به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین دوم

پرسش ۱	نام دانشجو	ثمین سلوکی
	شماره دانشجویی	[redacted]

فهرست

شکل‌ها

جدول ها

پرسش ۱) تشخیص آلزایمر با استفاده از تصویر برداری مغزی (ADNI)

۱-۱) معرفی مسئله

مجموعه دادگان شامل ۱۶۵۴ تصویر از دو کلاس متفاوت "AD یا MCI" است. کلاس AD، مخفف Alzheimer's disease، مربوط به بیماران تشخیص داده شده با بیماری آلزایمر است. کلاس MCI، مخفف Mild Cognitive Impairment، مربوط به بیمارانی است که کاهش جزئی اما قابل توجهی در توانایی های شناختی، مانند حافظه و مهارت های تفکری داشته اند؛ این افراد در معرض افزایش خطر ابتلا به بیماری آلزایمر یا سایر انواع زوال عقل هستند.

هدف از انجام این پروژه، طبقه بندی بیماران به دو دسته ی AD و MCI است و در شناسایی علائم اولیه زوال شناختی ضرورت دارد و برای تشخیص زودهنگام و مداخله در بیماری های عصبی حائز اهمیت است.

۱-۲) پیش پردازش تصاویر

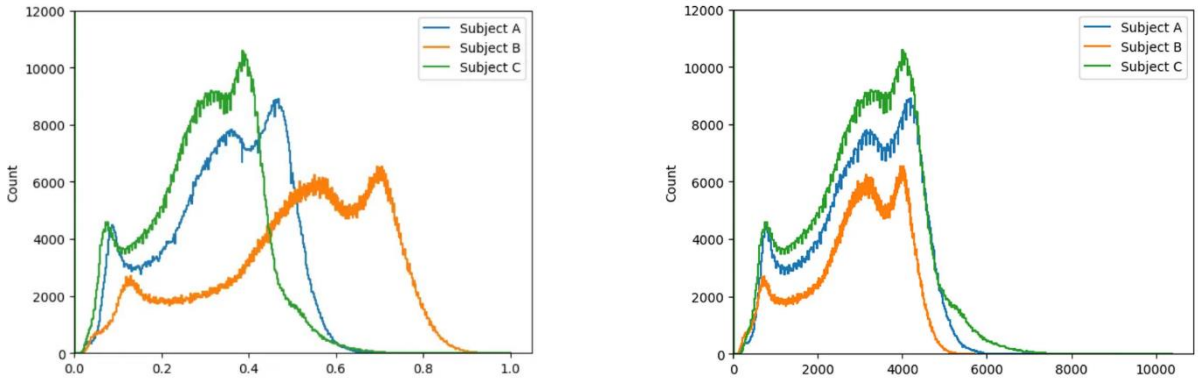
پس از **فراخوانی دادگان** مربوط به پروژه و **تغییر سایز** تمامی تصاویر به ۱۲۸×۱۲۸ ، و استفاده از **GRAYSCALE**_cv2.IMREAD برای شناساندن عکس های سیاه و سفید، اقدام به **نرمالسازی** می کنیم. (لازم به ذکر است که سایز عکس ها برگرفته از مدل پیشنهادی مقاله ی پایه است. در مقاله ذکر می شود که بهترین مدل آموزش دیده با سایز (۱۲۸و۱۲۸) است.)

به صورت کلی intensity normalization برای آموزش شبکه لازم است زیرا که به همگرایی کارآمدتر شبکه کمک می‌کند. مقادیر بین ۰ و ۱ ورودی به شبکه اجازه می دهد که وزن ها را آزادانه مقدار دهی و به روز کند و تحت تاثیر مقادیر خود نورون ها نباشد، و باعث جلوگیری از رخداد exploding gradients می شود. برای این مرحله، Min-Max normalization روشی است که معمولا برای پیش پردازش تصاویر استفاده می شود، اما به گفته ی وبسایت medium[1] ، intensity normalization تصاویر پزشکی، مانند MRI مغز با نرمالسازی دیگر تصاویر متفاوت است، در این مقاله ذکر می شود که:

- در برخورد با تصاویر عادی، نرمالسازی بین کران ۰ و ۲۵۵ کاری معمول است، اما هنگام برخورد با تصاویر پزشکی باید رویکرد متفاوتی اتخاذ کرد زیرا که شدت تصویر image intensity منعکس کننده ی نوع بافت است و این شدت نسبی است و محدوده ی شدت، محدود نیست!

[1] https://medium.com/@susanne.schmid/image-normalization-in-medical-imaging-f586c8526bd1

به طور مثال در شکل ۱، عکس سمت راست، هیستوگرام ۳ عکس متفاوت بدست آمده از اسکنرGE را مشاهده می کنید که پس از نرمالسازی با روش Min-Max به شکل سمت چپ در آمده است. همانطور که مشخص است CSF (قله های مختلف هیستوگرام) ، ماده سفید و ماده خاکستری هنگامی که برای آموزش در شبکه قرار می گیرند، مقادیر شدت متفاوتی دارند. ناهماهنگی هیستوگرام ها پس از نرمال سازی Min-Max به دلیل موارد پرت مانند وجود/فقدان بافت خاصی است، مثلا: برداشتن جمجمه معیوب که بافت چربی یا متغیرهای وابسته به اسکنر را حفظ می کند (مثلا نویز یا طراحی توالی پالس).



