

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه مهندسی فناوری اطلاعات

پایان نامه کارشناسی

رشته مهندسی کامپیوتر گرایش فناوری اطلاعات

عنوان پایان نامه

تحلیل ویژگی های رادیومیکس برای تشخیص عفونت در ریه

استاد راهنما:

دکتر آزاده محمدی

پژوهشگر:

سحر سرکار

شهریور ۱۴۰۰



دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه مهندسی فناوری اطلاعات

پروژه کارشناسی رشته‌ی مهندسی کامپیوتر گرایش فناوری اطلاعات

خانم سحر سرکار

تحت عنوان

تحلیل ویژگی‌های رادیومیکس برای تشخیص عفونت در ریه

در تاریخ / / ۱۳ توسط هیأت داوران زیر بررسی و با نمره به تصویب نهایی رسید.

۱- استاد راهنمای پروژه:

دکتر آزاده محمدی

امضا

۲- استاد داور :

دکتر سیدفخرالدین نوربهبهانی

امضا

امضای مدیر گروه

تشکر و قدردانی

سپاس خدای بزرگ را که مرا یاری رساند تا بتوانم این مقطع تحصیلی را به پایان رسانده و گامی در راستای اعتلای علم بردارم. از پدر و مادرم، مهربان‌ترین همراهان زندگی‌ام، تشکر می‌کنم که حضورشان همیشه گرمابخش وجود من بوده است. از استاد بزرگوارم خانم دکتر محمدی برای تمام راهنمایی‌ها و زحمات بی‌دریغشان سپاس گزارم. از جناب آقای دکتر حسن‌پور که در انجام این پروژه همواره الهام‌بخش و راهنمای من بودند کمال تشکر را دارم. در نهایت سپاس فراوان از تمام اساتیدی که در طول سال‌های دانشگاه، معنی واقعی استاد را تداعی کرده و بذر دانش را در وجودمان کاشتند.

تقدیم به

پدر و مادر عزیز و مهربانم

که در سختی‌ها و دشواری‌های زندگی

همواره یآوری دلسوز و فداکار و پشتیبانی

محکم و مطمئن برایم بوده‌اند.

چکیده:

بیماری کرونا نامی است که امروزه بیش از هر چیز به گوش می‌رسد. بیماری که کل جهان را فراگرفته و جلوگیری و درمان آن بزرگ‌ترین دغدغه در جوامع امروزی است. این بیماری روزانه جان هزاران نفر در دنیا را می‌گیرد. ریه‌ها هدف اصلی این ویروس هستند. برای درمان و جلوگیری از پیشرفت این بیماری، یکی از مهم‌ترین موارد، تشخیص درست و زودهنگام آن است. در این راستا نظر بر آن شد تا گامی در جهت کمک به تشخیص عفونت ریه در این همه‌گیری جهانی برداشته شود.

هدف از انجام این پروژه، تشخیص نواحی عفونی (درگیر عفونت) و غیرعفونی است. در این پروژه، از ویژگی‌های رادیومیکس استفاده می‌شود. رادیومیکس یک رویکرد نوظهور در تصویربرداری پزشکی است که از تجزیه و تحلیل‌های ریاضی پیشرفته استفاده می‌کند.

حال هدف بر آن است تا با اجرای الگوریتم‌های انتخاب ویژگی، بتوان تشخیص داد که کدام ویژگی‌ها در جهت افتراق بین ناحیه‌ی عفونی و اطراف آن مؤثرتر هستند.

در این راستا از مجموعه داده‌ای شامل تصاویر توموگرافی ۶۰ بیمار درگیر ویروس کرونا در بیمارستان غیائی تهران استفاده می‌شود. این مجموعه داده شامل تصاویر اصلی و تصاویر ماسک برچسب‌گذاری شده توسط رادیولوژیست است. پس از به دست آوردن ویژگی‌های رادیومیکس، با اجرای الگوریتم‌های مختلف انتخاب ویژگی و مقایسه‌ی نتایج حاصل از اجرای آن‌ها، الگوریتم‌های ElasticNet و XGBoost به عنوان قدرتمندترین الگوریتم‌ها در این زمینه انتخاب شدند.

واژگان کلیدی: انتخاب ویژگی، تصاویر توموگرافی، عفونت ریه، ویروس کرونا، ویژگی‌های رادیومیکس،

هوش مصنوعی

فهرست مطالب

عنوان	صفحه
فصل اول مقدمه.....	۵
۱-۱ هدف پروژه.....	۵
۲-۱ کاربردهای پروژه.....	۶
۳-۱ ساختار پایان نامه.....	۶
فصل دوم مفاهیم.....	۷
۱-۲ مقدمه.....	۷
۲-۲ عفونت در ریه.....	۷
۳-۲ رادیومیکس.....	۱۱
۴-۲ انتخاب ویژگی.....	۲۴
۵-۲ جمع بندی.....	۲۹
فصل سوم شرح پروژه.....	۳۰
۱-۳ مقدمه.....	۳۰
۲-۳ روش پیشنهادی.....	۳۰
۳-۳ جمع بندی.....	۳۶
فصل چهارم نتایج.....	۳۷
۱-۴ مقدمه.....	۳۷
۲-۴ مجموعه داده ها.....	۳۷
۳-۴ معیارهای ارزیابی.....	۳۷
۴-۴ ابزارها و کتابخانه ها.....	۳۹
۵-۴ نتایج.....	۴۱
۶-۴ ارزیابی نتایج.....	۴۶
۷-۴ جمع بندی.....	۴۸
فصل پنجم نتیجه گیری و پیشنهادها.....	۴۹
منابع.....	۵۱

فهرست شکل‌ها

صفحه	عنوان
۱۰.....	شکل ۱-۲: کووید-۱۹.....
۱۴.....	شکل ۲-۲: دستگاه CT.....
۱۵.....	شکل ۳-۲: دستگاه MRI.....
۱۵.....	شکل ۴-۲: دستگاه PET.....
۳۰.....	شکل ۱-۳: چارچوب روش پیشنهادی.....
۴۲.....	شکل ۱-۴: نمودار حاصا از اجرای الگوریتم LASSO.....
۴۳.....	شکل ۲-۴: نمودار حاصل از اجرای الگوریتم TreeClassifier.....
۴۴.....	شکل ۳-۴: نمودار حاصل از اجرای الگوریتم ElasticNet.....
۴۵.....	شکل ۴-۴: نمودار حاصل از اجرای الگوریتم XGBoost.....
۴۷.....	شکل ۵-۴: مقایسه‌ی میزان هم‌پوشانی الگوریتم‌های انتخاب ویژگی.....

فهرست جدول‌ها

صفحه	عنوان
جدول ۴-۱: ماتریس درهم‌ریختگی.....	۳۸
جدول ۴-۲: نتایج حاصل از اجرای الگوریتم LASSO.....	۴۱
جدول ۴-۳: نتایج حاصل از اجرای الگوریتم TreeClassifier.....	۴۲
جدول ۴-۴: نتایج حاصل از اجرای الگوریتم ElasticNet.....	۴۳
جدول ۴-۵: نتایج حاصل از اجرای الگوریتم XGBoost.....	۴۵
جدول ۴-۶: میانگین F1 Score محاسبه شده در الگوریتم‌های انتخاب ویژگی.....	۴۶

PCR	Polymerase Chain Reaction
CT scan	Computer Tomography scan
AI	Artificial Intelligence
IPF	Idiopathic Pulmonary Fibrosis
3D	Three-dimensional
CPU	Central Processing Unit
GPU	Graphics Processing Unit
NLP	Natural Language Processing
MRI	Magnetic Resonance Images
PET	Positron Emission Tomography
ROI	Region Of Interest
IBSI	Image Biomarker Standardization Initiative
GLCM	Gray Level Co-occurrence Matrix
GLSZM	Gray Level Size Zone Matrix
GLRLM	Gray Level Run Length Matrix
NGTDM	Neighbouring Gray Tone Difference Matrix
GLDM	Gray Level Dependence Matrix
MAD	Mean Absolute Deviation
RMAD	Relative Madian Absolute Deviation
RMS	Root Mean Square
MCC	Maximal Correlation Coefficient
IDN	Inverse Diffrence Normalized
SAE	Small Area Emphasis
LAE	Larg Area Emphasis
GLN	Gray Level Non-Uniformity
SZN	Size-Zone Non-Uniformity
ZP	Zone Percentage
SRE	Short Run Emphasis
LRE	Long Run Emphasis
OLS	Ordinary Least Squares
PLS	Partial Least Squares
SSV	Separability of Split Value

فصل اول

مقدمه

۱-۱ هدف پروژه

تشخیص به موقع و درست عفونت^۱ ریه یکی از موارد مهمی است که امروزه نیز به دلیل همه‌گیری جهانی کووید-۱۹^۲ از اهمیت ویژه‌ای برخوردار شده است. در این راستا نظر بر آن شد تا گامی در جهت کمک به تشخیص عفونت ریه در این همه‌گیری جهانی برداشته شود. بیماری ناشی از کووید-۱۹ یک بیماری تنفسی است. بیشتر موارد ویروس کرونا^۳ از طریق آزمایش واکنش زنجیره‌ای پلیمراز^۴ تشخیص داده می‌شود. علاوه بر آن، از تصاویر توموگرافی کامپیوتری یا سی‌تی‌اسکن^۵ نیز برای تشخیص احتمال ابتلای فرد و همچنین تشخیص میزان عفونت در ریه‌ی فرد استفاده می‌شود. این تصاویر CT باید توسط متخصصین رادیولوژیست^۶ مورد بررسی قرار گیرند تا میزان عفونت و نواحی دارای عفونت در آن‌ها تشخیص داده شود. حال به دلیل آنکه ممکن است رادیولوژیست دچار خطا شود و یا بررسی تعداد بالای تصاویر توسط متخصصین رادیولوژیست زمان‌بر باشد و علاوه بر این موارد، دانش افراد متخصص در این زمینه به یک میزان نباشد، تصمیم بر آن شد تا این کار با استفاده از ویژگی‌های رادیومیکس^۷ در جهت افتراق بین ناحیه‌ی عفونی و اطراف آن صورت گیرد.

