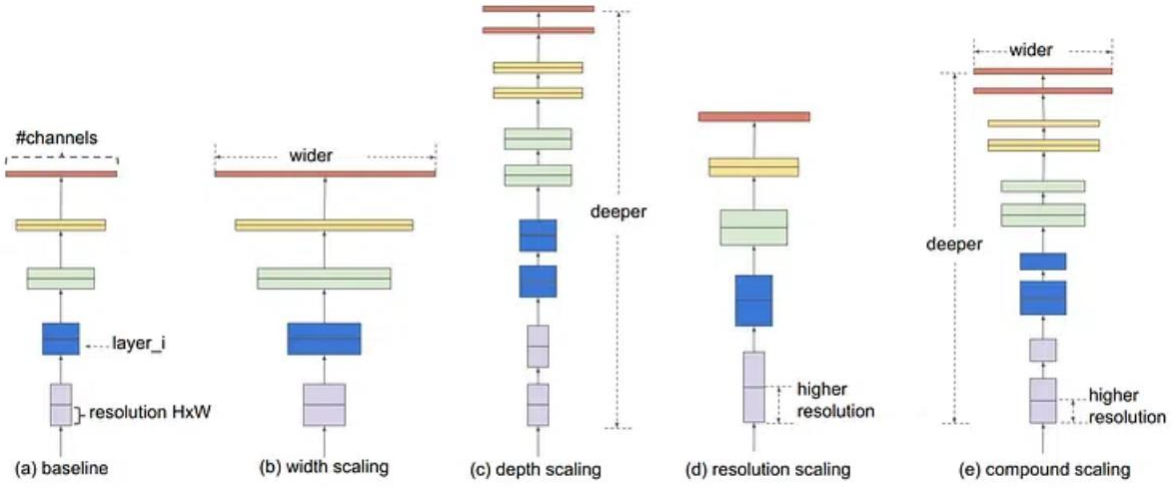
۳-۲ معماری EfficientNet

شبکه EfficientNet یک **شبکه** عصبی کانولوشنی است که بر اساس مفهومی به نام "مقیاس‌بندی ترکیبی" ساخته شده است. این مفهوم به مصالحه دیرینه بین اندازه مدل، دقت و کارایی محاسباتی می‌پردازد. ایده پشت مقیاس‌بندی ترکیبی، مقیاس‌بندی سه بعد اساسی یک شبکه عصبی است: عرض، عمق و وضوح.

۱. **عرض**: مقیاس‌بندی عرض به تعداد کانال‌ها در هر لایه شبکه عصبی اشاره دارد. با افزایش عرض، مدل می‌تواند الگوها و ویژگی‌های پیچیده‌تری را ثبت کند که منجر به بهبود دقت می‌شود. در مقابل، کاهش عرض منجر به مدلی سبک‌تر می‌شود که برای محیط‌های با منابع محدود مناسب است.

۲. **عمق**: مقیاس‌بندی عمق به تعداد کل لایه‌ها در شبکه مربوط می‌شود. مدل‌های عمیق‌تر می‌توانند نمایش‌های پیچیده‌تری از داده‌ها را ثبت کنند، اما آنها همچنین به منابع محاسباتی بیشتری نیاز دارند. از طرف دیگر، مدل‌های کم‌عمق‌تر از نظر محاسباتی کارآمدتر هستند اما ممکن است دقت را فدا کنند.

۳. **وضوح**: مقیاس‌بندی وضوح شامل تنظیم اندازه تصویر ورودی است. تصاویر با وضوح بالاتر اطلاعات جزئی‌تری ارائه می‌دهند که می‌تواند منجر به عملکرد بهتر شود. با این حال، آنها همچنین به حافظه و قدرت محاسباتی بیشتری نیاز دارند. از طرف دیگر، تصاویر با وضوح پایین‌تر منابع کمتری مصرف می‌کنند اما ممکن است منجر به از دست دادن جزئیات ظریف شوند.

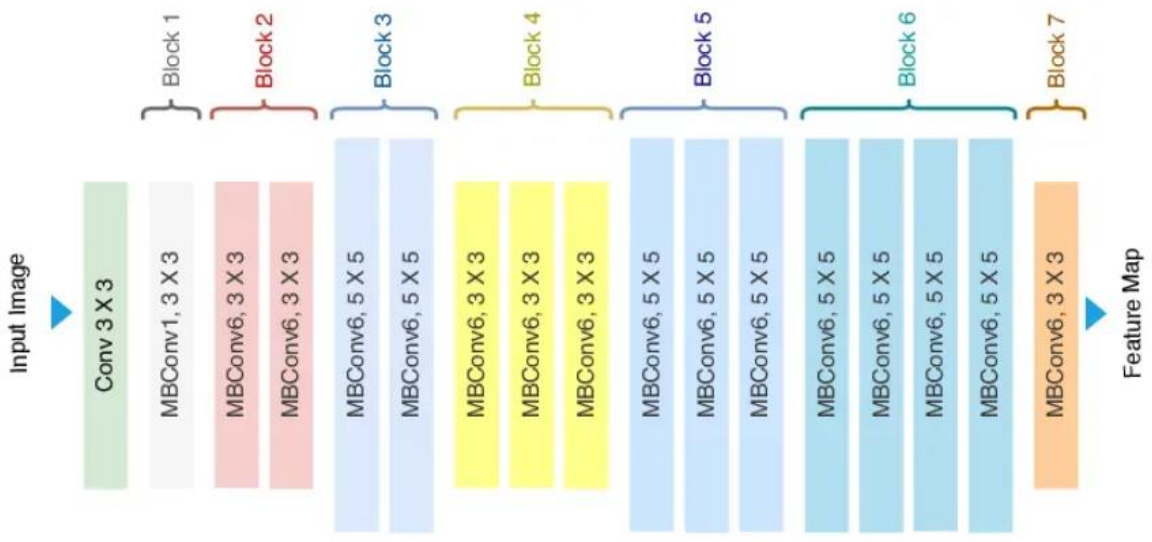


شکل ۳-۴ مقیاس بندی سه بعد اصلی و مقیاس بندی ترکیبی

یکی از نقاط قوت EfficientNet در توانایی آن برای متعادل کردن این سه بعد از طریق یک رویکرد اصولی است. محققان با شروع از یک مدل پایه، یک جستجوی شبکه‌ای سیستماتیک انجام می‌دهند تا ترکیب بهینه عرض، عمق و وضوح را پیدا کنند. این جستجو توسط یک ضریب ترکیبی، که با ϕ نشان داده می‌شود، هدایت می‌شود که ابعاد مدل را به طور یکنواخت مقیاس می‌کند. این مقدار ϕ به عنوان یک پارامتر تعریف شده توسط کاربر عمل می‌کند که پیچیدگی کلی مدل و نیازهای منابع آن را تعیین می‌کند.

۳-۲-۱ بلوک‌های EfficientNet

این شبکه از لایه‌های Mobile Inverted Bottleneck یا (MBConv) استفاده می‌کند که ترکیبی از کانولوشن‌های عمق‌پذیر جداپذیر و بلوک‌های باقیمانده معکوس هستند. علاوه بر این، معماری مدل از بهینه‌سازی Squeeze-and-Excitation، (SE) برای بهبود بیشتر عملکرد مدل استفاده می‌کند. شکل ۳-۵ نمایی از معماری EfficientNet را نشان می‌دهد.



شکل ۳-۵ بلوک‌های کلی شبکه EfficientNet

فرآیند با یک **مدل پایه** شروع می‌شود که به عنوان نقطه شروع عمل می‌کند. این مدل پایه معمولاً یک شبکه عصبی با اندازه مناسب است که در یک وظیفه خاص عملکرد خوبی دارد، اما ممکن است برای کارایی محاسباتی بهینه نشده باشد.

سپس، یک ضریب ترکیبی به عنوان پارامتر تعریف شده توسط کاربر معرفی می‌شود که تعیین می‌کند ابعاد شبکه عصبی چقدر مقیاس‌بندی شوند. این یک مقدار اسکالر واحد است که به طور یکنواخت عرض، عمق و وضوح مدل را مقیاس می‌کند. با تنظیم این مقدار ϕ پیچیدگی کلی و نیازهای منابع مدل می‌تواند کنترل شود.

از اینجا، ابعاد مقیاس‌بندی می‌شوند. ایده اصلی پشت مقیاس‌بندی ترکیبی، مقیاس‌بندی ابعاد مدل پایه (عرض، عمق و وضوح) به صورت متعادل و هماهنگ است. ضرایب مقیاس‌بندی برای هر بعد از ضریب ترکیبی ϕ مشتق می‌شوند.

- مقیاس‌بندی عرض :عرض شبکه عصبی به طور متناسب با به توان رساندن ϕ به یک نمای خاص معمولاً با α نشان داده می‌شود مقیاس می‌شود.
- مقیاس‌بندی عمق :به طور مشابه، عمق شبکه با به توان رساندن ϕ به نمای دیگری معمولاً با β نشان داده می‌شود مقیاس می‌شود.
- مقیاس‌بندی وضوح :وضوح یا اندازه تصویر ورودی با ضرب وضوح اصلی (r) در ϕ به توان نمای متفاوت معمولاً با γ نشان داده می‌شود مقیاس می‌شود.

$$\text{Depth:} \quad d = \alpha^{\phi}$$

$$\text{Width:} \quad w = \beta^{\phi}$$

$$\text{Resolution:} \quad r = \gamma^{\phi}$$

$$\text{s.t.} \quad \alpha \cdot \beta^2 \cdot \gamma^2 \simeq 2, \quad \alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

$$\alpha = 1.2, \beta = 1.1, \gamma = 1.15,$$

در ادامه، باید **نماهای بهینه** تعیین شوند. نماهای α، β، γ ثابت‌هایی هستند که باید برای دستیابی به بهترین مقیاس‌بندی تعیین شوند. مقادیر این نماها معمولاً از طریق یک جستجوی شبکه‌ای تجربی یا فرآیند بهینه‌سازی

به دست می‌آیند. هدف شناسایی ترکیبی از نماها است که به بهترین تعادل بین دقت مدل و کارایی محاسباتی منجر شود.

پس از تعیین ضرایب مقیاس‌بندی برای عرض، عمق و وضوح، آنها به طور متناسب به مدل پایه اعمال می‌شوند. مدل حاصل اکنون EfficientNet با یک مقدار ϕ خاص است.

بسته به مورد استفاده خاص و منابع محاسباتی موجود، محققان و متخصصان می‌توانند از طیفی از مدل‌های EfficientNet انتخاب کنند، که هر کدام مربوط به یک مقدار ϕ متفاوت است. مقادیر ϕ کوچکتر منجر به مدل‌های سبک‌تر و کارآمدتر از نظر منابع می‌شوند، در حالی که مقادیر ϕ بزرگتر منجر به مدل‌های قدرتمندتر اما از نظر محاسباتی پیچیده‌تر می‌شوند.