شکل ۱- **Effect of Min-Max Scaling**

به توضیح بیشتر می پردازیم:

- تصاویر MRI با contrastای که برای بافت نرم قائل می شود به به رادیولوژیست اجازه می دهد تا بین انواع مختلف بافت تمایز قائل شود. voxel intensity به ترکیب بافت خاص اشاره دارد در حالی که در مقایسه با تصاویر طبیعی این contrast، مقدار شدت روشنایی جسم را منعکس می کند و مفهوم متفاوتی دارد.
- شدت تصویر MR نسبی است. MRI معمولی کیفی است به این معنی که حتی اگر مقدار شدت به یک بافت خاص اشاره دارد، یک اسکن MRI مکرر شدت وکسل های متفاوتی را به دست می آورد. با این حال، شدت مطلق در MRI اهمیتی ندارد زیرا تصاویر MR از کنتراست بین بافت‌های مختلف بهره می‌برند و نه از مقدار مطلق. مقدار شدت به خودی خود معنای فیزیکی ندارد بلکه جریان اندازه گیری شده در سیم پیچ گیرنده است که با استفاده از تبدیل فوریه تبدیل شده است و به عواملی چون توالی پالس، سخت افزار و تغییرات دما و غیره بستگی دارد. مقدار شدت مطلق در MRI متفاوت است و نه چشم انسان و نه تجسم کامپیوتر نمی تواند بین هزاران مقدار خاکستری تفاوت قائل شوند؛ در این هنگام توصیه می شود از Window & Level برای تجزیه و تحلیل شدت ناحیه مورد نظر استفاده شود تا هنگام باز کردن یک تصویر در یک نمایشگر پزشکی، Window & Level به طور خودکار برای دستیابی به بهترین کنتراست تنظیم شوند.(این در حالی است که هنگام ترسیم تصاویر با استفاده از کتابخانه هایی مانند matplotlib، دامنه شدت کامل به طور مساوی به bin ها تقسیم می شود.)

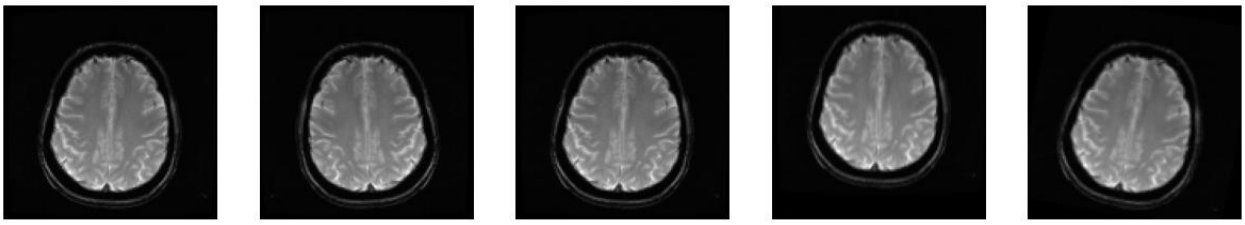
- محدوده شدت تصویر محدود نیست. شدت تصویر در MRI متفاوت است و هیچ حد بالایی واقعی وجود ندارد به طور مثال، برخی از اسکنرها تصاویری با شدت مطلق ۱۰۰۰۰ دارند، برخی دیگر دارای شدت ۶۰۰۰ هستند، لذا دامنه ی شدت متفاوت است و کران بالای مشخصی نمی توان تعریف کرد زیرا که این کران به بسته به بیماران و عملکرد متفاوت دستگاه ها، تغییر می کند.

با توجه به توضیحات ذکر شده، روش متفاوتی برای نرمالسازی به جای Min-Max normalization انتخاب شد: نرمالسازی دسته: نرمالسازی کل دسته با میانگین و انحراف معیار مربوطه. در این نرمالسازی، پارامترهای آماری مانند میانگین و انحراف استاندارد کل دسته برای نرمالسازی تصاویر با تفریق میانگین و تقسیم تصویر بر انحراف استاندارد استفاده می شود. از آنجایی که پارامترها در دسته ثابت هستند، عملیات خطی است و منجر به هم‌ترازی هیستوگرام می‌شود.

در مرحله ی بعد، مجموعه داده به ۳ بخش **آموزش، اعتبارسنجی و تست** تقسیم می شود. بهترین تقسیم طبق نتایج مقاله ی پایه انتخاب شد، این انتخاب بدین شرح است: تقسیم داده های آموزش و تست به ترتیب ٪۹۵ و ۵٪ و تقسیم داده ی آموزش به ۹۰٪ صرفا برای آموزش و ۱۰٪ صرفا برای اعتبارسنجی.