تمامی مطالعات و پژوهش‌هایی که تاکنون درمورد استفاده از رادیومیکس و ویژگی‌های آن برای ریه صورت گرفته، پیرامون تشخیص تومور سرطانی و نوع عفونت ریه بوده است؛ درحالی‌که هدف از انجام این پروژه، تشخیص و جداسازی نواحی عفونی و غیرعفونی ریه است و بررسی خواهد شد که ویژگی‌های رادیومیکس تا چه میزان می‌توانند به افتراق بین ناحیه‌ی عفونی و اطراف آن کمک کنند.

¹ Infection

² COVID-19

³ Coronavirus

⁴ Polymerase Chain Reaction (PCR)

⁵ Computer Tomography scan (CT scan)

⁶ Radiologist

⁷ Radiomics

البته این نکته را باید در نظر داشت که استفاده از ویژگی‌های رادیومیکس در جهت افتراق بین ناحیه‌ی عفونی و اطراف آن برای عفونت‌های ریه، غیر از کووید-۱۹ نیز کاربرد دارد، اما در این پروژه به دلیل اهمیت مسئله‌ی همه‌گیری این بیماری در جهان و در اختیار داشتن تصاویر CT بیماران دارای کووید-۱۹، تمرکز ما بیشتر بر روی بررسی این نوع عفونت خواهد بود.

۲-۱ کاربردهای پروژه

تشخیص عفونت در ریه باتوجه به شیوع ویروس کرونا در جهان، یک امر ضروری و مهم می‌باشد. تشخیص دیر هنگام و نادرست ناحیه‌ی عفونت در ریه، موجب به‌وجود آمدن علائم زیادی از جمله ذات‌الریه و دشوار شدن تنفس، نارسایی عضوهای مختلف، مشکلات قلبی و تشکیل لخته خون در عروق می‌شود. این پروژه می‌تواند به‌عنوان ابزاری در جهت کمک به رادیولوژیست‌ها استفاده شود و از به‌وجود آمدن خطا در افتراق بین دو ناحیه‌ی دارای عفونت و اطراف آن جلوگیری کند.

۳-۱ ساختار پایان‌نامه

ابتدا در فصل دوم به بررسی مفاهیم اولیه در رابطه با عفونت در ریه، رادیومیکس، انتخاب ویژگی و پژوهش‌های انجام شده در این زمینه‌ها پرداخته می‌شود. در فصل سوم، روش پیشنهادی که شامل چهار گام اصلی آماده‌سازی داده‌ها جهت پردازش، ساخت تصاویر ماسک جدید، استخراج ویژگی‌های رادیومیکس و انتخاب ویژگی است، بررسی و به شرح مراحل موجود در هر یک از این گام‌ها پرداخته می‌شود. در فصل چهارم نیز به معرفی مجموعه‌داده‌ها، معیارهای ارزیابی و ابزارها و کتابخانه‌ها پرداخته شده و علاوه بر آن نیز به بررسی نتایج حاصل از روش پیشنهادی و ارزیابی آن‌ها پرداخته می‌شود. در نهایت، در فصل پنجم نیز نتیجه‌گیری و پیشنهادهایی برای آینده بیان می‌شود.

فصل دوم

مفاهیم

۱-۲ مقدمه

در بخش اول این فصل ابتدا توضیح مختصری در رابطه با عفونت ریه داده می‌شود. جهت آشنایی و بررسی عفونت ریه، ابتدا لازم است تا با بخش‌بندی‌های ریه و انواع عفونت آشنایی پیدا کرد. در ادامه توضیحاتی درباره‌ی بیماری کووید-۱۹ ارائه می‌شود. بیماری که امروزه سرتاسر جهان را فراگرفته و هرروزه صحبت از آن بیشتر می‌شود.

در بخش دوم، به بررسی رادیومیکس و آشنایی با ویژگی‌های مختلف آن پرداخته می‌شود. علاوه‌برآن، ارتباط رادیومیکس و هوش مصنوعی^۱ نیز مورد بحث و بررسی قرار می‌گیرد. در آخر نیز، ابتدا به تعریف انتخاب ویژگی^۲ و سپس به معرفی روش‌ها و الگوریتم‌های آن پرداخته می‌شود.

۲-۲ عفونت در ریه

هنگامی که قسمتی از بافت ریه و کیسه‌های راه‌های هوایی التهاب یابند و در آن‌ها خلط و ترشحات، مسیر تنفسی را ببندد و علاوه‌برآن در تنفس فرد مشکل ایجاد کند، عفونت ریه رخ داده است. عفونت ریه ممکن است در اثر حضور سازواره‌های^۳ مختلفی در ریه صورت گیرد. این سازواره‌ها می‌توانند شامل انواع ویروس‌ها، انگل‌ها و باکتری‌ها باشند [۱].

سازواره‌های کوچک از طریق جریان خون به ریه‌ها رسیده و سبب ایجاد عفونت می‌گردند؛ اما در اغلب اوقات سازواره‌ها توسط قطرات ریز آب موجود در هوای تنفسی یا همان قطرات تنفسی که از فردی به فرد دیگری منتقل می‌شوند به ریه‌ها منتقل می‌گردند [۱].

عفونت به وجود آمده در ریه معمولاً کیسه‌های هوایی که در ریه موجود می‌باشند را درگیر کرده و

^۱ Artificial Intelligence (AI)

^۲ Feature Selection

^۳ Organisms

مجاری هوایی انتقال‌دهنده‌ی هوا به ریه‌ها را به میزان کمتری درگیر می‌کند. بخشی از ریه که دچار عفونت بسیار شدیدی شده باشد، توسط مایعی پر می‌گردد. این مایع حاوی گلبول‌های سفید می‌باشد که برای مقابله با عفونت تولید شده‌اند.

دمای معمولی بدن افراد معمولاً در حدود ۳۷ درجه سانتی‌گراد می‌باشد که اگر فرد دچار عفونت‌های باکتریایی در ریه شده باشد، ممکن است دمای بدن او تا حدود ۴۱ درجه سانتی‌گراد نیز بالا رود که این میزان تب بالا بسیار خطرناک است [۱].

شیوع عفونت ریه همان گونه که اشاره شد بسیار بالا می‌باشد. هر ساله حدود یک درصد از افراد بالغ گرفتار این بیماری می‌شوند و ممکن است برخی از آن‌ها نیز دچار مرگ شوند. امروز به دلیل شیوع ویروس کرونا این مقدار افزایش یافته است [۱].

۲-۱-۲- ریه و بخش‌بندی‌های آن

ریه‌ها ارگان‌هایی هستند که عمده‌ترین نقش را در تنفس دارند. آن‌ها در سینه و هر دو طرف میان‌سینه^۱ قرار گرفته‌اند. ریه‌ها تقریباً مخروطی شکل هستند، با یک قله، قاعده، سه سطح و سه لبه. ریه سمت چپ به دلیل وجود قلب کمی کوچک‌تر از ریه راست است [۲].

ریه‌ها از لوب‌ها تشکیل شده‌اند که این لوب‌ها باعث می‌شوند تا ریه به بخش‌هایی تقسیم شود که در سطح علمی و پزشکی، مطالعه و درک آن را تسهیل می‌کند. لوب‌ها توسط حفراتی به داخل ریه‌ها نفوذ کرده و از هم جدا می‌شوند. ساختار لوب‌های ریه‌ی راست و چپ یکسان نیست. ریه‌ی راست سه لوب فوقانی، میانی و تحتانی دارد. این لوب‌ها با دو شکاف لوبار (شکاف مایل و شکاف افقی) از یکدیگر جدا شده‌اند. ریه‌ی چپ نیز دارای لوب فوقانی و تحتانی است که توسط شکاف مایل از هم جدا می‌شوند [۲].

بنابراین ریه‌ی انسان از نظر آناتومیکی^۲ و عملکردی به پنج قسمت مجزا به نام لوب تقسیم می‌شود که با شکاف‌های لوبار از هم جدا می‌شوند. از آن جا که سیستم راه هوایی و درخت رگ به طور مستقل در هر لوب توزیع شده است، بسیاری از بیماری‌های ریوی اغلب بر سطح لوبار خاصی تأثیر می‌گذارند [۳].