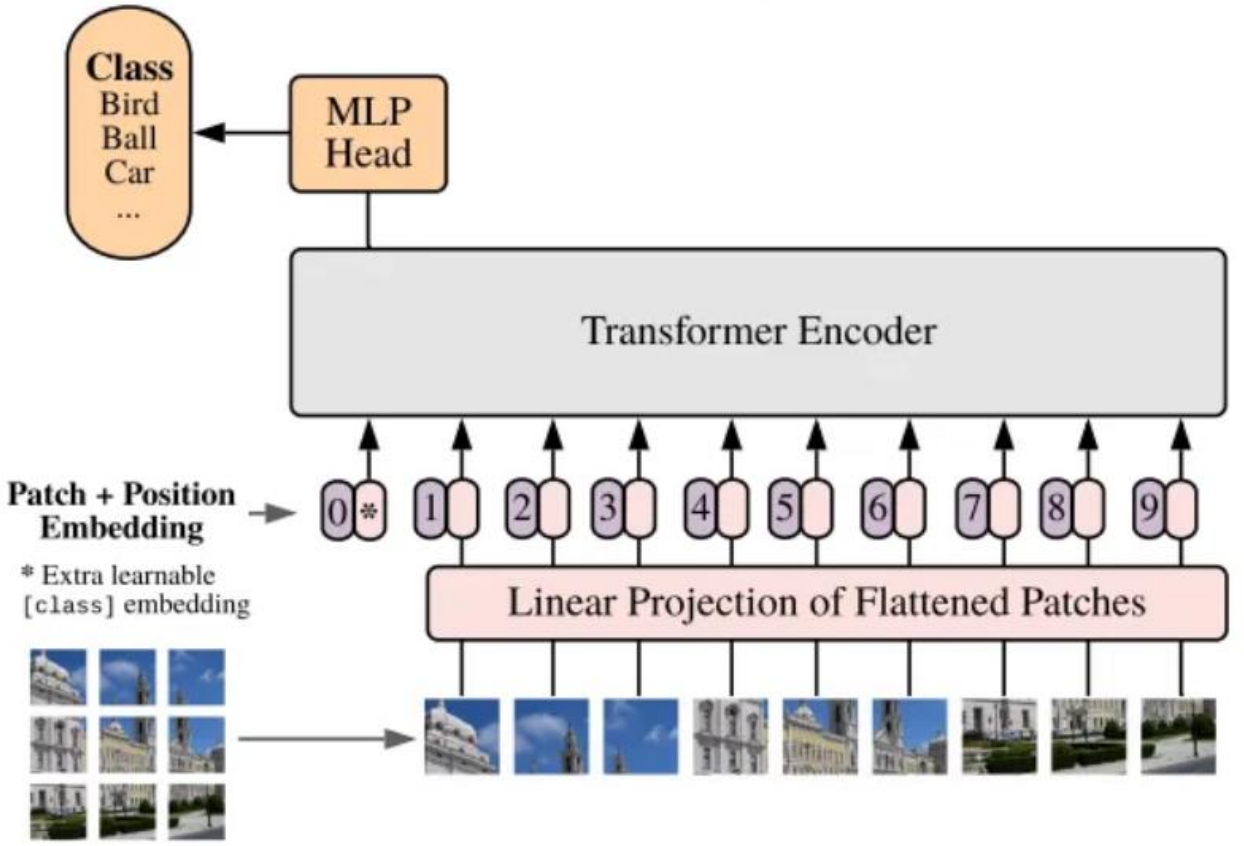
۳-۳ ترانسفورمر بینایی

ترانسفورمر بینایی (Vision Transformer) یک معماری شبکه عصبی انقلابی است که نحوه پردازش و درک تصاویر را از نو تصور می‌کند. مدل ViT در سال ۲۰۲۱ در یک مقاله تحقیقاتی توسط گوگل معرفی شد. ViT با الهام از موفقیت ترانسفورمرها در پردازش زبان طبیعی، روش جدیدی را برای تحلیل تصاویر معرفی می‌کند که با تقسیم آنها به بخش‌های کوچکتر و استفاده از مکانیسم‌های توجه خودکار (self-attention) انجام می‌شود. این به مدل اجازه می‌دهد تا روابط محلی و جهانی درون تصاویر را ثبت کند، که منجر به عملکرد چشمگیر در وظایف مختلف بینایی کامپیوتری می‌شود.

ترانسفورمر بینایی در چندین جنبه کلیدی با شبکه‌های عصبی کانولوشنی متفاوت است:

- نمایش ورودی: در حالی که CNN مستقیماً مقادیر پیکسل خام را پردازش می‌کند، ViT تصویر ورودی را به بخش‌هایی تقسیم کرده و آنها را به توکن‌ها تبدیل می‌کند.
- مکانیسم پردازش: CNN از لایه‌های کانولوشن و پولینگ برای ثبت سلسله مراتبی ویژگی‌ها در مقیاس‌های فضایی مختلف استفاده می‌کند ViT .از مکانیسم‌های توجه خودکار برای در نظر گرفتن روابط بین تمام بخش‌ها استفاده می‌کند.
- درک کلی از تصویر: ViT به طور ذاتی درک کلی را از طریق توجه خودکار ثبت می‌کند، که به شناسایی روابط بین بخش‌های دور از هم کمک می‌کند. CNN برای اطلاعات گلوبال به لایه‌های پولینگ متکی است.
- کارایی داده: CNNها اغلب به مقادیر زیادی داده برچسب‌گذاری شده برای آموزش نیاز دارند، در حالی که ViT می‌تواند از پیش‌آموزش روی مجموعه داده‌های بزرگ و سپس تنظیم دقیق برای وظایف خاص بهره ببرد.

ترانسفورمرهای بینایی چگونه کار می‌کنند؟

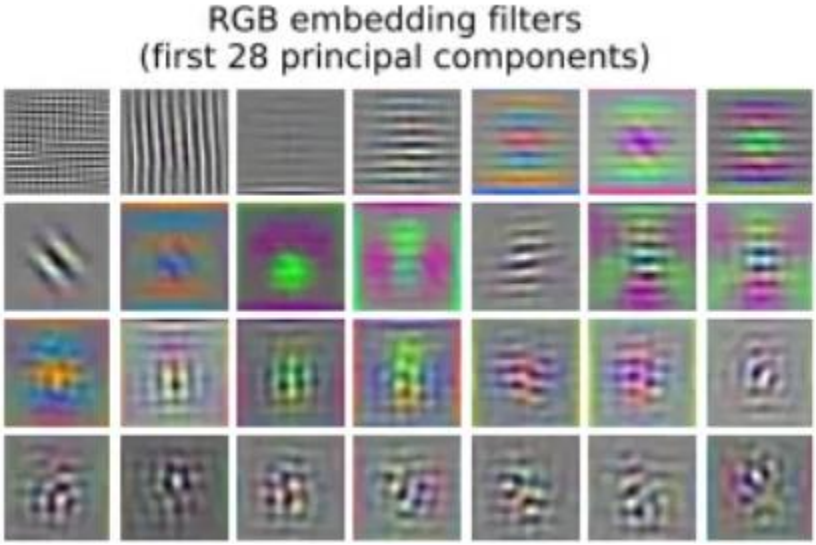


شکل ۳-۶ ترانسفورمر بینایی

عملکرد ViT را می‌توان به چندین مرحله تقسیم کرد که هر کدام نقش مهمی در عملکرد کلی آن ایفا می‌کنند:

۱. جاسازی پچ Patch Embedding:

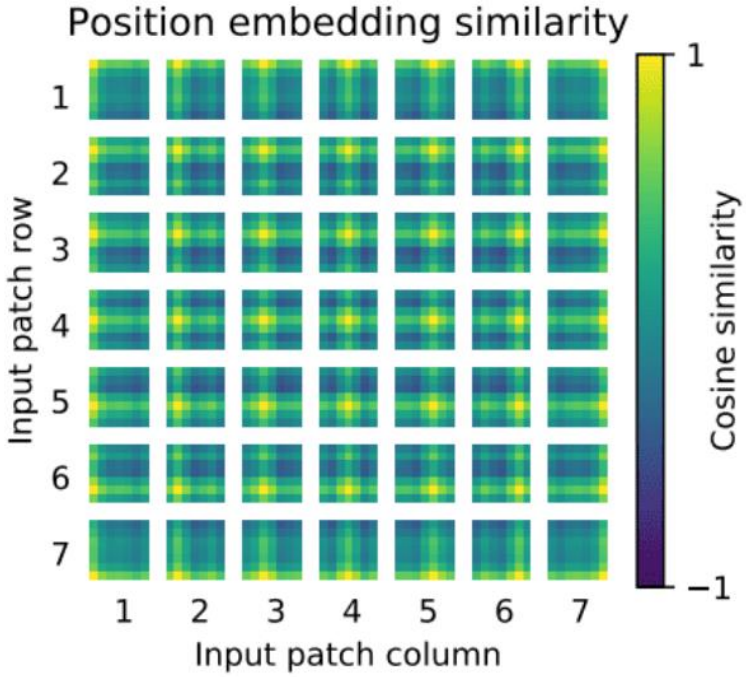
- تصویر ورودی به بخش‌های مربعی با اندازه ثابت تقسیم می‌شود. سپس هر بخش با استفاده از یک عملیات تصویر خطی قابل یادگیری به یک بردار تبدیل می‌شود. این منجر به یک توالی از پچ‌های جاسازی شده می‌شود که به عنوان توکن‌های ورودی برای لایه‌های بعدی عمل می‌کنند.



شکل ۳-۷ جاسازی پچ

۲. جاسازی موقعیت Positional Embedding:

- از آنجا که ViT فاقد هرگونه درک ذاتی از روابط فضایی است، اطلاعات مربوط به موقعیت هر پچ باید به صورت صریح ارائه شود. این کار با افزودن کدگذاری‌های موقعیتی به جاسازی‌های پچ انجام می‌شود.
- کدگذاری‌های موقعیتی به مدل کمک می‌کنند تا بین موقعیت‌های مختلف در تصویر تمایز قائل شود و روابط بین آنها را ثبت کند. آنها معمولاً یاد گرفته می‌شوند و در مرحله ورودی به جاسازی‌های پچ اضافه می‌شوند.



شکل ۳-۸ جاسازی موقعیت

۳. لایه‌های رمزگذارEncoder Layers:

- هسته اصلی ViT از چندین لایه رمزگذار تشکیل شده است که هر کدام شامل دو زیرلایه اصلی هستند: توجه خودکار چندسری (multi-head self-attention) و شبکه‌های عصبی پیش‌رونده (feedforward neural networks).

۴. توجه خودکار چندسری multi-head self-attention

- مکانیسم توجه خودکار روابط بین بخش‌های مختلف در توالی ورودی را ثبت می‌کند.
- برای هر جاسازی پچ، توجه خودکار یک مجموع وزنی از تمام جاسازی‌های بخش را محاسبه می‌کند، که در آن وزن‌ها بر اساس ارتباط هر بخش با بخش فعلی تعیین می‌شوند.
- این مکانیسم به مدل اجازه می‌دهد تا بر بخش‌های مهم تمرکز کند و در عین حال زمینه‌های لوکال و گلوبال را در نظر بگیرد.

- توجه چندسری از چندین مجموعه پارامتر قابل یادگیری (سرهای توجه) برای ثبت انواع مختلف روابط استفاده می‌کند.

شکل ۳-۹ نحوه تمرکز مکانیزم توجه

۵. شبکه‌های عصبی پیش‌رونده feedforward neural networks:

- پس از توجه خودکار، خروجی از مکانیسم توجه خودکار هر بخش از طریق یک شبکه عصبی پیش‌رونده عبور داده می‌شود.
- این شبکه معمولاً از یک لایه کاملاً متصل به همراه یک تابع فعال‌سازی مانند ReLU تشکیل شده است.
- هدف از شبکه پیش‌رونده، معرفی غیرخطی و اجازه دادن به مدل برای یادگیری روابط پیچیده بین بخش‌ها است.

۶. لایه نرمال‌سازی (Layer Normalization and Residual Connections):

- هم خروجی مکانیسم توجه خودکار و هم خروجی شبکه پیش‌خور با نرمال‌سازی لایه و اتصالات Residual دنبال می‌شوند.
- نرمال‌سازی لایه با نرمال کردن ورودی‌ها به هر زیرلایه، به تثبیت و تسریع آموزش کمک می‌کند.
- اتصالات Residual، که به عنوان اتصالات میان‌بر نیز شناخته می‌شوند، جاسازی‌های ورودی اصلی را به خروجی هر زیرلایه اضافه می‌کنند. این به جریان گرادیان‌ها در طول آموزش کمک می‌کند و از مشکل رو به صفر میل گردن گرادیان جلوگیری می‌کند.