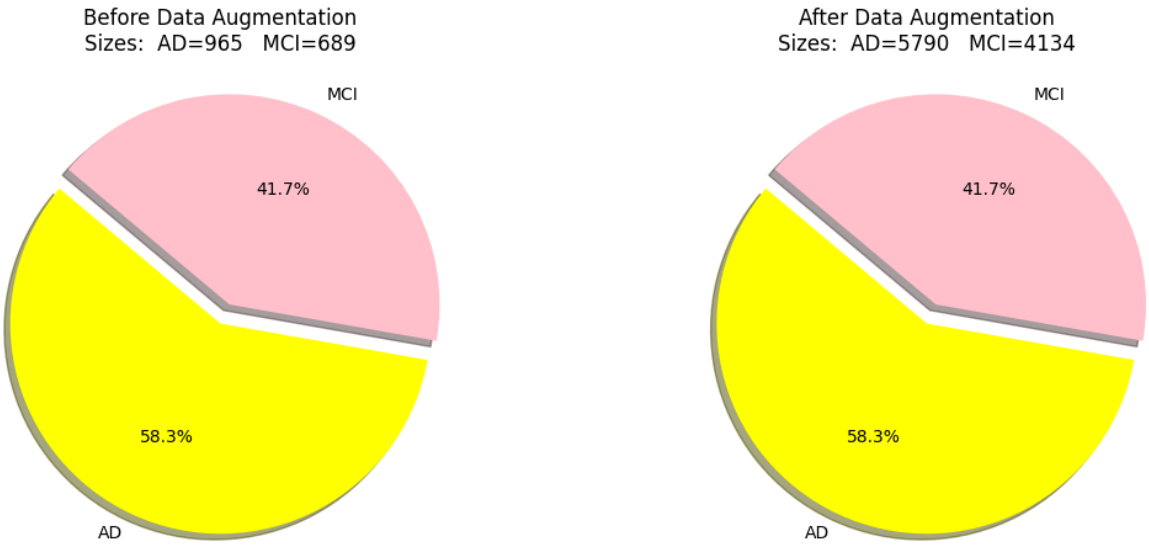
۱-۳) داده افزایی Data Augmentation

در مقاله ی پایه، از هر تصویر موجود پنج تصویر دیگر تولید شده است، این داده افزایی به صورت چرخش افقی، برش با دامنه ۰.۲، جابجایی با دامنه ۰.۱، چرخش با ۱۵ درجه و زوم با دامنه ۰.۲ انجام شد ، به طور مثال نتیجه ی یک تصویر را در زیر می توانید مشاهده کنید:



شکل ۲- نمونه ای از داده های افزوده شده

همچنین در دو نمودار زیر، توزیع کلاس ها قبل و بعد از انجام Data Augmentation، نمودار است:



شکل ۳- توزیع داده ها قبل و بعد از **data augmentation**

افزایش داده ها، تعداد و تنوع داده های آموزشی را افزایش می دهد و توزیع کلاس یا فراوانی کلاس ها را در کلاس مربوطه تغییر نمی دهد و نسبت کلاس اصلی در طول فرآیند افزایش حفظ می شود. در نمودار هم مشخص است که توزیع آماری کلاس ها قبل و بعد تغییر نکرده است. (درصد فرکانس قبل وبعد ثابت است.)

نکته ی لازم به ذکر در این مرحله، این است که ابتدا برای هم افزایی داده ، روش های هم افزایی همگی با هم به ImageDataGenerator داده شد، نتیجه ی این هم افزایی، باعث تقویت مدل نشد! به طوری که دقت مدل بدون انجام افزایی حدود ۸۰ درصد و با انجام هم افزایی حدود ۶۰ درصد می شد! با بررسی دقیق تر مقاله ی پایه و نتایج مدل اجرا شده، یافتیم که روش های هم افزایی داده ی ذکر شده، می بایست که جدا جدا انجام شود به طوری که نباید ترکیب چند روش هم افزایی روی داده ها اجرا شود. پس از پیاده سازی هر تکنیک داده افزایی به صورت جداگانه، دقت مدل به ۸۰ درصد رسید و مطلوب بود. تحلیل و نتایج بیشتر داده افزایی در بخش تحلیل نتایج آورده شده است.

۱-۴) پیاده سازی

پس از آماده سازی داده و تقسیم داده ها به ۳ گروه آموزش، تست، اعتبار سنجی ضمن اطمینان از توزیع کلاس ها، به سراغ آماده سازی خود شبکه می رویم: مقدار دهی اولیه وزن های شبکه بنا بر مقاله ی پایه ، با استفاده از Glorot Uniform weight initializer انجام شد. این مقداری دهی اولیه به گونه ای وزن های شبکه را مقداردهی می کند که توابع فعال سازی نورون در مناطق اشباع یا مرده شروع نشوند و نهایتا باعث همگرایی سریعتر و دقت بالاتر می شود. به طور دقیق تر، در این روش، وزن های شبکه از یک توزیع یکنواخت

در یک محدوده خاص که طوری انتخاب می شوند که واریانس فعالساز ها و گرادیان ها ندرتا ثابت بمانند که از وقوع vanishing gradients یا vanishing gradients جلوگیری می کند.

برای انجام طبقه بندی در این پروژه، سه معماری استفاده شده در مقاله پیاده شد، تابع هزینه استفاده شده، Binary Cross Entropy طبق پیاده سازی دقیق مقاله ی پایه بوده زیرا عملکرد این تابع هزینه نسبت به توابع دیگر، در آموزش مدل بسیار بهتر بوده . Optimizer استفاده شده آدام است و نرخ یادگیری ۰.۱ (طبق مدل پایه) مقدار دهی شد.