آمفیژم^۳، سل^۴ و سیلیکوز^۵ معمولاً لوب‌های فوقانی را تحت تأثیر قرار می‌دهند؛ درحالی‌که فیروز ریوی ایدیوپاتیک^۶ معمولاً با لوب‌های تحتانی همراه است. علاوه بر این، گزینه‌های درمانی شامل جراحی سرطان ریه و جراحی کاهش ریه در سطح لوبار انجام می‌شود. تقسیم‌بندی لوب‌های ریه در CT نقش مهمی در اندازه‌گیری عملکرد ریه و تجزیه و تحلیل تصویر منطقه‌ای شامل تجزیه و تحلیل بافت و شکل در سطح لوبار و

¹ Mediastinum

² Anatomy

³ Emphysema

⁴ Tuberculosis

⁵ Silicosis

⁶ Idiopathic Pulmonary Fibrosis (IPF)

برنامه‌ریزی جراحی دارد [۳].

۲-۲-۲- دلایل و منشأ به وجود آمدن عفونت‌های ریه

منبع این سازواره‌های کوچک می‌تواند افرادی باشد که به دلیل بیماری خود با سرفه و عطسه سبب انتشار قطرات تنفسی خود به دیگران می‌شوند. برخی مواقع این میکروب‌ها از مکان‌های خاصی منشأ می‌گیرند. به عنوان مثال سازواره‌هایی که سبب ایجاد یک نوع عفونت ریوی خاص به نام پنومونی لژیونلایی^۱ می‌گردند، سازواره‌هایی هستند که در آب‌های گرم زندگی می‌کنند [۱].

پسیتاکوز^۲ را می‌توان به عنوان مثال دیگری نام برد که معمولاً از پرندگان به خصوصى مانند طوطی‌ها منشأ می‌گیرد. همچنین استافیلوکوک‌ها^۳ نیز معمولاً در بیمارستان‌ها یافت می‌شوند و می‌توانند در افرادی که به علت‌های متفاوت در بیمارستان پذیرش شده‌اند ایجاد عفونت ریه کنند [۱].

ریه‌ی افرادی که دارای مشاغل خاص هستند ممکن است در هنگام کار با موادی که در محیط کاری آن‌ها قرار دارد، دچار التهاب شود. به عنوان مثال می‌توان برخی از کشاورزان را نام برد که در زمان تنفس گردوغبارهای حاصل از غلات کپک‌زده ممکن است دچار التهابات ریوی گردند.

۲-۲-۳- علائم عفونت ریه

علائم عفونت ریه معمولاً به صورت ناگهانی بروز می‌کند. از علائم شایع عفونت ریه می‌توان به موارد زیر اشاره کرد:

- تنگی نفس
- درد قفسه‌ی سینه
- سرفه و تب‌لرزه‌های گاه‌گاهی

علائم عفونت ریه می‌تواند از علائم و نشانه‌های خفیف تا نشانه‌های شدید متفاوت باشد. بروز علائم و شدت آن‌ها به عوامل مختلفی بستگی دارد که می‌توان از این عوامل به سن فرد، سلامتی عمومی بدن وی و این که آیا بیماری توسط باکتری، ویروس و یا قارچ ایجاد شده، بستگی دارد. علائم ممکن است مانند علائم ساده شبیه به علائم سرماخوردگی و آنفولانزا باشند؛ اما تمایل به ماندگاری بیشتری در بدن دارند [۱].

۲-۲-۴- کووید ۱۹

کووید ۱۹- با نام علمی SARS-CoV-2 در اواخر دسامبر سال ۲۰۱۹ در ووهان^۴ چین آغاز شد که در شکل ۱-۲ اجتماع تعدادی از این ویروس‌ها نشان داده شده است. این بیماری از طریق استنشاق یا تماس با

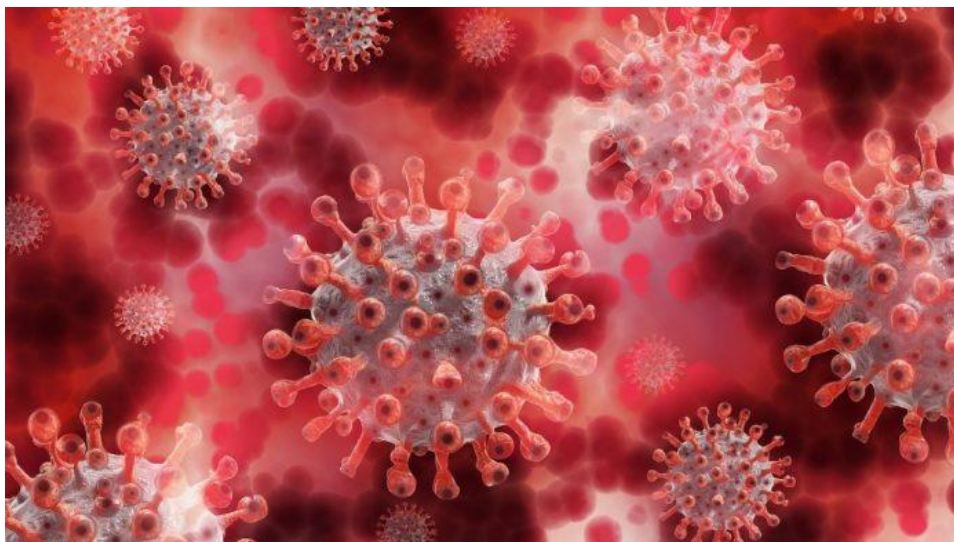
^۱ Legionella pneumophila

^۲ Psitacosis

^۳ Staphylococci

^۴ Wuhan

قطرات عفونی منتقل می‌شود. خوشبختانه ذراتی که در سیستم تنفسی ایجاد می‌کند تقریباً کمی سنگین هستند و هنگام عطسه و سرفه‌ی فرد بیمار، ذرات در فاصله‌ی یک تا یک و نیم متری رسوب می‌کنند و با رعایت حد فاصله‌ی دو متر، می‌توان از انتقال آن‌ها جلوگیری کرد [۴].



شکل ۲-۱: کووید-۱۹ [۵]

دوره‌ی نهفتگی این بیماری از ۲ تا ۱۴ روز متغیر است. علائم این بیماری معمولاً تب، سرفه، گلودرد، تنگی نفس، خستگی و ضعف است. کووید-۱۹ می‌تواند باعث بروز عوارض ریوی مانند ذات‌الریه^۱ و در موارد شدیدتر، سندروم دیسترس تنفسی حاد^۲ شود. سپسیس^۳ نیز یکی دیگر از عوارض احتمالی کووید-۱۹ است که می‌تواند به ریه‌ها و سایر اندام‌ها آسیب دائمی بزند [۴].

هنگامی که فرد به کووید-۱۹ مبتلا می‌شود، سیستم ایمنی بدن او برای مبارزه با مهاجم به‌صورت سخت کار می‌کند. این موضوع می‌تواند بدن را در برابر عفونت با یک باکتری یا ویروس دیگر در معرض خطر قرار دهد. عفونت اضافی نیز منجر به آسیب بیشتر ریه می‌شود.

سازمان بهداشت جهانی گزارش داد که اغلب بیماران مبتلا به کووید-۱۹ شکل خفیف تا متوسط عفونت را داشته‌اند که این موارد نیازی به بستری شدن ندارند. این موارد به این دلیل است که سیستم ایمنی بدن در پاسخ قادر به حفظ ویروس در دستگاه تنفسی فوقانی است. بیماران جوان نسبت به بیماران مسن، سیستم ایمنی قوی‌تری دارند.

موارد شدید و وخیم به علت نشت ویروس به نای و ورود آن به دستگاه تنفسی تحتانی است. مشکلات تنفسی مانند ذات‌الریه در این زمان به وجود می‌آید.

^۱ Pneumonie

^۲ ARDS

^۳ Sepsis

ذات‌الریه با تنگی نفس همراه با سرفه مشخص می‌شود و بر کیسه‌های هوایی ریزی که آلوئول^۱ نام دارند تأثیر می‌گذارد. آلوئول‌ها جایی هستند که اکسیژن و دی‌اکسیدکربن ردوبدل می‌شود. در ذات‌الریه، لایه‌ی نازک سلول‌های آلوئولی توسط ویروس آسیب می‌بینند. بدن با ارسال سلول‌های ایمنی به ریه برای مقابله با آن واکنش نشان می‌دهد و این منجر به ضخیم شدن پوشش‌ها نسبت به حالت طبیعی می‌شود. با ضخیم شدن پوشش‌ها، کیسه‌های هوایی کوچکی که برای رسیدن اکسیژن خون لازم است، دچار انسداد می‌شوند. از عواملی که احتمال بهبودی و بازیابی عملکرد ریه در فرد را افزایش می‌دهد، می‌توان به تشخیص زودهنگام و انجام اقدامات به‌موقع در بیمارستان برای بیماران اشاره کرد. انجام به‌موقع اقدامات لازم باعث به حداقل رسیدن آسیب به ریه‌ها می‌شود.