۳-۳-۱ کاربردهای ViT

۱. طبقه‌بندی تصویر: دسته‌بندی تصاویر به دسته‌های از پیش تعریف شده

۲. تشخیص اشیاء: شناسایی و تعیین موقعیت اشیاء در داخل تصاویر

۳. قطعه‌بندی معنایی: اختصاص یک برچسب به هر پیکسل در تصویر برای شناسایی مرزهای اشیاء

۴. تولید تصویر: ایجاد تصاویر جدید بر اساس یک زمینه یا توصیف داده شده

۳-۳-۲ محدودیت‌های ViT

۱. مجموعه داده‌های بزرگ: آموزش موثر Vision Transformers اغلب نیاز به مجموعه داده‌های بزرگ دارد که ممکن است برای همه حوزه‌ها در دسترس نباشد.

۲. نیازهای محاسباتی: آموزش ViT به دلیل مکانیسم‌های توجه خودکار می‌تواند از نظر محاسباتی پرهزینه باشد.

۳. اطلاعات فضایی: پردازش ترتیبی ViT ممکن است الگوهای فضایی دقیق را به اندازه CNN ها برای وظایفی مانند قطعه‌بندی به طور موثر ثبت نکند.

فصل چهارم: طراحی و به کارگیری مدل

۴-۱ دیتاست

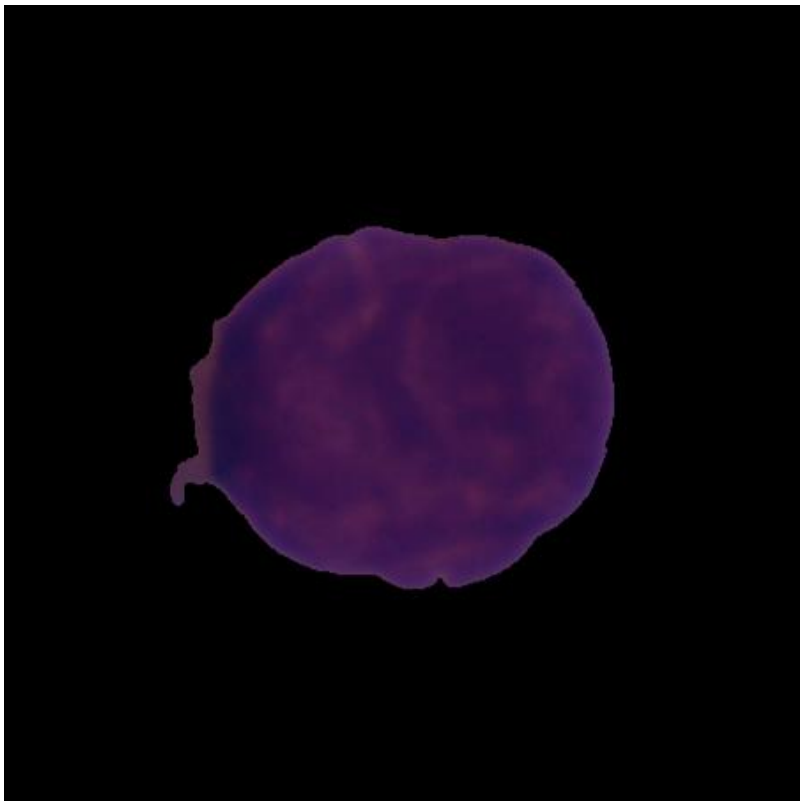
مجموعه داده مورد استفاده برای ساخت مدل تشخیصی، مجموعه داده C-NMC-2019 است، و ۱۰۶۶۱ تصویر سلول‌های سفید از ۷۳ فرد انتخاب شده است، شامل ۷۲۷۲ تصویر از سلول‌های B-lymphoblast لوسمی (سلول‌های سرطانی) از ۴۷ بیمار دارای لوسمی و ۳۳۸۹ تصویر از B-lymphoid (سلول‌های نرمال) از ۲۶ فرد سالم. این سلول‌ها از تصاویر میکروسکوپی جدا شده‌اند و هر تصویر سلولی یک تصویر واقعی پس از جمع‌آوری است. برخی نویزهای رنگ‌آمیزی و خطاهای روشنایی ایجاد شده در طول فرآیند جمع‌آوری تا حد زیادی اصلاح شده‌اند. همانطور که در شکل ۱-۲ نشان داده شد، مورفولوژی دو سلول بسیار مشابه است، بنابراین یک متخصص برچسب تصویر را حاشیه‌نویسی خواهد کرد. برچسب‌های تصاویر سلول‌های نرمال نمونه‌های مثبت هستند و برچسب‌های تصاویر سلول‌های سرطانی نمونه‌های منفی هستند.

تصاویر در ابعاد ۴۵۰× ۴۵۰ هستند با توجه شکل ۴-۱، می‌توان دید که در درصد بسیار بالایی از نمونه‌ها بخش قابل توجهی از حاشیه تصویر فاقد اطلاعات مفید بوده و کاملا سیاه می‌باشد. از آنجا که ابعاد ورودی یک شبکه یادگیری عمیق بشدت در سرعت آموزش و حجم محاسبات تاثیر گذار است، کاهش ابعاد بسیار کمک کننده خواهد بود. این کاهش ابعاد در دو مرحله انجام میپذیرد؛

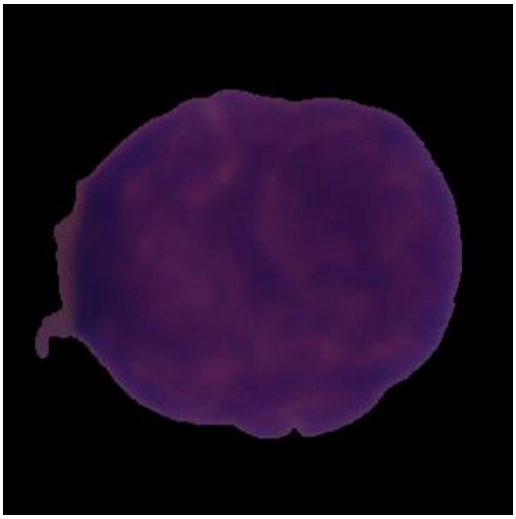
۱. ابتدا حاشیه‌های بدون اطلاعات تصاویر را به کمک تابع crop در کتابخانه PIL برش میزنیم. در انتهای این مرحله تصاویر به ابعاد ۳۰۰× ۳۰۰ خواهند رسید.

۲. با توجه به اینکه هر دو شبکه مورد نظر که در مراحل بعدی از آنها برای اموزش مدل استفاده خواهیم کرد انتظار ورودی در ابعاد ۲۲۴× ۲۲۴ را دارند، باید تغییرات لازم را انجام دهیم. این تغییرات به کمک تابع resize در کتابخانه PIL انجام می‌شوند. در پایان این مرحله دیتاست قابلیت کار با هر دو مدل EfficientNet و Vision Transformer را خواهد داشت.

شکل ۴-۱ و ۴-۲ نشان دهنده تغییرات اعمال شده بر یکی از تصاویر مجموعه می‌باشند.



شکل ۴-۱ تصویر قبل از هر گونه پیش پردازش

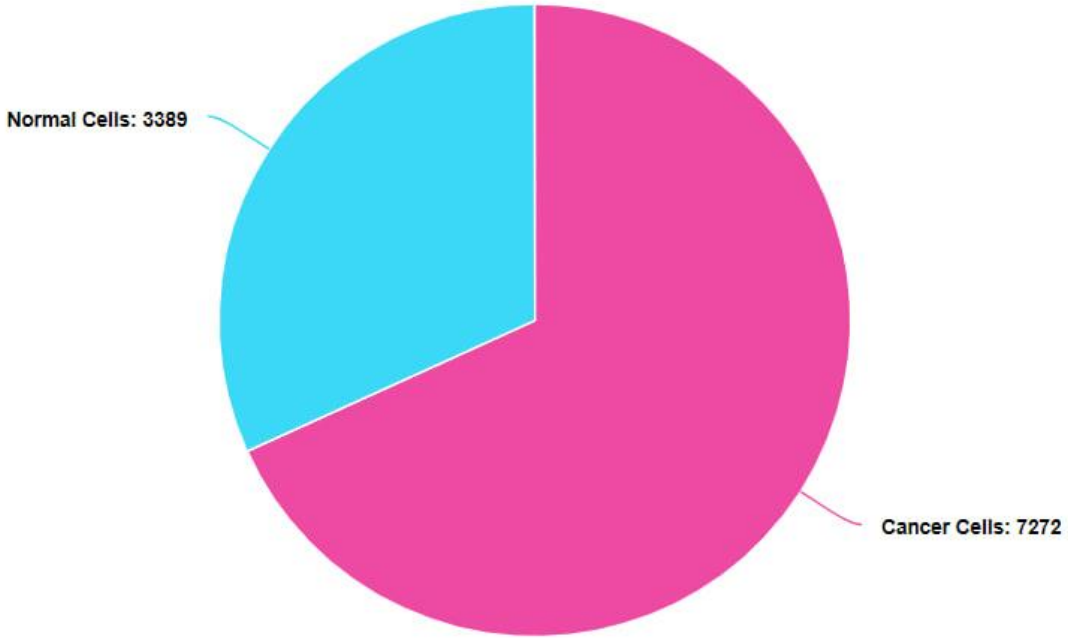


شکل ۴-۲ تصویر بعد از مرحله پیش پردازش

۴-۱-۱ متوازن سازی دیتاست

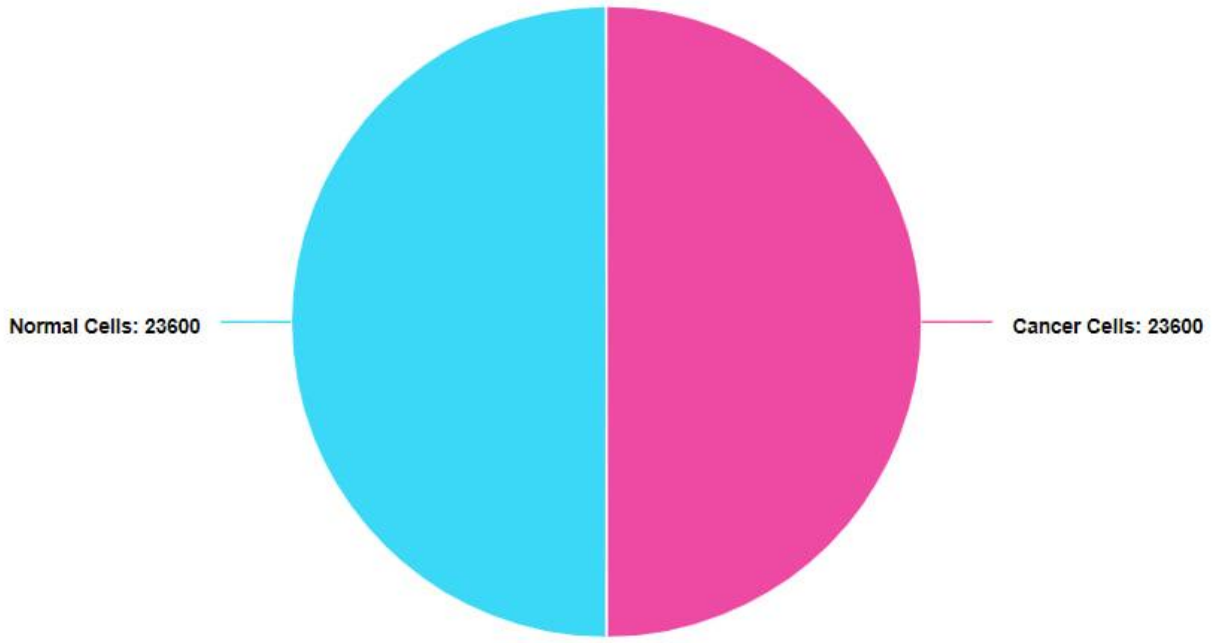
آموزش مستقیم بر روی یک مجموعه داده نامتوازن می‌تواند به راحتی باعث شود که مدل دچار overfitting شود یا قابلیت تعمیم مدل ضعیف شود. برای حل مشکل عدم تعادل داده‌ها، یک روش افزایش داده مبتنی بر difference enhancement-random sampling (به اختصار DERS) پیشنهاد می‌کنیم. فرض کنید مجموعه داده نامتوازن D، شامل a تصویر از دسته A و b تصویر از دسته B است، که در آن b بزرگتر از a است. فرض کنید N نوع افزایش داده بر روی دسته A و M نوع افزایش داده بر روی دسته B انجام می‌شود به طوری که تعداد a×N و b×M نسبتاً نزدیک باشند. سپس L تصویر از a×N تصویر دسته A و L تصویر از b×M تصویر دسته B انتخاب می‌شود، بنابراین اطمینان حاصل می‌شود که تعداد دسته A و تعداد دسته B در مجموعه داده جدید یکسان است، به طوری که مجموعه داده جدید به یک مجموعه داده متوازن تبدیل می‌شود.