نهایتا، . مدل نهایی ای که ما به تحلیل آن پرداختیم، دارای لایه ی ورودی کانولوشنی با ۳۲ فیلتر و اندازه کرنل (۳و۳) با GlorotUniform به عنوان kernel initializer است. در ادامه با لایه ی BatchNormalization که در مقاله ی پایه هم بدان اشاره شد، مقادیر نرمال شده و در لایه ی بعد مجدد لایه ی کانولوشنی با ۳۲ فیلتر و اندازه کرنل (۳و۳) و سپس لایه ی MaxPooling2D به سایز (۲و۲) اضافه شد. در لایه ی بعدی مجدد همین لایه ها اضافه شدند و در بخش fully connected لایه ی با ۱۲۸ سپس ۶۴ نورون سپس ۲ نورون به عنوان لایه ی خروجی قرار گرفت. توابع فعالساز بین لایه های مخفی ReLU بوده و برای لایه ی آخر Softmax. جزییات این معماری در شکل زیر به نمایش درآمده:

```
proposed_model = tf.keras.Sequential([
    Conv2D(32, (3,3), input_shape=X_train.shape[1:], kernel_initializer=GlorotUniform()),
    BatchNormalization(),
    Conv2D(32, (3,3)),
    BatchNormalization(),
    MaxPooling2D(pool_size=(2,2)),
    Conv2D(32, (3,3)),
    BatchNormalization(),
    Conv2D(32, (3,3)),
    BatchNormalization(),
    MaxPooling2D(pool_size=(2,2)),
    Flatten(),
    Dense(128),
    Dense(64),
    Dense(2),
    Activation('softmax')])
```

شکل ۴- معماری مدل

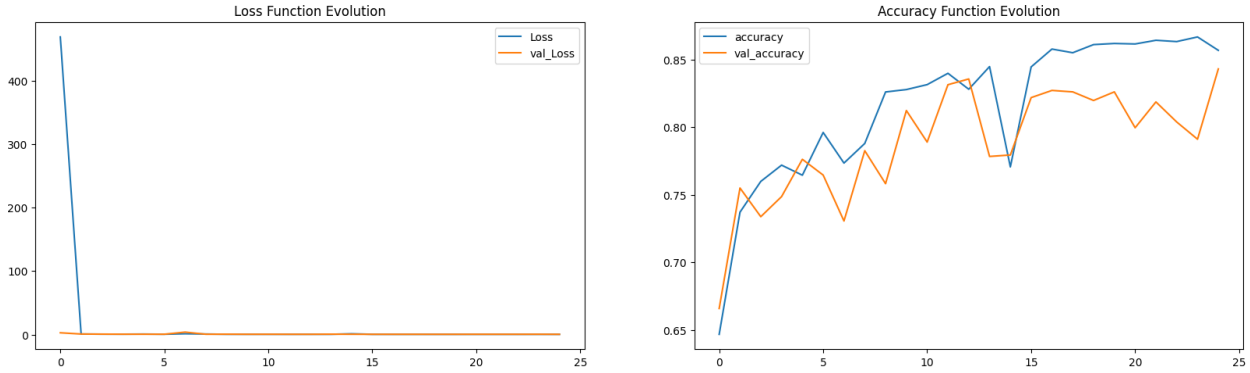
سپس این مدل با نرخ یادگیری ۰.۱ با بهینه ساز آدام آموزش و تابع هزینه ی binary_crossentropy کامپایل شد.

۱-۵) تحلیل نتایج

در این مرحله به ارزیابی مدل می پردازیم.

نمودار خطا و دقت

در شکل زیر، نمودار خطا و دقت در ۲۵ ایپاک را مشاهده می کنید.

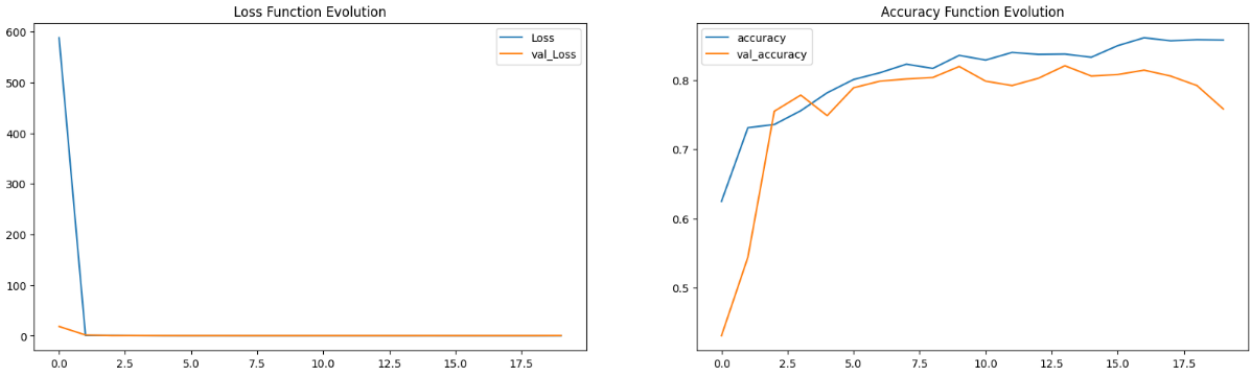


شکل ۵- دقت و خطای مدل با **epoch=25**

دقت مدل در ایپاک آخر بر روی داده های آموزش ٪۸۵.۶۸ بوده و بر روی داده های اعتبارسنجی، ٪۸۴.۳۱ است. همانطور که در نمودار loss مشخص است، سیر نزولی تابع هزینه تا ایپاک ۲۵ برقرار بوده است.