۲-۳ رادیومیکس

مانند بسیاری دیگر از زمینه‌های فعالیت بشری در دهه‌های گذشته، پزشکی نیز همواره در حال افزایش دیجیتالی کردن اطلاعات تولید شده در طی روال بالینی بوده است. علاوه بر اینکه داده‌های پزشکی بیشتری در قالب دیجیتالی در دسترس قرار گرفت و نرم‌افزارهای جدید و پیچیده‌تری نیز برای تجزیه و تحلیل آن‌ها ایجاد شد. در این بین تحقیقات در زمینه‌ی هوش مصنوعی نیز به جایی رسیده که روش‌ها و ابزارهای نرم‌افزاری آن نه تنها قدرتمند شده‌اند، بلکه به اندازه‌ی کافی در دسترس هستند تا از بخش‌های علوم کامپیوتر خارج شده و برنامه‌های کاربردی را در دامنه‌های گسترده‌تری پیدا کنند. این موارد در سال‌های اخیر نیز بیشتر مشاهده می‌شود [۶].

باهدف تسهیل کارهای تکراری که پزشکان در جریان کارهای بالینی روزانه‌ی خود با آن روبه‌رو می‌شوند، کاربردهای هوش مصنوعی در این زمینه افزایش یافته است. روش‌های مختلف مورد استفاده در حوزه‌ی هوش مصنوعی، به‌ویژه یادگیری ماشین^۲ و یادگیری عمیق^۳، هنگامی که به حوزه‌ی داده‌های بزرگ^۴ می‌آیند، بسیار مفید هستند. داده‌های بزرگ به معنای دارایی‌های اطلاعاتی است که حجم بالایی دارند و با سرعت زیاد تولید می‌شوند. این داده‌ها به تکنیک‌ها و فناوری‌های پیشرفته‌ای نیاز دارند تا بتوانند اطلاعات را ذخیره، توزیع، مدیریت و تجزیه و تحلیل کنند. به دلیل حجم بالای اطلاعات چندبعدی، روش‌های حوزه‌ی هوش مصنوعی برای استخراج اطلاعات از داده‌های بزرگ مورد نیاز است [۶].

در پزشکی، روش‌های مختلفی برای ایجاد داده‌های بزرگ وجود دارد. از جمله زمینه‌های شناخته شده می‌توان به ژنومیک^۵، پروتئومیک^۱ و متابولومیک^۲ اشاره کرد [۶]. پسوند "omics" اصطلاحی است که در

^۱ Alveolus

^۲ Machine Learning

^۳ Deep Learning

^۴ Big Data

^۵ Genomics

رشته‌های زیست‌شناسی مولکولی^۳ برای توصیف ویژگی‌های دقیق مولکول‌های زیست‌شناسی ایجاد شده است. در حال حاضر این واژه در سایر زمینه‌های تحقیقات پزشکی که داده‌های پیچیده با ابعاد بالا را از اجسام یا نمونه‌های واحد تولید می‌کنند، استفاده می‌شود. یکی از ویژگی‌های مطلوب داده‌های omics این است که این داده‌ها قابل استخراج هستند؛ بنابراین می‌توانند برای اکتشاف و ایجاد فرضیه استفاده شوند [۷].

مشابه این گروه‌های omics از تصویربرداری، می‌توان به گروه دیگری به نام رادیومیکس اشاره کرد. رادیومیکس، یک رویکرد کمی برای تصویربرداری پزشکی است که با استفاده از تجزیه و تحلیل‌های ریاضی پیشرفته، باعث افزایش داده‌های موجود و در دسترس برای پزشکان می‌شود [۶]. به عبارت دیگر، با افزایش تعداد ابزارهای تشخیص الگو و افزایش اندازه‌ی مجموعه داده‌های تصاعدی، حوزه‌ی تجزیه و تحلیل تصاویر پزشکی به صورت تصاعدی رشد کرده است. این پیشرفت‌ها موجب افزایش توان در استخراج ویژگی‌ها و تبدیل تصاویر به داده‌های قابل استخراج، جهت انجام تجزیه و تحلیل روی آن‌ها گردید که به این عمل رادیومیکس گفته می‌شود. این روش، برخلاف روش سنتی پردازش تصاویر پزشکی است. در روش سنتی، از تفسیر بصری تصاویر بهره‌گیری می‌شود [۷].

این مفهوم که به طور وسیع (اما نه منحصرأ) در زمینه‌ی سرطان‌شناسی به کار گرفته شده است، بر این فرض استوار است که تصاویر زیست‌پزشکی حاوی اطلاعاتی از فرایندهای خاص بیماری هستند که برای چشم انسان نامحسوس می‌باشند و در نتیجه از طریق بازرسی بصری تصاویر ایجاد شده، قابل دسترس نیستند [۶]. رادیومیکس از طریق استخراج ریاضی، توزیع فضایی شدت سیگنال، روابط متقابل بین پیکسل‌ها^۴ و همچنین با استفاده از روش‌های حوزه‌ی هوش مصنوعی، اطلاعات بافتی را کمی می‌کند. علاوه بر این، تفاوت‌های قابل ملاحظه‌ی بصری در شدت، شکل یا بافت تصاویر را می‌توان با استفاده از رادیومیکس تعیین کرد [۶].

داده‌های رادیومیکس شامل آمار درجه اول، دوم و مرتبه‌ی بالاتر است. این داده‌ها با سایر داده‌های بیمار ترکیب شده و با ابزارهای بیوانفورماتیکی^۵ پیچیده استخراج می‌شوند تا مدل‌هایی را توسعه دهند که دقت تشخیصی، پیش‌آگهی و پیش‌بینی را بهبود بخشد [۷].

مفهوم omics به آسانی در مورد تصویربرداری کمی توموگرافی در سطوح مختلف اعمال می‌شود. یک تصویر چند قسمتی یا سه‌بعدی^۶ از یک بیمار به راحتی می‌تواند میلیون‌ها وکسل^۷ را شامل شود. وکسل،

¹ Proteomics

² Metabolomics

³ Molecular Biology

⁴ Pixels

⁵ Bioinformatic

⁶ Three-dimensional (3D)

⁷ Voxel

کوچک‌ترین جز ساختاری^۱ یک تصویر سه‌بعدی است. در تجزیه و تحلیل‌های رادیومیکس نشانگرهای مولکولی و دیگر نشانگرهای زیستی برای پیش‌بینی درمان مناسب در زمان مناسب استفاده می‌شوند [۷].

رادیومیکس به معنای اتوماسیون فرایندهای تشخیصی نیست، بلکه داده‌های اضافی را در اختیار موارد موجود قرار می‌دهد. تجزیه و تحلیل رادیومیکس را می‌توان بر روی تصاویر پزشکی با روش‌های مختلف انجام داد [۶].

۲-۳-۱- رادیومیکس و هوش مصنوعی

هوش مصنوعی ماشین‌ها را قادر می‌سازد که هوش انسان را تقلید کنند و در نتیجه اثر پایین‌دست در تبدیل صنایعی مانند پزشکی دارد. یادگیری عمیق زیرمجموعه‌ای از یادگیری ماشین است. یادگیری ماشین نیز یک زیرمجموعه از هوش مصنوعی است [۸].

هوش مصنوعی به دلیل پیشرفت در واحدهای پردازش مرکزی^۲ و واحدهای پردازش گرافیکی^۳ و همچنین در دسترس بودن مجموعه داده‌های بزرگ، موفقیت بی‌سابقه‌ای در زمینه‌های مختلف داشته است. علاوه بر این، روش‌های هوش مصنوعی پیشرفت قابل‌توجهی در زمینه‌ی تفسیر داده‌های پیچیده و چندبعدی در کاربردهای مختلف، از تجارت سهام گرفته تا بینایی رایانه‌ای^۴ و پردازش زبان طبیعی^۵ انجام داده‌اند. امروزه، هوش مصنوعی پیشرفت‌هایی در زمینه‌ی مراقبت‌های بهداشتی، کشف دارو، تشخیص بیماری، تشخیص بیمار و پیش‌بینی تصاویر پزشکی نیز داشته است [۸].

هوش مصنوعی در تحلیل تصاویر پزشکی می‌تواند به دو روش اعمال شود [۸]:

- ۱) ویژگی‌های رادیومیکس استخراج شده می‌توانند به عنوان ورودی الگوریتم‌های یادگیری ماشین استفاده شوند. از این ویژگی‌ها جهت بهبود طبقه‌بندی استفاده می‌شود.
- ۲) یک مجموعه از تصاویر به عنوان ورودی یک شبکه‌ی یادگیری عمیق، برای توسعه‌ی طبقه‌بندی‌ها باشد.

در حال حاضر روش‌های یادگیری عمیق در تشخیص، توصیف و نظارت بر سرطان و عفونت ریه اهمیت فراوانی دارند [۸].