شکل ۴-۳ تعداد سلول‌های نرمال و سرطانی را در مجموعه داده نشان می‌دهد. می‌توان دید که این مجموعه داده، یک مجموعه داده نامتوازن است. برای مجموعه داده مورد استفاده در این پژوهش، دو دسته از تصاویر سلولی وجود دارد که تصاویر سلول‌های نرمال و سرطانی هستند. تعداد تصاویر سلول‌های سرطانی بیش از دو برابر تعداد تصاویر سلول‌های نرمال است.



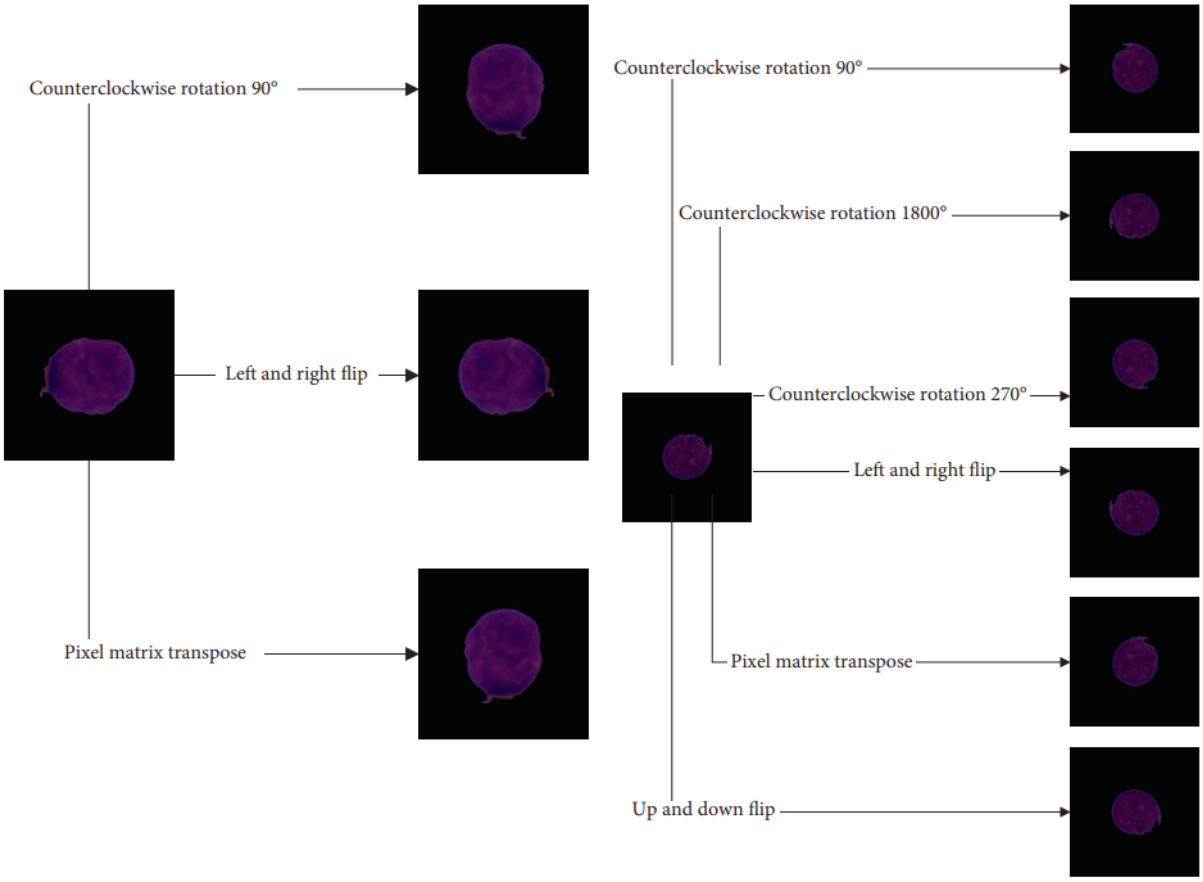
شکل ۴-۳ تعداد داده‌ها در هر کلاس

ما از روش DERS برای پردازش مجموعه داده استفاده می‌کنیم. سه روش افزایش داده شامل برگرداندن در جهت چپ و راست، چرخش ۹۰ درجه خلاف عقربه‌های ساعت، و ماتریس ترانهاده برای تولید تصاویر جدید سلول‌های سرطانی استفاده می‌شود. شش روش افزایش داده شامل برگرداندن در جهت چپ و راست، برگرداندن در جهت بالا و پایین ، چرخش ۹۰ درجه خلاف عقربه‌های ساعت، چرخش ۱۸۰ درجه خلاف عقربه‌های ساعت، چرخش ۲۷۰ درجه خلاف عقربه‌های ساعت، و ماتریس ترانهاده برای تولید تصاویر جدید سلول‌های نرمال استفاده می‌شود. تعداد تصاویر سلول‌های سرطانی ۲۹,۰۸۸ و تعداد تصاویر سلول‌های نرمال ۲۳,۷۰۲ بود. یک مجموعه داده جدید با ۲۳,۶۰۰ تصویر به طور تصادفی از دو تصویر تازه تولید شده ایجاد می‌شود. مجموعه داده جدید یک مجموعه داده کاملاً متوازن است. قبل از آموزش مدل، تصاویر نیاز به پیش‌پردازش دارند. اندازه اصلی تصاویر سلول‌ها ۴۵۰ × ۴۵۰ است، اندازه تصاویر سلول‌ها به ۲۲۴ × ۲۲۴ تنظیم می‌شود، و تصویر برای جلوگیری از overfitting مدل نرمال‌سازی می‌شود.



شکل ۴-۴ تعداد داده‌ها در هر کلاس پس از افزایش داده به روش DERS

شکل ۴-۵، روش‌های استفاده شده برای افزایش تعداد داده‌ها در هر کلاس را نشان می‌دهد؛



شکل ۴-۵ روش‌های استفاده شده برای افزایش تعداد داده‌ها در هر کلاس

در مرحله آخر داده‌ها به سه قسمت آموزش، سحت سنجی و تست تقسیم بندی می‌شوند به گونه‌ای که ۸۰ درصد به مجموعه آموزش، ۱۰ درصد به مجموعه صحت سنجی و ۱۰ درصد باقی مانده نیز به مجموعه تست اختصاص داده شوند.

۴-۲ مدل اول با استفاده از معماری EfficientNet

برای آموزش مدل از مدل موجود در کتابخانه tensorflow استفاده خواهیم کرد برای استفاده از این ماژول باید دیتاست در یک پوشه قرار بگیرد به گونه‌ای که آن پوشه حاوی سه پوشه، هر کدام برای یک قسمت از داده‌ها (داده‌های آموزش، صحت سنجی، و تست) باشد. هر یک از این پوشه‌ها نیز شامل دو پوشه متناسب با دو کلاس لوسمی یا سالم خواهد بود.

انواع مدل‌های EfficientNet: این معماری شامل ورژن‌های مختلف با تعداد پارامترهای متفاوت است، که در جدول ۴-۱ قابل مشاهده است.

جدول ۴-۱

Model	Params (Million)	Input Resolution
EfficientNet-B0	5.3	224 x 224
EfficientNet-B1	7.8	240 x 240
EfficientNet-B2	9.2	260 x 260
EfficientNet-B3	12	300 x 300
EfficientNet-B4	19	380 x 380
EfficientNet-B5	30	456 x 456
EfficientNet-B6	43	528 x 528
EfficientNet-B7	66	600 x 600

در این جدول، مدل‌ها به ترتیب پیچیدگی مرتب شده‌اند. مدل B0 ساده ترین مدل با ۵/۳ میلیون پارامتر می‌باشد درحالی که پیچیده ترین مدل دارای ۶۶ میلیون پارامتر است. قابل انتظار است که مدل B0 با کمترین مقدار پارامتر پیچیدگی محاسباتی کمتری خواهد داشت و در زمان آموزش سریع تر عمل خواهد کرد. با توجه به اینکه تعداد داده‌های دیتاست استفاده شده حتی پس از تکنیک‌های افزایش داده نیز به اندازه‌ای نیست که بتوان مدل‌های خیلی پیچیده را به خوبی آموزش داد و همینطور آموزش این مدل‌ها با سرعت بسیار کمتری انجام می‌شود، مدل انتخابی در این پژوهش مدل B0 می‌باشد.

مدل EfficientNet بر روی مجموعه داده ImageNet آموزش داده شده است مجموعه داده‌ای با بیش از ۱/۲ میلیون داده آموزشی. یکی از تکنیک‌های متداول و موثر در بینایی ماشین و یادگیری عمیق به ویژه در پردازش تصاویر پزشکی، یادگیری انتقالی می‌باشد. به این معنا که بجای استفاده کردن از وزن‌های تصادفی و به روز رسانی آنها در هر دور آموزش، از وزن‌های قبلی که هنگام آموزش بر روی دیتاست ImageNet بدست آمده‌اند استفاده کنیم و در هر مرحله آنها را بهبود ببخشیم. یادگیری عیمق بدلیل کمبود داده در زمینه پردازش تصاویر پزشکی بسیار حائز اهمیت است زیرا برای آموزش شبکه‌های پیچیده مانند EfficientNet به تعداد بسیار زیادی داده نیاز است، که بنا بر حساسیت‌ها در حوزه پزشکی قابل دسترسی نیست.

نحوه فراخانی ماژول EfficientNet در پایتون:

```
base_model = tf.keras.applications.EfficientNetB0(input_shape=(224, 224, 3), include_top=False, weights='imagenet')
```

توجه شود که EfficientNet بر روی مجموعه داده ImageNet آموزش داده شده است که شامل ۱۰۰۰ کلاس می‌باشد. برای همین لایه آخر که شامل هزار نورون است را حذف می‌کنیم و در ادامه یک لایه دیگر با دو نورون به عنوان لایه خروجی قرار خواهیم داد. حذف کردن لایه آخر توسط آرگمان Include_top انجام شده است.