برای نمودار دقت باید گفت دقت از ٪۶۴.۶۹ در ایپاک ۱ شروع می شود و به تدریج در طول ایپاک ها بهبود می یابد. در ایپاک ۱۰، دقت به ٪۸۱.۲۳ افزایش یافته است که نشان از بهبود مداوم است. در ایپاک ۱۱ اندکی کاهش دقت وجود دارد، اما در دوره های بعدی به سرعت بهبود می یابد. دقت در ایپاک ۲۵ به ۸۴.۳۱ درصد می رسد که بهبود قابل توجهی را نسبت به دقت اولیه نشان می دهد. به طور کلی، یک روند مثبت در دقت در طول ایپاک ها وجود دارد (با برخی از نوسانات در طول تکامل) اما در نهایت شاهد بهبود هستیم.

از نمودار دریافت می شود که مدل ممکن است در ایپاک ۱۷ تا ۲۳ کمی بیش از حد برازش شده باشد. یا افت عملکرد در ایپاک ۱۴ می تواند نشان از برازش بیش از حد باشد زیرا که ممکن است انطباق بیش از حد بر داده های آموزشی اتفاق افتاده و عملکرد در تعمیم پذیری به داده ها در ایپاک ۱۴ رخ داده است. به همین دلیل (امکان برازش بیش از حد) در اجرای بعدی مدل، تعداد ایپاک ها را از ۲۵ به ۲۰ کاهش دادیم.نتیجه ی مدل به شکل زیر بود:



شکل ۶ـدقت و خطای مدل با **epoch=20**

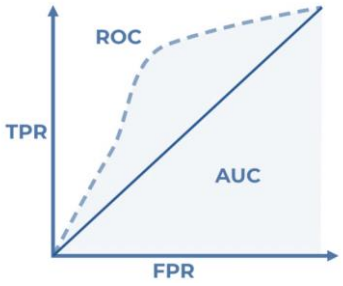
دقت بر داده ی تست قبل از کاهش(با ایپاک ۲۵)، ۸۲.۲۹٪ بوده و با تعداد ایپاک ۲۰، به ۷۹٪ رسید، precision ٪۷۵ بوده و به ۶۹٪ رسید. Recall٪۸۳ بوده و به ۸۸٪ رسید. F1 score٪۷۹ بوده و به ۷۷٪ رسید. بعد از بررسی کاهش عملکرد مدل هنگامی که تعداد ایپاک ها به ۲۰ می رسند ، تصمیم گرفتیم که مدل اول را مبنا قرار دهیم.

Confusion Matrix

در ارزیابی مدل از شاخص هایی چون accuracy, precision, recall, AUC-ROC استفاده می کنیم؛ ابتدا به توضیح منحنی AUC-ROC (مخفف Area Under the Receiver Operating Characteristic curve) که برای ارزیابی توانایی یک مدل برای تمایز بین دو کلاس استفاده می شود، می پردازیم.

منحنیROC (مخفف Receiver Operating Characteristics) نمایش گرافیکی اثربخشی مدل طبقه بندی باینری است و نرخ مثبت واقعی (TPR) در مقابل نرخ مثبت کاذب (FPR) را در آستانه های طبقه بندی مختلف ترسیم می کند.

منحنی AUC (مخفف Area Under the Curve) نشان دهنده سطح زیر منحنی ROC است و عملکرد کلی مدل طبقه بندی باینری را اندازه گیری می کند. از آنجایی که هر دو TPR و FPR بین ۰ تا ۱ قرار دارند، بنابراین، مقدارآن همیشه بین ۰ و ۱ قرار می گیرد و مقدار بیشتر AUC نشان دهنده عملکرد بهتر مدل است. هدف اصلی، به حداکثر رساندن این ناحیه به منظور داشتن بالاترین TPR و کمترین FPR در آستانه معین است. AUC این احتمال را می سنجد که مدل به یک نمونه مثبت تصادفی انتخاب شده، احتمال پیش بینی شده بالاتری را در مقایسه با یک نمونه منفی تصادفی می دهد. این نشان دهنده احتمالی است که مدل می تواند بین دو کلاس موجود، تمایز قائل شود.



شکل ۷- **ROC**

- TPR و FRP، (مخففFalse Positive Rate- True Positive Rate)
  - True positive : مثبت واقعی و به عنوان مثبت پیش بینی شده است.
  - True Negative: منفی واقعی و پیش بینی شده به عنوان منفی.
  - False Positive (خطای نوع اول): منفی واقعی اما مثبت پیش بینی شده است.
  - False Negative (خطای نوع دوم): مثبت واقعی اما به عنوان منفی پیش بینی شده است.