۲-۳-۲- رادیومیکس و تصاویر پزشکی

رادیومیکس با داشتن توان بالای محاسباتی، می‌تواند به سرعت ویژگی‌های کمی بی‌شماری را از تصاویر

¹ Element

² Central Processing Unit (CPU)

³ Graphics Processing Unit (GPU)

⁴ Computer Vision

⁵ Natural Language Processing (NLP)

توموگرافی کامپیوتری، تصاویر تشدید مغناطیسی^۱ و تصاویر توموگرافی تابش پوزیترون^۲ استخراج کند [۷].

- **توموگرافی کامپیوتری:** توموگرافی کامپیوتری، یک آزمایش بدون درد و با استفاده از اشعه‌ی ایکس^۳ است که جزئیات ساختارهای داخل بدن را به تصویر می‌کشد. تصاویر مقطعی ایجاد شده در طی یک CT می‌توانند در چندین صفحه بازسازی شوند و حتی می‌توانند تصاویر سه‌بعدی ایجاد کنند. CT یکی از سریع‌ترین و دقیق‌ترین ابزارها برای بررسی قفسه‌ی سینه، شکم و لگن است. این تصاویر جزئیات دقیق و مقدماتی را در مورد انواع بافت ارائه می‌دهند. در شکل ۲-۲ نمونه‌ای از دستگاه CT نشان داده شده است.



شکل ۲-۲: دستگاه CT [۹]

- **تصاویر تشدید مغناطیسی:** این تصاویر با استفاده از یک روش پیشرفته‌ی تصویربرداری که با ایجاد میدان مغناطیسی به تولید تصویر می‌پردازد، تولید می‌شوند. در این روش با استفاده از یک آهنربای بسیار قوی، امواج رادیویی و یک رایانه‌ی قوی، تصاویری با جزئیات دقیق از اندام‌ها و بافت‌های بدن ایجاد می‌شود. در MRI برخلاف CT از اشعه‌ی ایکس برای تصویربرداری استفاده نمی‌شود. این موضوع باعث شده است تا MRI نسبت به سایر روش‌های تصویربرداری از ایمنی بیشتری برخوردار باشد. در شکل ۲-۳ این دستگاه نشان داده شده است.
- MRI بهترین روش تشخیصی برای تصویربرداری از بافت‌های نرم مانند غضروف، تاندون، رگ و عصب‌ها است. MRI نسبت به CT اطلاعات دقیق‌تری در مورد اندام‌های داخلی (بافت‌های نرم) مانند رباط‌ها را ارائه می‌دهد؛ اما برای ریه‌ها، تصاویر CT بهترین عملکرد را دارند.

^۱ Magnetic Resonance Images (MRI)

^۲ Positron Emission Tomography (PET)

^۳ X-ray



شکل ۲-۳: دستگاه MRI [۱۰]

- تصاویر توموگرافی تابش پوزیترون: این تصاویر با استفاده از روش برش‌نگاری با گسیل پوزیترون یا پت اسکن^۱ به دست می‌آید که روشی نوین در علوم تشخیصی، فیزیک پزشکی و به‌ویژه در پزشکی هسته‌ای است. پت اسکن سیستمی است که می‌تواند تصاویر مقطعی از توزیع پوزیترون انتشاری حاصل از داروهای رادیواکتیو تزریق شده در بدن را دریافت کند. این تصاویر حاوی اطلاعاتی راجع به فعالیت‌های بیوشیمیایی، متابولیسم‌های^۲ سلولی، اندام‌شناسی^۳ و آسیب‌شناسی^۴ اندام و بافت‌های مختلف بدن هستند. در حالی که CT و MRI، تنها تصاویر مربوط به ویژگی‌های آناتومیک اندام و بافت‌ها را در اختیار پزشک قرار می‌دهند. در شکل ۲-۴ می‌توان این دستگاه را مشاهده کرد.



شکل ۲-۴: دستگاه PET [۱۱]

امروزه دستگاه‌های تصویربرداری و توموگرافی برای تشخیص بیماری‌ها، کاربردهای بسیار وسیعی در امر پزشکی دارند. بدون استفاده از این تجهیزات، تشخیص و درمان بیماری‌ها مشکلات بسیار زیادی به همراه دارد و خطراتی را نیز برای بیماران از جهت درمان صحیح و دقیق ایجاد می‌نماید. از این‌رو رادیومیکس و ویژگی‌های

^۱ PET Scan

^۲ Metabolism

^۳ Physiology

^۴ Pathology

آن، در تشخیص بیماری‌ها از روی این تصاویر به پزشکان و رادیولوژیست‌ها کمک می‌کند. سالانه میلیون‌ها مورد جدید سرطان گزارش می‌شود. اکثر این بیماران تحت چندین CT، MRI و PET قرار می‌گیرند. در آینده ممکن است تفسیر تصویر برای همه‌ی این مطالعات با استفاده از رادیومیکس افزایش یابد. درحالی‌که رادیومیکس امکان تشخیص بهتر بیماران و بیماری‌های آن‌ها را از طریق کاربردهای جدید ژنومیک ایجاد می‌کند، به چالش‌ها در زمینه‌ی مدیریت داده‌ها نیز می‌افزاید [۷].

۲-۳-۳- رادیومیکس و ریه

در سراسر دنیا، سرطان ریه شایع‌ترین سرطان تشخیص داده‌شده است؛ بنابراین تشخیص زودهنگام آن به کنترل و درمان کمک فراوانی می‌کند. با افزایش بیماران درگیر سرطان ریه، میزان مطالعه بر روی تصاویر پزشکی بیماران افزایش یافت. نشان داده‌شده است که رادیومیکس، ویژگی‌های زیستی سریع و دقیق را برای پیش‌بینی خطر سرطان ریه، تشخیص، درمان، نظارت بر واکنش درمان و زیست تومور نشان می‌دهد. در این میان، ویژگی‌های رادیومیکس و روش‌های حوزه‌ی هوش مصنوعی موردبحث و بررسی قرار گرفته‌اند [۸].

۲-۳-۱- رابطه‌ی میان عفونت ریه و سرطان ریه

عفونت ریه و سرطان ریه هر دو در ریه‌ها ایجاد می‌شوند و علائم مشابهی با هم دارند. عفونت ریه منجر به مشکلات تنفسی و به‌وجودآمدن مایع در ریه‌ها می‌شود. ویروس‌ها، باکتری‌ها و قارچ‌های مختلف می‌توانند باعث به‌وجودآمدن عفونت ریه شوند. سرطان ریه به دلیل رشد بیش از حد سلول‌های ریه ایجاد می‌شود که می‌تواند تومور ایجاد کند. طبق گزارش‌ها، سرطان ریه یکی از شایع‌ترین سرطان‌ها است که منجر به مرگ افراد زیادی در جهان می‌شود.

سرطان ریه با تضعیف سیستم ایمنی، خطر ابتلا به عفونت ریه را افزایش می‌دهد؛ بنابراین عفونت ریه یکی از عوارض سرطان ریه تلقی می‌شود. معمولاً بین ۵۰ تا ۷۰ درصد از افرادی که سیستم ایمنی ضعیفی دارند و دچار سرطان ریه نیز هستند، مبتلا به عفونت ریه نیز می‌شوند. علاوه بر این، درمان‌های فشرده‌ای که پزشکان برای درمان سرطان ریه استفاده می‌کنند، اغلب عملکرد سیستم ایمنی را به‌شدت کاهش می‌دهد. این بدان معناست که افراد ممکن است کمتر بتوانند از ورود عوامل عفونی به بدن جلوگیری کنند و حتی ممکن است در زمان مبارزه با عفونت نیز دچار مشکلات بیشتری شوند و بدنشان به داروها خوب پاسخ ندهد. برای این افراد، عفونت یک خطر جدی است و تشخیص درست و زودهنگام آن می‌تواند برای آن‌ها تأثیر بسیار زیادی داشته باشد.

برخی از علائم سرطان ریه و عفونت ریه یکسان است. به طور معمول علائم عفونت ریه شدیدتر است. سرطان ریه به‌طورکلی کندتر از عفونت ریه توسعه می‌یابد و تا زمانی که به مرحله‌ی پیشرفته نرسد، علائمی نشان نمی‌دهد.

۲-۳-۴- نحوه‌ی انجام رادیومیکس

رادیومیکس شامل مراحل مجزایی است که هرکدام چالش‌های خاص خود را دارند. این مراحل به ترتیب در زیر توضیح داده شده‌اند.

۲-۳-۴-۱- تقسیم‌بندی تصویر

اولین گام در این مرحله، ایجاد منطقه‌ی موردنظر^۱ (ROI) است. ROI، ناحیه‌ای از تصویر است که برای تجزیه و تحلیل و یا پردازش بیشتر تعریف شده است. از ROI جهت افزایش بهینگی استفاده می‌شود. برای مثال تصاویر به دست آمده از CT ریه دارای ابعاد بزرگی هستند و به همین دلیل هنگام پردازش آن‌ها با تعداد پیکسل‌های زیادی مواجه هستیم. علاوه بر آن قسمتی که دارای عفونت می‌باشد و برای پردازش به آن نیاز است، منطقه‌ی محدودی را در بر دارد. از این رو بهتر است به جای آنکه تمام تصویر در نظر گرفته شود، تنها همان منطقه‌ی موردنیاز را مشخص کرده و پردازش روی آن انجام گیرد. ویژگی‌های رادیومیکس در ROIهای مشخص شده محاسبه می‌شوند [۶].