در ادامه لایه خروجی اضافه می‌شود.

```
hp_layer_1 = hp.Int('layer_1', min_value=64, max_value=256, step=64)
hp_learning_rate = hp.Choice('learning_rate', values=[1e-2, 1e-3, 1e-4])

model = tf.keras.Sequential([
    base_model,
    tf.keras.layers.GlobalAveragePooling2D(),
    tf.keras.layers.Dense(units=hp_layer_1, activation='relu'),
    tf.keras.layers.Dense(2, activation='softmax')  # 2 classes, so final layer with softmax activation
])
```

همچنین در این مرحله به کمک تابع keras_tuner هایپرپارامترهای شبکه را تنظیم می‌کنیم. این تابع برای هر حالت از هایپرپارامترها (در اینجا ضریب یادگیری و تعداد نورون‌های لایه قبل از آخر) یک بار برای مدت

محدودی شکبه را آموزش می‌دهد و بهترین حالت هایپرپارامترها را به روش سعی و خطا بدست می‌آورد. در اینجا بهترین مقدار برای ضریب یادگیری ۰/۰۰۰۱ و تعداد نورون‌ها ۲۵۶ بدست می‌آید. اگر با این مقادیر شبکه را آموزش دهیم. دقت بر داده‌های صحت سنجی به بیش از ۹۵٪ خواهد رسید.

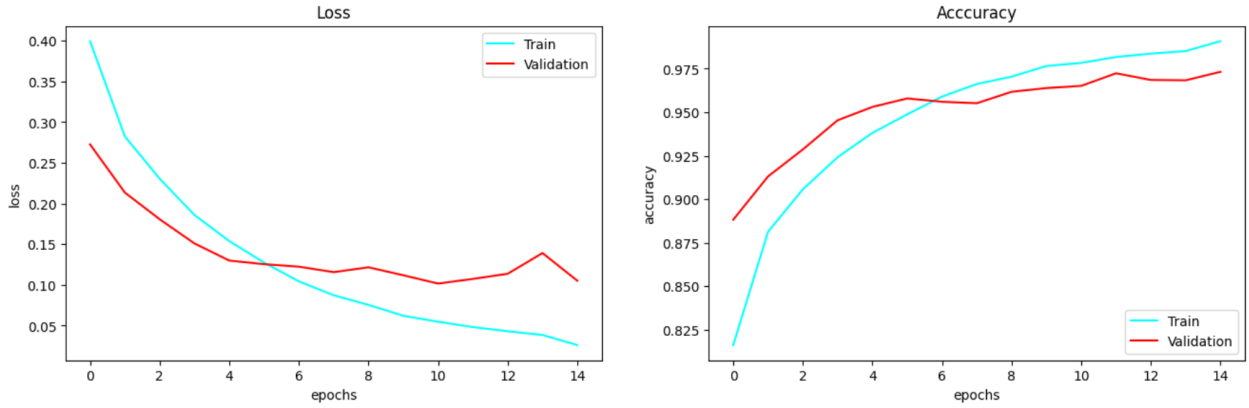
در مرحله بعد از فریز کردن لایه‌ها استفاده می‌کنیم. معماری EfficientNet دارای حدود ۲۵۵ لایه است. به دلیل آنکه تعداد داده‌ها در دیتاست استفاده شده کم است، ممکن است این میزان داده برای آموزش دادن تمامی لایه‌ها کافی نباشد. روال کلی فریز کردن لایه‌ها به این صورت است که ابتدا تمامی لایه‌ها را فریز یا غیر قابل به روزرسانی کنیم و سپس چند لایه از آخر شبکه را قابل اپدیت کنیم. طبیعتا نمی‌توان به طور مشخص گفت چه تعداد لایه باید قابل آموزش و به‌روزرسانی شوند و این یکی از هایپرپارامترهای دیگر شبکه خواهد بود. لذا، دوباره از تابع keras_tuner کمک می‌گیریم.

```
# Freeze layers in the base model
base_model.trainable = True
fine_tune_at = hp.Int('freeze_at', min_value=50, max_value=240, step=50)
for layer in base_model.layers[:fine_tune_at]:
    layer.trainable = False
```

کد بالا مشخص می‌کند که چند لایه از اول شبکه باید فریز شوند. بدین منظور ۵۰، ۱۰۰، ۱۵۰ و ۲۰۰ لایه را به طور مجزا فریز کرده و هر بار نتایج را بررسی میکند. برای شبکه طراحی شده اگر ۵۰ لایه ابتدایی فریز شوند بهترین نتیجه حاصل می‌شود. به دلیل اینکه ۵۰ کمترین عدد داده شده به keras_tuner برای بررسی بود و عملا مقادیر کمتر را بررسی نکردیم لازم است بررسی‌های بیشتر نیز انجام شود تا به بهترین نتیجه برسیم.

پس از سعی و خطای بسیار، مقدار ۳۰ بهترین گزینه ممکن برای این شبکه انتخاب شد. همینطور یکی از نکات مهمی که ممکن است به راحتی باعث خطا شود این است که لایه‌های BatchNormalization به هیچ عنوان نباید قابل آموزش باشند در غیر این صورت عملکرد مدل تا حد زیادی تحت شعاع قرار می‌گیرد.

در نهایت مدل را با پارامترهای بهینه آموزش میدهیم. نتایج مدل بر داده‌های آموزش و صحت سنجی مطابق نمودار زیر خواهد بود.



شکل ۴-۶ منحنی‌های دقت و خطا در مدل اول

۴-۲-۱ نکات تکمیلی

مدل بر روی GPU-T4 در کولب آموزش داده شده است.

بدلیل محدودیت حافظه رم نمی‌توان تمامی داده‌ها را در یک مرحله بارگذاری کرد. و به کمک تابع ImageDataGenerator داده‌ها را در بچ‌های ۳۲ تایی بارگذاری می‌کنیم. اینکار با کد زیر انجام می‌شود.

```
# using data generator to load data in batches
train_datagen = tf.keras.preprocessing.image.ImageDataGenerator()
validation_datagen = tf.keras.preprocessing.image.ImageDataGenerator()
test_datagen = tf.keras.preprocessing.image.ImageDataGenerator()


train_generator = train_datagen.flow_from_directory(directory=train_dir, target_size=(224, 224), color_mode="rgb",
                                                    batch_size=32, class_mode="categorical", shuffle=True, seed=42)
validation_generator = validation_datagen.flow_from_directory(directory=validation_dir, target_size=(224, 224), color_mode="rgb",
                                                              batch_size=32, class_mode="categorical", shuffle=True, seed=42)
test_generator = test_datagen.flow_from_directory(directory=test_dir, target_size=(224, 224), color_mode="rgb",
                                                  batch_size=32, class_mode="categorical", shuffle=True, seed=42)
```

هنگام آموزش از کاهش دهنده ضریب یادگیری استفاده شده است. بدین صورت که اگر دقت بر روی داده‌های صحت سنجی طی سه ایپاک متوالی بهبود نیابد ضریب یادگیری نیم برابر می‌شود. اینکار به افزایش دقت مدل کمک خواهد کرد.

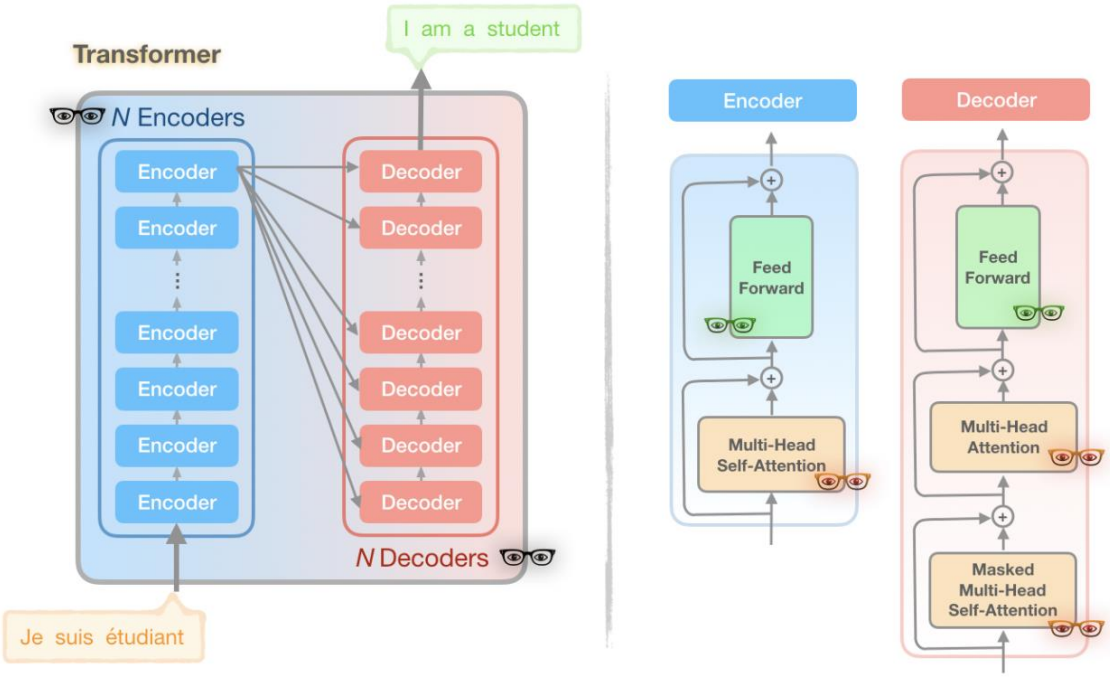
```
reduce_lr = tf.keras.callbacks.ReduceLROnPlateau(monitor='val_loss', factor=0.5, patience=3, verbose=1)
```

همینطور اگر مدل در هر ایپاک عملکرد بهتری روی داده‌های صحت سنجی داشته باشد به عنوان مدل نهایی ذخیره می‌شود.

دقت مدل نهایی EfficientNet بر داده‌های تست ۹۵/۴٪ می‌باشد.