در بیان ساده تر، False Positive را می توان هشدار غلط و False Negative را یک از دست رفته معنا کرد.

- Sensitivity= TP/(TP+FN)

این شاخص توانایی مدل در شناسایی صحیح موارد مثبت را نشان می دهد.

- Specificity= TN/(TN+FP)

این شاخص، نسبت نمونه های منفی واقعی را که به درستی به عنوان منفی شناسایی شده اند را اندازه میگیرد و نشان دهنده ی توانایی مدل برای شناسایی صحیح موارد منفی است.

- FPR=FP/(TN+FP)

این شاخص، نسبت نمونه های متفی که به اشتباه طبقه بندی شده اند را نشان میدهد.

همانطور که از فرمول شاخص های برمی آید، بین Specificity و Sensitivity رابطه ی معکوس بر قرار است و می بایست یک trade-off بین مثبت واقعی و منفی واقعی ایجاد کرد. با تعیین یک آستانه این بالانس تعریف می شود بدین صورت که آستانه ی پایین، مقادیر بالاتر Sensitivity و تعداد بیشتر مثبت واقعی به قیمت تعداد بیشتر false positive را باعث می شود و آستانه ی بالا، مقدار بالاتر Specificity را باعث می شود که به قیمت false positive کمتر است اما بهمراه false negative بیشتر !

در مدل اجرا شده، نتایج confusion matrix به شکل زیر است:

		Actual Values	
		Positive(MCI)	Negative(AD)
Prediction	Positive(MCI)	TP=266	FP=24
	Negative(AD)	FN=64	TN=143

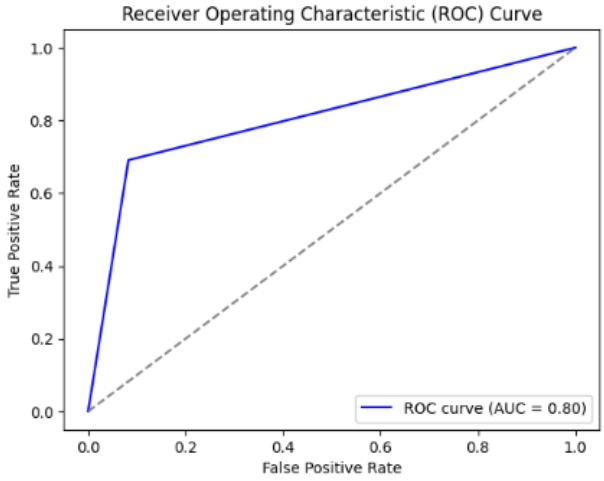
گزارش چهار شاخص دیگر به شرح زیر است:

- Accuracy: 0.8229
- Precision: 0.8562
- Recall: 0.69082
- F1 score: 0.76470

Accuracy ۰.۸۲۲۹ نشان می دهد که مدل تقریباً ۸۲.۲۹٪ موارد را در داده های تست به درستی طبقه بندی کرده است. این معیار مهم، یک معیار کلی از اثربخشی مدل در تمایز بین بیماران AD و MCI بر اساس تصاویر MRI است که عملکرد خوبی را نشان میدهد. Precision ، که نسبت پیش بینی های مثبت واقعی را در بین تمام پیش بینی های مثبت انجام شده توسط مدل اندازه گیری می کند، به امتیاز بالای ۰.۸۵۶۲ دست یافت. این نشان می‌دهد که وقتی مدل پیش‌بینی می‌کرد که بیمار مبتلا به AD یا MCI باشد، تقریباً ۸۵.۶۲٪ مواقع درست بود. Precision بالا مطلوب بوده زیرا نشان دهنده توانایی مدل در به حداقل رساندن پیش بینی های مثبت کاذب است. Recall که نشان دهنده نسبت موارد مثبت واقعی است که توسط مدل به درستی شناسایی شده نشان میدهد که تقریباً ۶۹.۰۸٪ از موارد واقعی AD و MCI را در داده های آزمایشی به دقت شناسایی کرد. در حالی که این معیار در مقایسه با دقت نسبتاً پایین‌تر است، اما همچنان برای ثبت موارد مثبت واقعی مهم است. F1 score که ارزیابی متعادلی از عملکرد مدل ارائه می دهد با مقدار ۷۶.۴۷٪ نشان می دهد که این مدل به تعادل خوبی بین دقت و Recall در پیش بینی های خود روی داده های آزمایشی دست یافته است. این نشان می دهد که این مدل هم در به حداقل رساندن مثبت های کاذب و هم در گرفتن مثبت های واقعی موثر است. به طور کلی، مدل CNN عملکرد قوی در طبقه بندی بیماران مبتلا به AD و MCI بر اساس تصاویر MRI نشان می دهد. دقت بالا، دقت و امتیاز F1 نشان می دهد که مدل قادر به پیش بینی دقیق است.