تقسیم‌بندی تصویر به دو صورت نیمه خودکار (دستی) و خودکار صورت می‌گیرد. تقسیم‌بندی نیمه خودکار با استفاده از الگوریتم‌های استاندارد تقسیم‌بندی تصویر مانند رشد منطقه^۲ و تقسیم‌بندی خودکار با استفاده از الگوریتم‌های یادگیری عمیق صورت می‌گیرد [۶].

تقسیم‌بندی نیمه خودکار اشکالاتی دارد. اولاً، زمان‌بر است. ثانیاً، تفکرات ناظر ممکن در انجام تقسیم‌بندی تأثیری ایجاد کند؛ بنابراین میزان استفاده از این روش کاهش یافته است. امروزه در بیشتر مطالعاتی که انجام می‌گیرد، از تقسیم‌بندی خودکار استفاده می‌شود و روزبه‌روز الگوریتم‌های مختلفی برای تقسیم‌بندی تصویر اندام‌های مختلف آموزش می‌بینند [۶].

۲-۳-۴-۲- پردازش تصویر^۳

دومین گام برای انجام پردازش‌ها از طریق رادیومیکس، پردازش تصویر است. در این مرحله سعی بر آن است تا تصاویر همگن شوند. طبق مطالعات انجام شده، این نتیجه حاصل شده است که نحوه‌ی پردازش تصاویر بر مقادیر به دست آمده از ویژگی‌های رادیومیکس، تأثیر فراوانی داشته است [۶].

۲-۳-۴-۳- استخراج ویژگی

مهم‌ترین گام در رادیومیکس، استخراج ویژگی‌ها است. فرمول‌های مختلفی برای محاسبه‌ی ویژگی‌ها در

^۱ Region Of Interest (ROI)

^۲ Region-growing

^۳ Image Processing

رادیومیکس وجود دارد و به همین دلیل بهترین کار، پیروی از دستورالعمل‌ها و فرمول‌های استانداردسازی شده توسط استانداردسازی نشانگر زیستی تصویر^۱ (IBSI) است [۶].

انواع مختلفی از ویژگی‌های رادیومیکس وجود دارد که بیشتر آن‌ها با ویژگی‌های مبتنی بر شدت، شکل، بافت و تغییر شکل ارتباط دارند. علاوه بر این، انواع مختلفی از فیلترها در این مرحله استفاده می‌شوند. به‌طور کلی می‌توان گفت که مرحله‌ی استخراج ویژگی در رادیومیکس، همان مرحله‌ی اجرا می‌باشد [۶].

۲-۳-۴- انتخاب ویژگی/کاهش ابعاد

آخرین گام، انتخاب ویژگی است. تعداد ویژگی‌های استخراج شده، باتوجه‌به توابع انتخاب شده در هنگام اجرا مشخص می‌شوند. هر چه تعداد ویژگی‌ها کمتر باشد، خطر بیش‌برازش^۲ را افزایش می‌دهد. بیش‌برازش به این معناست که الگوریتم فقط داده‌هایی که در مجموعه‌ی آموزشی^۳ یاد گرفته است را می‌تواند به‌درستی پیش‌بینی کند ولی اگر داده‌ها کمی از مجموعه‌ی آموزشی فاصله داشته باشند، الگوریتمی که دچار بیش‌برازش شده باشد، نمی‌تواند به‌درستی پاسخی برای این داده‌های جدید بیابد و باعث خطا در طبقه‌بندی‌ها می‌شود. در نتیجه این مرحله از اهمیت بسیار بالایی برخوردار است.

کاهش ابعاد، منجر به انتخاب ویژگی‌های غیر قابل تکرار، زائد و غیرمرتبط از مجموعه‌داده می‌شود. در این زمینه روش‌های مختلفی وجود دارد که در ادامه به طور مفصل بررسی شده است [۶].

۲-۳-۵- ویژگی‌های رادیومیکس

در این بخش به تعریف برخی از ویژگی‌های رادیومیکس که اکثر آن‌ها توسط IBSI تعریف شده‌اند، می‌پردازیم. این ویژگی‌ها را می‌توان به کلاس‌های زیر تقسیم‌بندی کرد [۱۲]:

- آمار مرتبه اول^۴
- ویژگی‌های مبتنی بر شکل سه‌بعدی^۵
- ویژگی‌های مبتنی بر شکل دوبعدی^۶
- ویژگی‌های ماتریس هم‌جواری سطح خاکستری^۷
- ویژگی‌های طول اجرای ماتریس سطح خاکستری^۸
- ویژگی‌های سطح خاکستری ماتریس منطقه^۱

^۱ Image Biomarker Standardization Initiative (IBSI)

^۲ Overfitting

^۳ Train set

^۴ First Order Statistics

^۵ Shape Features (3D)

^۶ Shape Features (2D)

^۷ Gray Level Co-occurrence Matrix (GLCM)

^۸ Gray Level Size Zone Matrix (GLSZM)

- ویژگی‌های ماتریس تفاوت رنگ خاکستری همسایه^۲
 - ویژگی‌های ماتریس وابستگی به سطح خاکستری^۳
- رادیومیکس در مجموع ۱۲۸ ویژگی برای تصاویری که به‌عنوان ورودی می‌پذیرد، محاسبه می‌کند. در ادامه، توضیح مختصری پیرامون کلاس‌های نام‌برده، آورده شده است [۱۲].

۲-۳-۵-۱- آمار مرتبه اول

آمارهای مرتبه اول، توزیع شدت و کسل‌ها را در نواحی تعریف‌شده‌ی تصویر، از طریق معیارهای متداول توصیف می‌کند [۱۲].

- (۱) انرژی: مقدار ارزش یک و کسل در تصویر را اندازه‌گیری می‌کند.
- (۲) انرژی کل: مجموع مقدار انرژی که بر اساس حجم و کسل‌ها مقیاس‌بندی شده است را محاسبه می‌کند.
- (۳) آنترپی^۴: مشخص‌کننده‌ی عدم قطعیت/ تصادفی بودن در مقادیر تصاویر است. در واقع آنترپی، متوسط میزان اطلاعات موردنیاز برای شکل‌دهی مقادیر تصاویر را اندازه‌گیری می‌کند.
- (۴) کمینه: حداقل مقدار را محاسبه می‌کند.
- (۵) دهمین صدک^۵: در آمار صدک به مقداری گفته می‌شود که درصد خاصی از نمونه‌های یک متغیر تصادفی کمتر از آن هستند. دهمین صدک یک متغیر، مقداری است که ۱۰ درصد مشاهدات این متغیر کمتر از آن است. ۲۵امین صدک یک متغیر چارک اول نام دارد.
- (۶) نودمین صدک: نودمین صدک یک متغیر، مقداری است که ۹۰ درصد مشاهدات این متغیر کمتر از آن است.
- (۷) بیشینه: بیشترین مقدار را محاسبه می‌کند.
- (۸) میانگین: میانگین مقدار متغیرها را محاسبه می‌کند.
- (۹) میانه^۶: جمعیت آماری را با یک توزیع احتمالی به دو قسمت مساوی تقسیم می‌کند.
- (۱۰) محدوده‌ی بین چارک‌ها: محدوده‌ی بین چارک اول (۲۵امین صدک) و چارک سوم (۷۵امین صدک) را محاسبه می‌کند.
- (۱۱) محدوده: محدوده‌ی مقادیر خاکستری درون ROI را محاسبه می‌کند.

¹ Gray Level Run Length Matrix (GLRLM)

² Neighbouring Gray Tone Difference Matrix (NGTDM)

³ Gray Level Dependence Matrix (GLDM)

⁴ Entropy

⁵ Percentile

⁶ Median

- (۱۲) میانگین انحراف مطلق^۱: فاصله‌ی متوسط همه‌ی مقادیر از مقدار میانگین را محاسبه می‌کند.
- (۱۳) انحراف متوسط مطلق^۲: میانگین فاصله‌ی تمام مقادیر از مقدار متوسط را محاسبه می‌کند که در بین دهمین و نودمین صدک قرار دارند.
- (۱۴) میانگین مربعات^۳: پارامتری است که به اندازه‌ی سیگنال اشاره می‌کند. اندازه‌ی سیگنال برای نشان دادن قدرت سیگنال استفاده می‌شود.
- (۱۵) انحراف استاندارد: میزان پراکندگی مقدار متوسط را اندازه‌گیری می‌کند. انحراف معیار برابر با جذر واریانس است.
- (۱۶) کج شدن: میزان عدم تقارن توزیع متقارن را اندازه‌گیری می‌کند. این مقدار می‌تواند مثبت یا منفی باشد.
- (۱۷) کشیدگی^۴: معیاری از توزیع مقادیر در ROI است. هر چه مقدار کشیدگی بیشتر باشد، نشان‌دهنده‌ی این مطلب است که جرم توزیع به سمت دم‌های تصویر بیشتر متمرکز است و هر چه مقدار آن کمتر باشد، به این معناست که جرم توزیع به سمت میانگین متمرکز است.
- (۱۸) واریانس: مقدار متوسط مربع اختلاف مقادیر از میانگین را محاسبه می‌کند.
- (۱۹) یکنواختی: مجموع مربعات هر مقدار شدت را یکنواختی گویند. هر جا یکنواختی بیشتر باشد، به معنای همگنی بیشتر است.