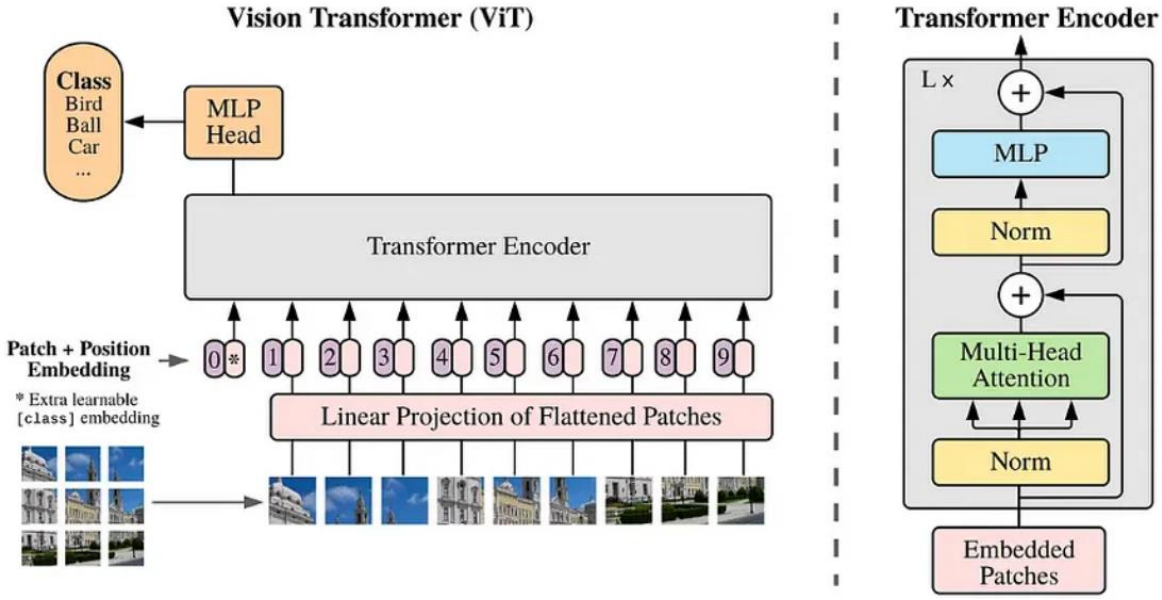
۴-۳ مدل ترانسفورمر بینایی

ترانسفورمرها نوعی مدل یادگیری عمیق در پردازش زبان طبیعی (NLP) هستند که از مکانیزمی به نام (self-attention) برای پردازش داده‌های ورودی استفاده می‌کنند. برخلاف مدل‌های سنتی ترتیبی مانند RNNها و LSTMها، ترانسفورمرها می‌توانند توالی‌های ورودی را به صورت موازی پردازش کنند که این امر آن‌ها را برای مجموعه داده‌های بزرگ بسیار سریع‌تر و کارآمدتر می‌سازد. مکانیزم توجه به خود به ترانسفورمرها اجازه می‌دهد تا اهمیت کلمات مختلف در یک جمله را بدون توجه به موقعیت آن‌ها ارزیابی کنند که این امر در به‌دست‌آوردن وابستگی‌های بلندمدت و اطلاعات متنی به‌طور مؤثر کمک می‌کند. این مدل‌ها در مقاله "توجه تمام چیزی است که نیاز دارید" در سال ۲۰۱۷ معرفی شدند و از معماری رمزگذار-رمزگشا (encoder-decoder) تشکیل شده‌اند. رمزگذار توالی ورودی را پردازش کرده و مجموعه‌ای از نمایش‌های مبتنی بر توجه تولید می‌کند، در حالی که رمزگشا از این نمایش‌ها برای تولید توالی خروجی استفاده می‌کند. این معماری منجر به توسعه مدل‌های قدرتمند زبانی مانند BERT، GPT و T5 شده است که عملکردی پیشرو در وظایف مختلف NLP نظیر ترجمه، خلاصه‌سازی و پاسخ به سوالات به نمایش گذاشته‌اند.



شکل ۴-۷ نمای کلی از ترانسفورمرها در پردازش زبان طبیعی

ساختار ترانسفورمر به‌طور گسترده‌ای در پردازش زبان طبیعی (NLP) استفاده می‌شود. مدل ویژن ترانسفورمر کاملاً بر اساس ساختار ترانسفورمر پیاده‌سازی شده و هیچ ساختار CNN در آن وجود ندارد. ساختار ترانسفورمر شامل مجموعه‌ای از اجزای رمزگذار و مجموعه‌ای از اجزای رمزگشا است، در حالی که مدل ویژن ترانسفورمر یک مدل طبقه‌بندی تصویر است و نیازی به رمزگشا ندارد. بنابراین، تنها یک جزء رمزگذار در ساختار ترانسفورمر ویژن ترانسفورمر وجود دارد. جزء رمزگذار شامل یک دسته از شش رمزگذار یکسان است. هر رمزگذار از یک لایه توجه چندگانه و یک لایه پیش‌خورنده تشکیل شده و هر دو لایه دارای ساختار اتصال باقیمانده و لایه LayerNorm هستند. ساختار یک جزء رمزگذار ویژن ترانسفورمر در شکل ۴-۸ نشان داده شده است.



شکل ۴-۸ ساختار ترانسفورمر بینایی و یک بخش انکودر

نگاهی به الگوریتم توجه چندگانه:

$$Q_i = QW_i^Q,$$

$$K_i = KW_i^K,$$

$$V_i = VW_i^V,$$

$$I = 1, \dots, 8$$

$$Headi = Attention(Q_i, K_i, V_i)$$

$$MultiHead(Q,K,V) = Concact(head_1, \dots, head_8)W^o$$

در این فرمول‌ها، Q به معنای بردار پرس‌وجو (query vector)، K به معنای بردار کلید (key vector)، V به معنای بردار مقدار (value vector) و W به معنای ماتریس وزن (weight matrix) است.

برای آموزش مدل از مدل موجود در کتابخانه pytorch استفاده خواهیم کرد برای استفاده از این ماژول باید دیتاست در یک پوشه قرار بگیرد به گونه‌ای که آن پوشه حاوی سه پوشه، هر کدام برای یک قسمت از داده‌ها (داده‌های آموزش، صحت سنجی، و تست) باشد. هر یک از این پوشه‌ها نیز شامل دو پوشه متناسب با دو کلاس لوسمی یا سالم خواهد بود.

برای استفاده کردن از ماژول ترانسفورمر بینایی از کد ریز استفاده می‌کنیم

```
# 1. Get pretrained weights for ViT-Base
pretrained_vit_weights = torchvision.models.ViT_B_16_Weights.DEFAULT

# 2. Setup a ViT model instance with pretrained weights
pretrained_vit = torchvision.models.vit_b_16(weights=pretrained_vit_weights).to(device)
```

مشاهده می‌شود که از وزن‌های از قبل آموزش داده شده استفاده میکنیم این وزن‌ها بدست آمده از آموزش شبکه بر روی داده‌های دیتاست ImageNet هستند، که دارای ۱۰۰۰ کلاس طبقه بندی است در حالی که دیتاست استفاده در این پژوهش دارای دو کلاس لوسمی یا سالم است.

استفاده از وزن‌های از پیش آموزش داده شده مصداق یادگیری انتقالی است که قبل‌تر بررسی شد در اینجا به طور جامع‌تر مزایای یادگیری انتقالی را بررسی میکنیم:

صرفه‌جویی در زمان و منابع محاسباتی: با استفاده از مدل‌های از پیش آموزش‌دیده، نیازی به آموزش مدل از صفر نیست و می‌توان از شبکه‌های عصبی که بر روی مجموعه داده‌های بزرگ و متنوع آموزش دیده‌اند، بهره برد. این امر به کاهش زمان و هزینه‌های محاسباتی کمک می‌کند.

عملکرد بهتر با داده‌های کمتر: در پروژه‌هایی که داده‌های برچسب‌گذاری‌شده کافی در دسترس نیست، یادگیری انتقالی می‌تواند بهبود قابل توجهی در عملکرد مدل ایجاد کند. با استفاده از وزن‌های از پیش آموزش‌دیده، مدل می‌تواند از دانشی که قبلاً به دست آورده استفاده کند و به داده‌های جدید با تعداد کمتری نمونه‌ها عمومی‌تر شود.

تسریع فرآیند توسعه: به دلیل اینکه نیازی به آموزش مدل از ابتدا نیست، توسعه‌دهندگان می‌توانند سریع‌تر به نتایج قابل قبول دست یابند و زمان بیشتری را صرف بهینه‌سازی و بهبود مدل‌ها کنند.

قابلیت استفاده در دامنه‌های مختلف: مدل‌های یادگیری انتقالی می‌توانند به راحتی به دامنه‌ها و وظایف مختلف منتقل شوند. برای مثال، یک مدل آموزش‌دیده برای تشخیص اشیاء در تصاویر می‌تواند به خوبی برای تشخیص چهره‌ها یا حتی تحلیل تصاویر پزشکی استفاده شود.

بهبود دقت و عملکرد: در بسیاری از موارد، مدل‌های یادگیری انتقالی دقت و عملکرد بهتری نسبت به مدل‌های آموزش‌دیده از ابتدا دارند. این به دلیل این است که مدل‌های از پیش آموزش‌دیده اغلب روی مجموعه داده‌های بزرگ و متنوعی آموزش دیده‌اند و بنابراین ویژگی‌های کلی‌تری را یاد گرفته‌اند.

کاهش خطر بیش‌برازش (Overfitting): با استفاده از مدل‌های از پیش آموزش‌دیده و تنظیم دقیق آنها بر روی داده‌های خاص، خطر بیش‌برازش کاهش می‌یابد. مدل‌ها از ویژگی‌های عمومی‌تری که در مراحل اولیه یادگیری به دست آمده‌اند، استفاده می‌کنند و در نتیجه کمتر به داده‌های خاص حساس می‌شوند.

قابلیت استفاده از دانش پیشین: یادگیری انتقالی به مدل‌ها این امکان را می‌دهد که از دانش و تجربیات قبلی استفاده کنند. این امر به ویژه در مواردی که داده‌های آموزشی محدود و گرانبها هستند، بسیار مفید است.

ساختار داخلی ترانسفورمر به صورت زیر است مشاهده می‌شود که دارای ۱۱ بلوک انکودر است که بالاتر نحوه عملکرد آن را بررسی کردیم.

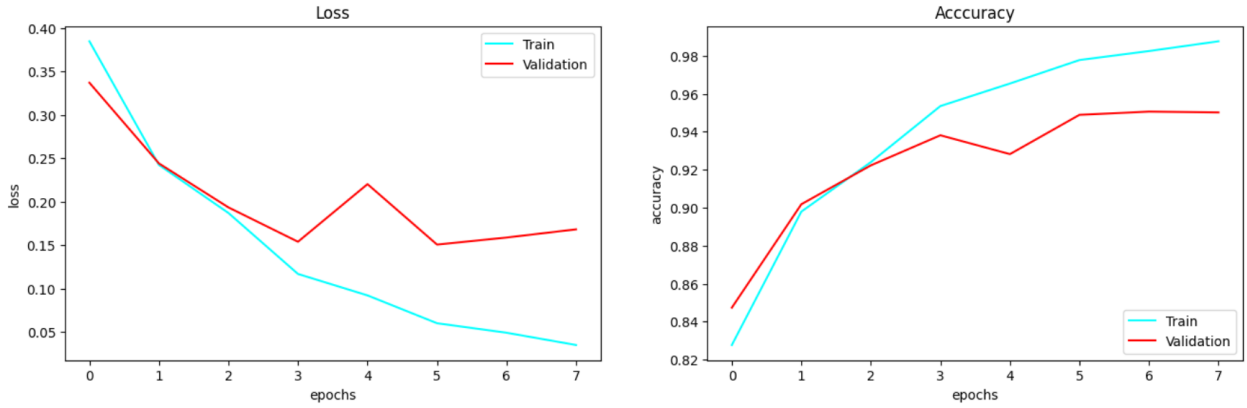
VisionTransformer (VisionTransformer)	[32, 3, 224, 224]	[32, 2]	768	Partial
├─Conv2d (conv_proj)	[32, 3, 224, 224]	[32, 768, 14, 14]	(590,592)	False
├─Encoder (encoder)	[32, 197, 768]	[32, 197, 768]	151,296	Partial
│ └─Dropout (dropout)	[32, 197, 768]	[32, 197, 768]	--	--
│ └─Sequential (layers)	[32, 197, 768]	[32, 197, 768]	--	Partial
│ │ └─EncoderBlock (encoder_layer_0)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_1)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_2)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_3)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_4)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_5)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_6)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_7)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_8)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_9)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_10)	[32, 197, 768]	[32, 197, 768]	(7,087,872)	False
│ │ └─EncoderBlock (encoder_layer_11)	[32, 197, 768]	[32, 197, 768]	7,087,872	True
│ └─LayerNorm (ln)	[32, 197, 768]	[32, 197, 768]	1,536	True
├─Linear (heads)	[32, 768]	[32, 2]	1,538	True

در این مرحله بجز آخرین ماژول انکودر، بقیه لایه‌ها را فریز و غیرقابل آموزش می‌کنیم.

```python
# 3. Freeze the base parameters
for parameter in pretrained_vit.parameters():
    parameter.requires_grad = False

# 3.1. Unfreeze the last three layers
num_layers = len(pretrained_vit.encoder.layers)
for block in pretrained_vit.encoder.layers[-1:]:
    for param in block.parameters():
        param.requires_grad = True

# 3.2. Unfreeze LayerNorm
for param in pretrained_vit.encoder.ln.parameters():
    param.requires_grad = True
```

با توجه به اینکه مدل ترانسفورمر بسیار کندتر از مدل آموزش می‌یابد و به دلیل محدودیت‌های استفاده از GPU کولب، مدل را با ۸ ایپاک آموزش می‌دهیم.