نکته ی دیگری که حائز اهمیت است، تفاوت معنای این شاخص هاست. همانطور که پیش از این ذکر شد، مقدار FP و FN نمیتواند همزمان با هم کاهش پیدا کند(مگر با افزایش داده ها که الان مورد بحث نیست)، لذا

در انتخاب مدل یا ارزیابی مدل، به مفهوم که توجه کنیم، باید در نظر بگیریم که پیامد ارتکاب خطا در کدام بیشتر است؟ FP یا FN؟ FN به این معناست که فرد واقعا سالم بوده اما به غلط مدل تشخیص آلزایمر داده است، که ۶۴ مورد بوده است. FP یعنی فرد واقعا بیمار بوده و از آلزایمر رنج می برده است اما مدل پیش بینی کرده که فرد سالم است. (۲۴ مورد) این اطلاعات برای تصمیم گیری پزشکان حایز اهمیت است. فرض کنید یک داروی خاص برای بیمارانی قلبی که آلزایمر هم دارند، تاثیر معکوس میگذارد. در این کیس، پیش بینی آلزایمری نبودن بسیار مهم است، مهمتر از اینکه فرد سالم را به غلط، AD تشخیص دهد. در اینجا که عواقب تشخیص غلط AD بسیار زیادتر است، می بایست که FP حداقل شود. و مدلی که بتواند این کار را انجام دهد، ارزش بیشتری دارد، مانند مدل بدست آمده در این پروژه.



شکل ۸**ROC-**

نمودار ROC مدل، نتیجه ی مطلوبی را نشان میدهد. سطح زیر نمودار مقدار مطلوب است. همچنین این نمودار با خط زاویه ۴۵ درجه فاصله ی مطلوبی نشان می دهد. برای این نمودار مجدد بسته به درجه اهمیت نتایج و عواقب متفاوت میتوان ارزیابی متفاوتی به عمل آورد. مثلا در threshold بالا، خطای FP را میتوانیم کمتر کنیم اما همچنان FN چالش برانگیز می ماند و باید به هدف پروژه و مسئله دقت کرد. لازم به ذکر است که تغییر این threshold تغییر confusion matrix را بدنبال دارد و در ادامه تغییر ROC.

۱-۶) مقایسه نتایج

در این بخش به مقایسه ی نتایج مدل های CNN با شاخصه های متفاوت می پردازیم.

مورد ۱) اثر نسبت تقسیم آموزش-تست ٪۵۰-٪۳۰: نتیجه ی آموزش شبکه با درصد جدید تقسیم داده ها (٪۳۰-٪۵۰) در زیر نمایش داده شده است:

جدول ۱-**Effect of Sample Size**

%	Accuracy	Precision	Recall	F1 score
Train-test:95-5	82.29	85.62	69.08	76.47
Train-test:30-70(50+20)	72.39	76.82	48.34	59.34

مشخصا مدل آموزش دیده شده با مقدار داده ی آموزش کمتر، عملکرد ضعیف تری نسبت به مدل آموزش-تست ۹۵-۵ با مقدار داده ی بیشتر دارد. وجود داده ی بیشتر، منجر به دیدن و یادگیری الگوهای بیشتر می شود و ویژگی های کم اهمیت تر در مدل کمرنگ تر شده و قابلیت تعمیم پذیری مدل را بهبود می دهد. این دریافت همسو با یافته های مقاله است که با تخصیص داده ی بیشتر به آموزش، عملکرد بهبود پیدا میکند.

مورد ۲) اثر dropout: dropout به‌طور تصادفی کسری از واحدهای ورودی را در طول آموزش صفر می‌کند، و به کاهش اتکای شبکه به ویژگی‌های خاص کمک می‌کند، و نسبت به جزئیات خاص داده های آموزش حساسیت کمتری پیدا می کند و احتمال کمتری دارد که نویز یا نقاط پرت را در داده ها به خاطر بسپارد، و نهایتا باعث بهبود عملکرد در مدل می شود. در مدل این پروژه، شبکه پیش از اعمال dropout، accuracy داده تست ۸۲.۲۹٪ بوده و Precision: ۸۵.۶۲٪ و Recall: ۶۹٪. پس از اعمال dropout شاهد بهبود دقت به ، accuracy داده تست ۸۴.۱ ٪ بوده و Precision: ۸۴.۰۴٪ و Recall: ۷۶٪ هستیم. این دستاورد هم سو با گفته ی مقاله ی پایه با بهبود عملکرد مدل به صورت میانگین از ۹۹.۹۸۱٪ به ۹۹.۹۸۷ بوده.

جدول ۲- اثر **dropout**

	Accuracy	Precision	Recall	F1 score
Without dropout	82.29	85.62	69.08	76.47
With dropout	84.1	84.04	76.32	80

مورد ۳)Glorot Initializer

با آغاز سازی وزن های شبکه بدون استفاده از Glorot Initializer ،تغییر زیر را در شاخص ها شاهد بودیم:

جدول ۳ -اثر **GlorotUniform**

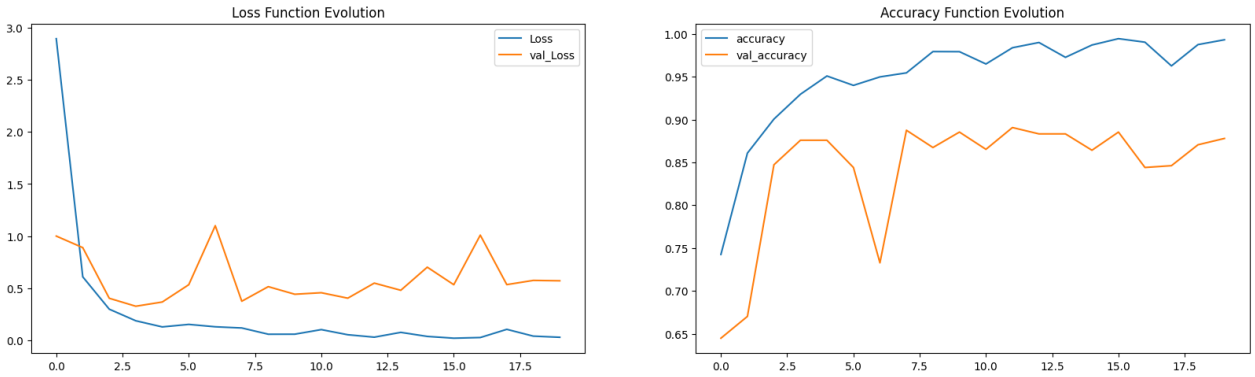
	Accuracy	Precision	Recall	F1 score
Default(Glorot Initializer)	82.29	85.62	69.08	76.47
RandomNormal(mean=0.0, stddev=0.05)	78.26	67.87	90.82	77.68

از داده های جدول، می توان گفت استفاده از GlorotUniform تاثیر مثبت بر آموزش شبکه و افزایش تعمیم پذیری دارد.

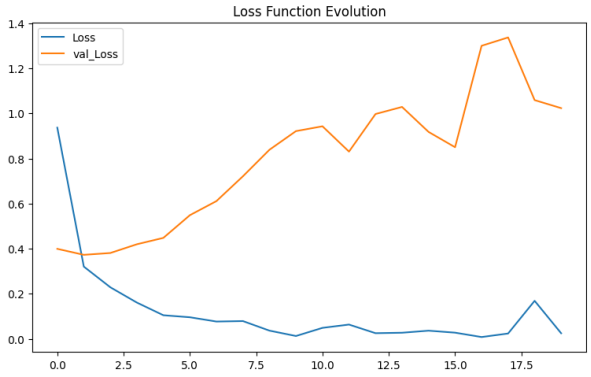
مورد ۴)Proposed model و Testing model 2 و Testing model 1

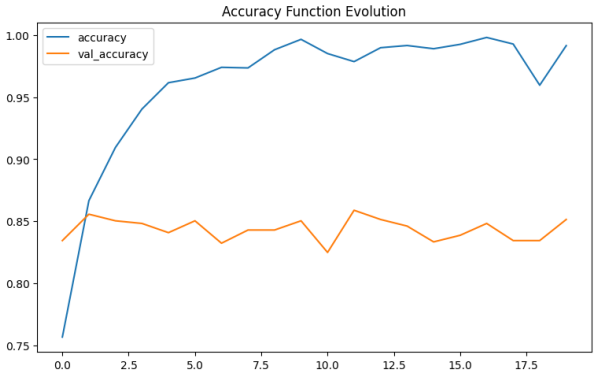
جدول ۴- مقایسه مدل پیشنهادی با مدل های تست

	Accuracy	Precision	Recall	F1 score
Proposed mode	82.29	85.62	69.08	76.47
Test 1	88.73	83.55	90.8	87.03
Test 2	87.32	89.56	78.74	83.8



شکل ۹ **- test model 1**





شکل ۲۱۰ **- Test model 2**

از جدول و نمودارهای فوق نتیجه می شود که مدل تست ۱ و مدل تست ۲ هردو احتمالا دچار برازش بیش از حد هستند و این قضیه تعمیم پذیری مدل را اندکی ضعیف می کند. سه مدل در این پروژه هر کدام در یک شاخص بر مدل دیگر برتری دارند. به طور مثال، شاخص Precision مدل پیشنهادی را از مدل تست ۱، برتر اعلام میکند اما شاخص Recall برعکس این موضوع را بیان می کند. تفاوت جالب در این بخش این است که این شاخص ها با شاخص های مقاله، بدلیل تفاوت دیتاست استفاده شده برای آموزش مدل است و طبیعی است که مدل های یکسان بر روی دیتاست های متفاوت، عملکرد متفاوتی از خود نشان دهند. اما در دنیای واقعی که یک دیتاست و چند مدل را داریم، سوال این است که کدام مدل برای ما بهتر است؟! برای پاسخ به این سوال مجدد به هدف مسئله و هدف انجام پروژه اشاره می کنیم. هر کدام از این شاخص ها دارای مفهوم متفاوت هستند و میتوانند بسته به موضوع، کاربرد متفاوت داشته باشند و منجر به انتخاب مدل متفاوت شود. لذا برای انتخاب بهترین مدل، مخصوصا در شرایطی مثل این جدول که مدل ها غالب پذیر نیستند (non-dominant) می بایست با خبرگان حوزه مشورت کرد و بنا بر بررسی درخواست ها و اهداف، بهترین مدل اعلام شود.