۲-۵-۳-۲ ویژگی‌های مبتنی بر شکل سه‌بعدی

در این گروه از ویژگی‌ها، توصیف اندازه و شکل سه‌بعدی ROI اهمیت دارد. این ویژگی‌ها مستقل از توزیع شدت سطح خاکستری در ROI هستند و تنها روی تصویر و ماسک ورودی محاسبه می‌شوند. در این ویژگی‌ها از مش مثلثی بهره‌گیری می‌شود [۱۲].

- (۱) حجم مش: حجم ROI از مش مثلثی را محاسبه می‌کند.
- (۲) حجم وکسل: باتوجه به تعداد وکسل‌های درون یک ROI، حجم هر وکسل را محاسبه می‌کند.
- (۳) مساحت سطح: ابتدا مساحت هر مثلث مش را محاسبه کرده و در انتها تمام این مقادیر را باهم جمع می‌کند.
- (۴) نسبت سطح به حجم: می‌توان گفت که این ویژگی، میزان فشردگی را مشخص می‌کند. به‌طوری که هرچه این مقدار کمتر باشد، به معنای این است که شکل فشرده‌تر است. این

¹ Mean Absolute Deviation (MAD)

² Relative Median Absolute Deviation (RMAD)

³ Root Mean Square (RMS)

⁴ Kurtosis

ویژگی تا حدی به حجم ROI نیز بستگی دارد.

- (۵) کروی بودن: میزان گرد بودن شکل ناحیه‌ی تومور نسبت به یک کره را مشخص می‌کند. این اندازه‌گیری مستقل از مقیاس و جهت است. اگر مقدار این ویژگی برابر با ۱ باشد، به معنای وجود یک کره‌ی کامل است. این ویژگی با ویژگی فشردگی ارتباط دارد.
- (۶) فشردگی ۱: میزان فشردگی شکل تومور را نشان می‌دهد.
- (۷) فشردگی ۲: همانند فشردگی ۱ است. این اندازه‌گیری بدون بعد و مستقل از مقیاس و جهت است.
- (۸) عدم تناسب کروی: نسبت سطح ناحیه‌ی تومور به مساحت یک کره با حجم یکسان منطقه‌ی ناحیه‌ی تومور.

۲-۳-۵-۳- ویژگی‌های مبتنی بر شکل دوبُعدی

در این گروه از ویژگی‌ها نیز مانند ویژگی‌های مبتنی بر شکل سه‌بُعدی، توصیف اندازه و شکل دوبُعدی ROI اهمیت دارد. این ویژگی‌ها نیز مستقل از توزیع شدت سطح خاکستری در ROI هستند و تنها روی تصویر و ماسک ورودی محاسبه می‌شوند [۱۲].

تمامی ویژگی‌های این گروه مانند گروه قبلی است و تنها تفاوت آن‌ها در دوبُعدی و سه‌بُعدی بودنشان است؛ بنابراین از توضیح مجدد صرف‌نظر شده است.

۲-۳-۵-۴- ویژگی‌های ماتریس هم‌جواری سطح خاکستری

GLCM، عملکرد احتمال مرتبه دوم یک ناحیه‌ی تصویر را که توسط ماسک محدود شده است، توصیف می‌کند. GLCM به صورت $P(i, j | \delta, \theta)$ تعریف می‌شود. عنصر (i, j) این ماتریس، تعداد دفعاتی را نشان می‌دهد که ترکیب سطوح i و j در دو پیکسل در تصویر رخ می‌دهد که با فاصله‌ی δ پیکسل در امتداد زاویه‌ی θ از یکدیگر جدا می‌شوند [۱۲].

- (۱) همبستگی^۱ خودکار: میزان ظرافت و درشتی بافت را اندازه‌گیری می‌کند.
- (۲) میانگین مشترک: شدت سطح خاکستری را برمی‌گرداند.
- (۳) برجستگی خوشه‌ای: معیاری برای عدم تقارن GLCM است. هر چه مقدار این ویژگی بیشتر باشد، به معنای عدم تقارن بیشتر در GLCM است و هر چه این مقدار کمتر باشد، نشان‌دهنده‌ی نزدیکی آن به مقدار متوسط است.
- (۴) سایه خوشه‌ای: میزان کج بودن و یکنواختی GLCM را اندازه‌گیری می‌کند. هر چه مقدار این

^۱ Correlation

- ویژگی بالاتر باشد، نشان‌دهنده‌ی عدم تقارن بیشتر درمورد میانگین است.
- (۵) گرایش خوشه: وکسل‌ها با مقادیر مشابه خاکستری را اندازه‌گیری می‌کند.
- (۶) تضاد^۱: اندازه‌گیری تنوع شدت محلی که مقادیر دورتر از قطر ($i=j$) را ترجیح می‌دهد.
- (۷) همبستگی: مقدار این ویژگی، عددی بین صفر (بدون ارتباط) و یک (ارتباط کامل) است که وابستگی خطی مقادیر سطح خاکستری را به وکسل‌های مربوط در GLCM نشان می‌دهد.
- (۸) انرژی مشترک: انرژی الگوهای همگن در تصویر را اندازه‌گیری می‌کند. هر چه مقدار بیشتر باشد به معنای آن است که تعداد جفت‌های همسایه در فرکانس‌های بالا بیشتر هستند.
- (۹) آنتروپی مشترک: میزان تغییرپذیری در مقادیر شدت منطقه را اندازه‌گیری می‌کند.
- (۱۰) ضریب همبستگی بیشینه^۲: بیشترین میزان پیچیدگی بافت را MCC گویند.
- (۱۱) معکوس تفاوت نرمال شده^۳: معیاری برای همگن کردن محلی یک تصویر است. تفاوت بین مقادیر شدت را با تقسیم بر کل مقادیر شدت، نرمال می‌کند.
- (۱۲) واریانس معکوس: ارتباط بین دو متغیر را به‌وسیله‌ی پراکندگی‌هایشان نسبت به میانگین، نشان می‌دهد.
- (۱۳) مجموع آنتروپی: مجموع تفاوت‌های مقدار شدت منطقه را محاسبه می‌کند.

۲-۳-۵- ویژگی‌های طول اجرای ماتریس سطح خاکستری

- GLSZM، مناطق سطح خاکستری را در یک تصویر تعیین می‌کند. منطقه‌ی سطح خاکستری، تعداد وکسل‌های متصل شده که شدت سطح خاکستری یکسانی دارند را شامل می‌شود [۱۲].
- (۱) تکیه بر منطقه‌ی کوچک^۴: این ویژگی، معیاری برای توزیع مناطق کوچک است. هر چه مقدار این ویژگی بیشتر باشد، نشان‌دهنده‌ی کوچک‌تر بودن مناطق و ظریف‌تر بودن بافت‌ها است.
- (۲) تکیه بر منطقه‌ی بزرگ^۵: معیاری برای توزیع مناطق بزرگ است و بیشتر شدن مقدار این ویژگی، نشان‌دهنده‌ی بزرگ‌تر بودن مناطق و درشت‌تر بودن بافت‌ها است.
- (۳) یکنواختی سطح خاکستری^۶: GLN، میزان تنوع در شدت سطح خاکستری تصاویر را اندازه‌گیری می‌کند. هر چه مقدار این ویژگی پایین‌تر باشد، به معنای همگن‌تر بودن مقادیر شدت سطح خاکستری تصاویر است.

¹ Contrast

² Maximal Correlation Coefficient (MCC)

³ Inverse Diffrence Normalized (IDN)

⁴ Small Area Emphasis (SAE)

⁵ Larg Area Emphasis (LAE)

⁶ Gray Level Non-Uniformity (GLN)

(۴) عدم یکنواختی اندازه‌ی منطقه^۱: SZN، میزان تنوع حجم ناحیه‌های تصویر را اندازه‌گیری می‌کند.

(۵) درصد منطقه^۲: این ویژگی، با در نظر گرفتن نسبت تعداد مناطق و تعداد وکسل‌ها در ROI، میزان درشتی بافت را اندازه‌گیری می‌کند.

(۶) واریانس سطح خاکستری: میزان واریانس شدت سطح خاکستری را برای تمام مناطق اندازه‌گیری می‌کند.

۲-۳-۵-۶- ویژگی‌های سطح خاکستری ماتریس منطقه

GLRLM، تعداد سطوح خاکستری که به‌عنوان طول تعریف شده‌اند را از پیکسل‌های متوالی که دارای مقدار سطح خاکستری یکسانی هستند، تعیین می‌کند [۱۲].

(۱) اهمیت کوتاه‌مدت^۳: این ویژگی، توزیع طول‌های کوتاه‌مدت را اندازه‌گیری می‌کند. هر چه مقدار آن بیشتر باشد، نشان‌دهنده‌ی طول کوتاه‌تر و بافت‌های بهتر است.