نتایج مدل ترانسفورمر بر داده‌های آموزش و صحت سنجی:



شکل ۴-۹ منحنی‌های خطا و دقت در مدل دوم

۴-۳-۱ نکات تکمیلی

مدل بر روی GPU T4 در کولب آموزش داده شده است.

بدلیل محدودیت حافظه رم نمی‌توان تمامی داده‌ها را در یک مرحله بارگذاری کرد. و به کمک تابع DataLoader داده‌ها را در بچ‌های ۳۲ تایی بارگذاری می‌کنیم. اینکار در پایتورچ با کد زیر انجام می‌شود.

```
# Turn images into data loaders
train_dataloader = DataLoader(
    train_data,
    batch_size=batch_size,
    shuffle=True,
    num_workers=num_workers,
    pin_memory=True,
)
val_dataloader = DataLoader(
    val_data,
    batch_size=batch_size,
    shuffle=False,
    num_workers=num_workers,
    pin_memory=True,
)
```

هنگام آموزش از کاهش دهنده ضریب یادگیری استفاده شده است. بدین صورت که اگر دقت بر روی داده‌های صحت سنجی طی ایپاک بعدی بهبود نیابد ضریب یادگیری ۰/۲ برابر می‌شود. اینکار به افزایش دقت مدل کمک خواهد کرد.

```
scheduler = ReduceLROnPlateau(optimizer, mode='min', factor=0.2, patience=1, verbose=True)
```

همینطور مدل با ضریب یادگیری ۰/۰۰۱ و ضریب رگولارایز ۰/۰۰۰۴ آموزش داده شده است.

الگوریتم بهینه‌ساز Adamax و تابع خطا CrossEntropyLoss هستند.

در مرحله آخر نیز بهترین مدل آموزش داده شده را ذخیره می‌کنیم و در فاز تست از آن برای بررسی عملکرد مدل روی داده‌هایی که تا بحال دیده نشده‌اند استفاده خواهیم کرد.

فصل پنجم: نتایج

فاز تست مدل بر داده‌های دیده نشده:

ابتدا پیش‌بینی مدل بر داده‌های تست را بدست می‌آوریم. برای این کار از کد زیر استفاده می‌شود.

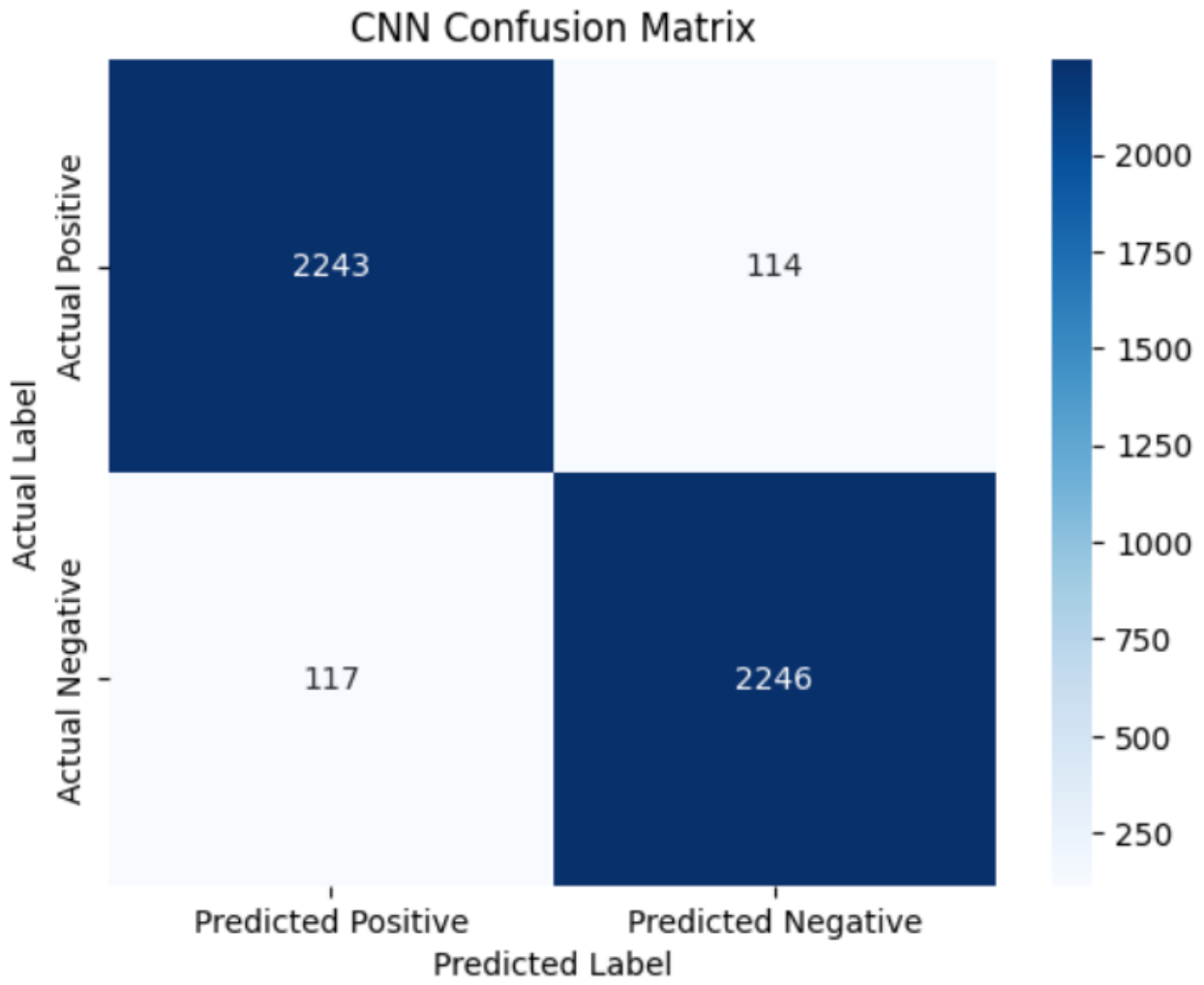
```
predictions = []

for sample in Data[1400:1600]:
  path_to_image = os.path.join('/content/C_NMC_2019_dataset/test/hem', sample)

  _, CNN_pred = make_predictions_CNN(path_to_CNN_model, path_to_image)
  _, ViT_pred = make_predictions_ViT(path_to_ViT_model, path_to_image)

  predictions.append([sample, CNN_pred, ViT_pred])

path_to_save = '/content/drive/MyDrive/Thesis/predictions.pkl/hem_1400-1600.pkl'

# Save to Pickle file
with open(path_to_save, 'wb') as pkl_file:
    pickle.dump(predictions, pkl_file)
```

مشاهده می‌شود که برای بخشی از داده‌ها نتیجه هر دو مدل آموزش داده شده را بدست می‌آوریم و به همراه اسم تصویر در یک لیست پایتون ذخیره می‌کنیم. در مرحله آخر نیز این نتایج در یک فایل pkl ذخیره می‌شوند تا در مراحل بعدی مورد استفاده قرار گیرند.

بدلیل محدودیت‌های سخت افزار داده‌های تست را به چند بخش تقسیم می‌کنیم و برای هر بخش اقدامات توضیح داده شده انجام می‌شوند.

۵-۱ نتیجه مدل مبتنی بر معماری EfficientNet بر داده‌های تست

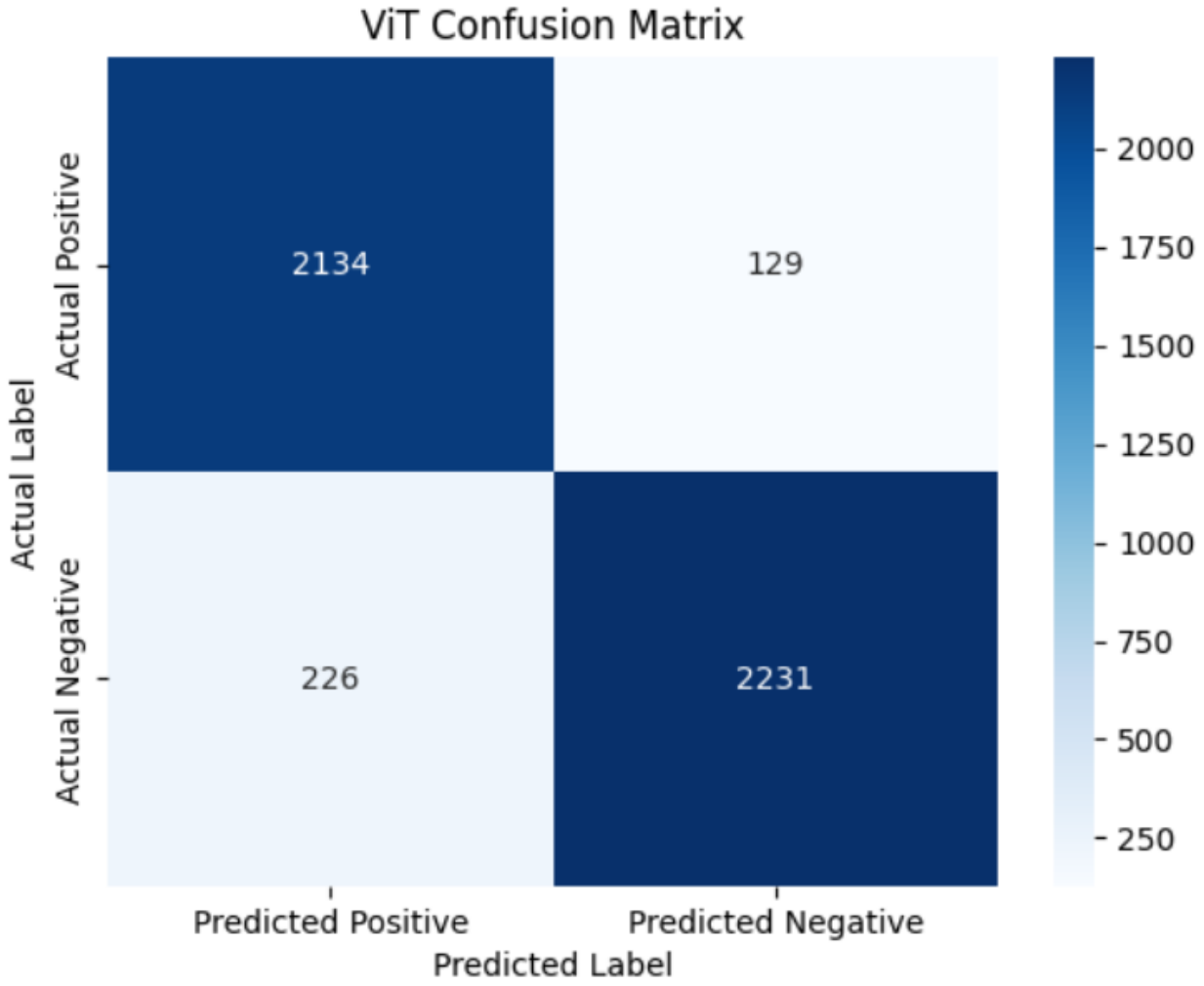
مدل شماره ۱ دقت ۹۵/۱٪ را بر داده‌های تست بدست می‌آورد. همچنین Confusion Matrix مربوطه در ادامه آورده شده است.



شکل ۵-۱ ماتریس کانفیوژن مدل اول

۵-۲ نتیجه مدل منبتی بر معماری ترانسفورمر بینایی بر داده‌های تست

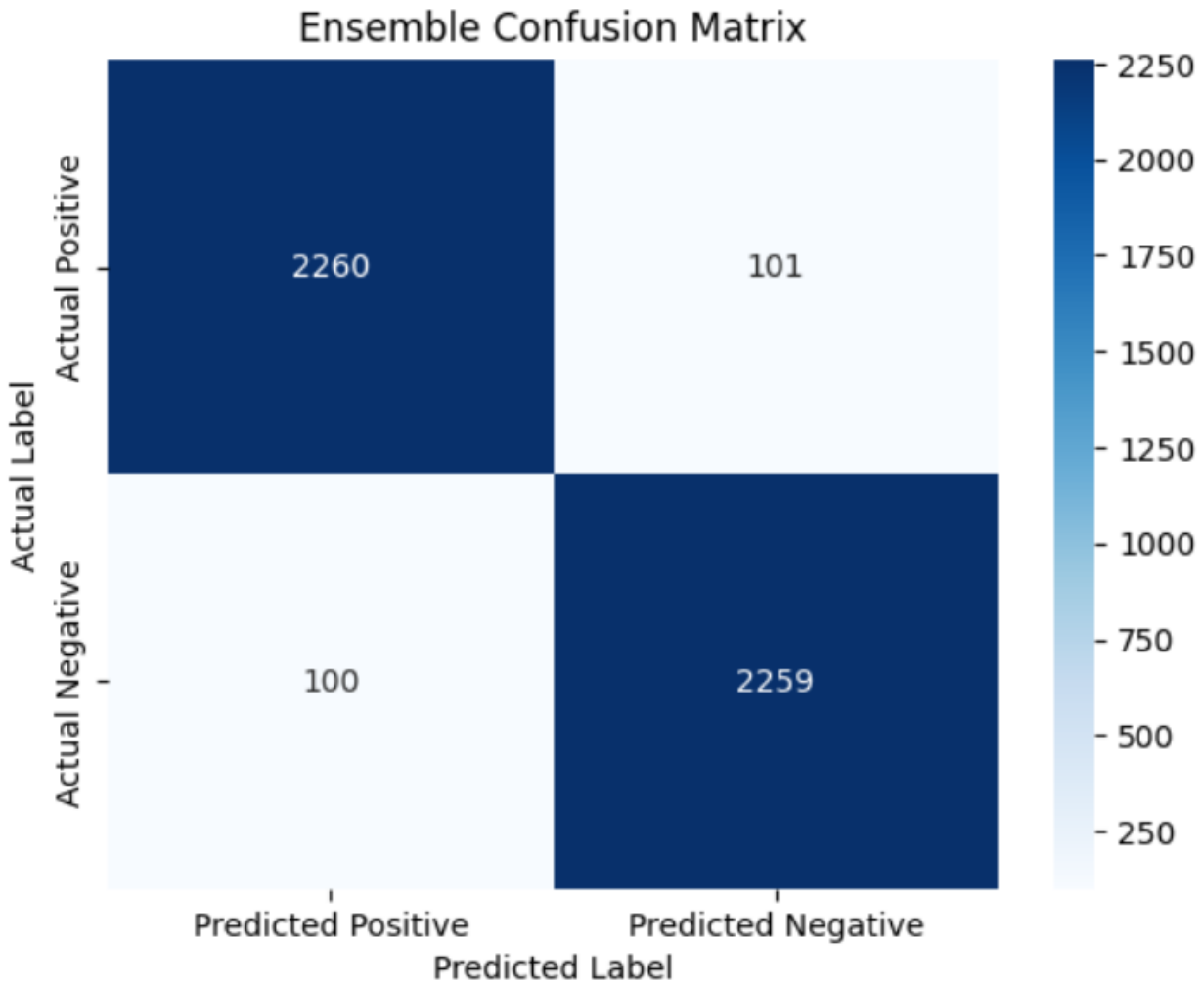
مدل شماره ۲ دقت ۹۲/۵٪ را بر داده‌های تست بدست می‌آورد. همچنین Confusion Matrix مربوطه در ادامه آورده شده است.



شکل ۵-۲ ماتریس کانفیوژن مدل دوم

۵-۳ نتیجه مدل ensemble بر داده‌های تست

این مدل که میانگین وزن دار از نتایج دو مدل قبل است دارای دقت ۹۵/۷٪ بر داده‌های تست است. میانگین وزن دار با وزن‌های ۵۵٪ از مدل اول و ۴۵٪ از مدل دوم انتخاب شده است که بهترین نتیجه را می‌دهد.
همچنین Confusion Matrix مربوطه در ادامه آورده شده است.



شکل ۵-۳ ماتریس کانفیوژن مدل سوم

۵-۴ مقایسه بین سه مدل ارائه شده و متریک‌های دیگر

مقایسه ای از این سه مدل در جدول ۵-۱ آورده شده است.

جدول ۵-۱

Models\Metrics	Accuracy	Precision	Recall	F1-Score
CNN	0.9511	0.9516	0.9504	0.9510
ViT	0.9248	0.9430	0.9042	0.9232
Ensemble	0.9574	0.9572	0.9576	0.9574

فصل ششم: جمع بندی

۶-۱ جمع بندی

در این مطالعه، ما یک رویکرد تشخیصی برای لوسمی لنفوبلاستیک حاد پیشنهاد کردیم که می‌تواند سلول‌های سرطانی و سلول‌های طبیعی را از طریق یک مدل ترکیبی طبقه‌بندی کند تا به پزشکان در تشخیص در واقعیت کمک کند. مجموعه داده C-NMC-2019 در این مقاله استفاده شد؛ ما روش DERS را برای حل مشکل عدم تعادل داده‌ها پیشنهاد کردیم. ما یک مدل ترکیبی طراحی کردیم که مدل ترانسفورمر بینایی و مدل EfficientNet را در مدل ترکیبی ViT-CNN ادغام می‌کند. دقت این مدل در طبقه‌بندی سلول‌های سالم و سرطانی ۹۵/۷۴٪ بود. مدل ترکیبی ViT-CNN عملکرد قابل مقایسه‌ای نسبت به این مدل‌های قبلی داشت. نتایج نشان داد که مدل پیشنهادی در این مقاله از نظر دقت برتر از سایر مدل‌ها بوده و توانایی طبقه‌بندی متعادلی داشت که می‌تواند به تشخیص لوسمی لنفوبلاستیک حاد کمک کند. لازم به ذکر است که مدل ترانسفورمر در صورت وجود سخت افزار مناسب و یافتن پارامترهای بهینه‌تر قادر است عملکرد بهتری را بدست بیاورد و در نهایت بر دقت مدل ترکیبی ViT-CNN تاثیرگذار باشد.

منابع و مراجع

[1] Talaat, F.M., Gamel, S.A. A2M-LEUK: attention-augmented algorithm for blood cancer detection in children. *Neural Comput & Applic* **35**, 18059–18071 (2023).

[2] Hegde RB, Prasad K, Hebbar H, Singh BMK, Sandhya I. Automated Decision Support System for Detection of Leukemia from Peripheral Blood Smear Images. J Digit Imaging. 2020 Apr;33(2):361-374. doi: 10.1007/s10278-019-00288-y. PMID: 31728805; PMCID: PMC7165227.

[3] Namayandeh SM, Khazaei Z, Lari Najafi M, Goodarzi E, Moslem A. GLOBAL Leukemia in Children 0-14 Statistics 2018, Incidence and Mortality and Human Development Index (HDI): GLOBOCAN Sources and Methods. Asian Pac J Cancer Prev. 2020 May 1;21(5):1487-1494. doi: 10.31557/APJCP.2020.21.5.1487. PMID: 32458660; PMCID: PMC7541866.

[4] Zolfaghari, M., Sajedi, H. A survey on automated detection and classification of acute leukemia and WBCs in microscopic blood cells. *Multimed Tools Appl* **81**, 6723–6753 (2022).

[5] Xing F, Yang L. Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review. IEEE Rev Biomed Eng. 2016;9:234-63. doi: 10.1109/RBME.2016.2515127. Epub 2016 Jan 6. PMID: 26742143; PMCID: PMC5233461.

[6] Rezayi S, Mohammadzadeh N, Bouraghi H, Saeedi S, Mohammadpour A. Timely Diagnosis of Acute Lymphoblastic Leukemia Using Artificial Intelligence-Oriented Deep Learning Methods. Comput Intell Neurosci. 2021 Nov 11;2021:5478157. doi: 10.1155/2021/5478157. PMID: 34804144; PMCID: PMC8601812.

[7] Zakir Ullah, M.; Zheng, Y.; Song, J.; Aslam, S.; Xu, C.; Kiazolu, G.D.; Wang, L. An Attention-Based Convolutional Neural Network for Acute Lymphoblastic Leukemia Classification. *Appl. Sci.* **2021**, *11*, 10662.

[8] S. Perveen, A. Alourani, M. Shahbaz, M. U. Ashraf and I. Hamid, "A Framework for Early Detection of Acute Lymphoblastic Leukemia and Its Subtypes From Peripheral Blood Smear Images Using Deep Ensemble Learning Technique," in *IEEE Access*, vol. 12, pp. 29252-29268, 2024.

[9] M. Ghaderzadeh, A. Hosseini, F. Asadi, H. Abolghasemi,D. Bashash, and A. Roshanpoor, “Automated detection modelin classification of B-lymphoblast cells from normal B-lymphoid precursors in blood smear microscopic imagesbased on the majority voting technique,” Scientific Program-ming, vol. 2022, Article ID 4801671, 2022.

[10] Mondal, Chayan & Hasan, Md. Kamrul & Jawad, Md Tasnim & Dutta, Aishwariya & Islam, Md & Awal, Md.abdul & Ahmad, Mohiuddin. (2021). Acute Lymphoblastic Leukemia Detection from Microscopic Images Using Weighted Ensemble of Convolutional Neural Networks. 10.20944/preprints202105.0429.v1.

[11] Atteia, G.; Alhussan, A.A.; Samee, N.A. BO-ALLCNN: Bayesian-Based Optimized CNN for Acute Lymphoblastic Leukemia Detection in Microscopic Blood Smear Images. *Sensors* **2022**, *22*, 5520. https://doi.org/10.3390/s22155520

Abstract

Leukemias are a deadly group of cancerous diseases that affect people of all ages, including children and adults, and are one of the major causes of mortality worldwide. Specifically, this disease is associated with an increase in immature lymphocytes and causes damage to the bone marrow or blood. Currently, manual analysis of blood samples obtained through microscopic images is used to diagnose this disease, which is often very slow, time-consuming, and inaccurate.

Automatic detection of leukemia or blood cancer is a challenging and much-needed task in medical centers. In recent decades, deep learning using deep neural networks has provided advanced approaches for image classification problems. However, there is still room for improvement in their efficiency, learning process, and performance. Therefore, in this research study, we have proposed a new version of the deep learning algorithm for the diagnosis of acute lymphoblastic leukemia (ALL) based on deep and machine learning with analysis of microscopic images of blood samples.

We propose the ViT-CNN hybrid model for classifying images of cancer cells and normal cells to aid in the diagnosis of acute lymphoblastic leukemia. The ViT-CNN model is a hybrid model that combines the Vision Transformer model and the Convolutional Neural Network model. The ViT-CNN hybrid model can extract features of cell images in two completely different ways to achieve better classification results. The classification accuracy of the ViT-CNN hybrid model on the test set has reached 95.74%, which is comparable to other methods. The proposed method can accurately distinguish cancer cells from normal cells and can be used as an effective method for computer-aided diagnosis in acute lymphoblastic leukemia.

Keywords: Acute lymphoblastic leukemia, medical image processing, deep learning, vision transformer.

Amirkabir University of Technology
(Tehran Polytechnic)

EE

Leukemia Cancer Detection in Microscopic Blood Samples using deep neural networks

Undergrad Thesis

By

Parsa Pilevar

Supervisor

Dr. Seyedin

August 2024