(۲) اهمیت طولانی‌مدت^۴: این ویژگی نیز برخلاف ویژگی قبل، توزیع طول بلندمدت را اندازه‌گیری می‌کند و هر چه مقدار آن بیشتر باشد، حاکی از طولانی‌تر بودن طول و درشت‌تر بودن ساختار بافت است.

(۳) یکنواختی سطح خاکستری: میزان شباهت شدت سطح خاکستری تصاویر را اندازه‌گیری می‌کند.

(۴) اجرای درصد: با در نظر گرفتن نسبت تعداد اجرا و تعداد وکسل‌ها در ROI، درشتی بافت را اندازه‌گیری می‌کند.

(۵) واریانس سطح خاکستری: واریانس شدت سطح خاکستری را اندازه‌گیری می‌کند.

(۶) اجرای واریانس: معیار واریانس در طول اجرا است.

۲-۳-۵-۷- ویژگی‌های ماتریس تفاوت رنگ خاکستری همسایه

این ماتریس، میزان تفاوت بین مقدار خاکستری و مقدار متوسط همسایگان آن ناحیه را در فاصله‌ی δ تعیین می‌کند [۱۲].

(۱) درشتی: درشتی را محاسبه کرده و برمی‌گرداند.

(۲) تضاد: مقدار تضاد را محاسبه کرده و برمی‌گرداند.

^۱ Size-Zone Non-Uniformity (SZN)

^۲ Zone Percentage (ZP)

^۳ Short Run Emphasis (SRE)

^۴ Long Run Emphasis (LRE)

(۳) میزان تغییر از یک پیکسل به همسایه‌ی آن را محاسبه می‌کند.

(۴) پیچیدگی را محاسبه کرده و برمی‌گرداند.

(۵) میزان بازده را محاسبه می‌کند.

۲-۳-۵-۸- ویژگی‌های ماتریس وابستگی به سطح خاکستری

این گروه از ویژگی‌ها، میزان وابستگی‌های سطح خاکستری در یک تصویر را تعیین می‌کنند. تعداد بسیاری از ویژگی‌های این گروه مانند ویژگی‌های قبلی است و تنها تفاوت بسیار کمی در معیارهای اندازه‌گیری دارد. به همین دلیل به شرح مجدد آن‌ها پرداخته نشده است [۱۲].

۲-۴ انتخاب ویژگی^۱

الگوریتم‌های یادگیری ماشین، با استفاده از ویژگی‌های به‌دست‌آمده، می‌توانند فرایند طبقه‌بندی را انجام دهند. در سال‌های اخیر، مجموعه‌داده‌های متعددی با ابعاد بالا در دسترس عموم قرار گرفته است. این امر چالش‌هایی را برای جوامع پژوهشی در پی داشته است، زیرا سروکار داشتن با حجم زیادی از ویژگی‌ها برای انجام طبقه‌بندی با استفاده از الگوریتم‌های یادگیری ماشین، کاری دشوار است. از این‌رو، روش‌های مختلفی برای کاهش تعداد ویژگی‌های نامربوط و اضافی به وجود آمده است. انتخاب ویژگی، به کاهش نیاز به محاسبه‌های طولانی و بهبود عملکرد پیش‌بینی کمک می‌کند [۱۳].

انتخاب ویژگی، انتخاب زیرمجموعه‌ای از ویژگی‌های ورودی است که می‌تواند داده‌های ورودی را به‌صورت مؤثرتری توصیف کند. این کار باعث بهبود پیش‌بینی‌های ارائه شده می‌شود [۱۳].

در این پروژه، کاربرد انتخاب ویژگی می‌تواند در تجزیه و تحلیل ویژگی‌های رادیومیکس باشد. این بدان معناست که کل محتوای اطلاعات را می‌توان از ویژگی‌های منحصربه‌فرد کمتری که حاوی حداکثر اطلاعات مفید هستند، بدست‌آورد؛ بنابراین با حذف ویژگی‌های وابسته، می‌توان میزان داده‌ها را کاهش داد که این امر منجر به بهبود عملکرد طبقه‌بندی می‌شود.

با استفاده از روش‌های انتخاب ویژگی، می‌توان دقت پیش‌بینی^۲ را بهبود بخشید. برای حذف یک ویژگی نامربوط، یک معیار انتخاب ویژگی موردنیاز است که می‌تواند ارتباط هر ویژگی را با برچسب^۳ خروجی اندازه‌گیری کند. انتخاب ویژگی، یک مرحله‌ی لازم جهت پردازش داده‌های بزرگ است، زیرا منجر به کمینه کردن زمان آموزش و حافظه‌ی تخصیص داده‌شده می‌شود [۱۳].

^۱ Feature Selection

^۲ Prediction accuracy

^۳ Label

۲-۴-۱- روش‌های انتخاب ویژگی

روش‌های انتخاب ویژگی به صورت کلی به سه دسته تقسیم می‌شوند که در ادامه توضیح کوتاهی پیرامون هر یک از این روش‌ها داده شده است [۱۳].

- روش‌های فیلتری^۱: فیلترها از روش‌های رتبه‌بندی ویژگی‌ها به عنوان معیار اصلی برای انتخاب ویژگی استفاده می‌کنند. استفاده از روش‌های رتبه‌بندی بسیار ساده است، به همین دلیل این روش‌ها به شدت مورد استقبال قرار گرفته و موفقیت چشم‌گیری در کارهای عملی کسب کرده‌اند. فیلترها، بر ویژگی‌های کلی مجموعه داده‌ها تکیه دارند و فرایند انتخاب ویژگی را به عنوان یک گام در فرایند پیش‌پردازش^۲ انجام می‌دهند. مزیت این مدل‌ها، هزینه‌ی محاسباتی پایین و توانایی تعمیم خوب آن‌ها محسوب می‌شود. مشخصه‌ی اصلی یک ویژگی منحصر به فرد این است که حاوی اطلاعات مفیدی در مورد کلاس‌های طبقه‌بندی شده در داده‌ها باشد. یکی از تعاریفی که می‌توان برای بحث مفید بودن دانست، این است که اگر یک ویژگی به طور مشروط مستقل از برجسب کلاس باشد، بی‌ربط و غیرمفید تلقی می‌شود. به عبارتی دیگر، می‌توان گفت که یک ویژگی مفید می‌تواند مستقل از داده‌های ورودی باشد؛ اما نمی‌تواند مستقل از برجسب‌های کلاس‌ها باشد؛ بنابراین ویژگی‌هایی که بر برجسب‌های کلاس‌ها تأثیری ندارند را می‌توان کنار گذاشت [۱۳].
 - روش‌های پوششی^۳: روش‌های پوششی از پیش‌بینی‌کننده به عنوان یک جعبه سیاه و از عملکرد پیش‌بینی‌کننده به عنوان تابع هدف، برای ارزیابی زیرمجموعه‌ی ویژگی‌ها استفاده می‌کند [۱۳]. این دسته از روش‌های انتخاب ویژگی، در هر مرحله، زیرمجموعه‌ای از ویژگی‌ها را انتخاب می‌کند و الگوریتم یادگیری ماشین روی این زیرمجموعه سنجیده می‌شود. از نتیجه‌ی عملکرد الگوریتم یادگیری ماشین، برای ارزیابی زیرمجموعه‌ی انتخاب شده از میان تمام ویژگی‌ها استفاده می‌شود. به عنوان مثال، مشخص می‌شود که کدام یک از زیرمجموعه‌های انتخاب شده، بیشترین تأثیر را در افزایش دقت یک مسئله‌ی دسته‌بندی خواهند داشت [۱۳].
- مشکل اصلی روش پوششی، حجم محاسبات مورد نیاز برای به دست آوردن زیرمجموعه‌ای از ویژگی‌ها است. این روش برای به دست آوردن هر زیرمجموعه، یک مدل ایجاد می‌کند؛ یعنی پیش‌بینی‌کننده برای هر زیرمجموعه به صورت جداگانه آموزش می‌بیند و برای به دست آوردن دقت طبقه‌بندی آزمایش می‌شود. حال اگر تعداد نمونه‌ها زیاد باشد، زمان زیادی از الگوریتم

^۱ Filters

^۲ Preprocessing

^۳ Wrappers

صرف آموزش می‌شود. یکی دیگر از اشکالات این روش، استفاده از عملکرد طبقه‌بندی‌کننده به‌عنوان تابع هدف است. اگر مدل طبقه‌بندی‌کننده داده‌ها را به‌خوبی یاد بگیرد ولی قابلیت تعمیم آن ضعیف باشد، بیش‌برازش ایجاد می‌شود. این بیش‌برازش می‌تواند منجر به انتخاب نادرست یک زیرمجموعه از ویژگی‌ها شود [۱۳].

- تعبیه‌شده^۱: روش‌های تعبیه شده، فرایند انتخاب ویژگی را در فرایند آموزش انجام می‌دهند و معمولاً برای ماشین‌های یادگیری خاصی مورد استفاده قرار می‌گیرند. این روش‌ها قادر به ثبت وابستگی‌ها با هزینه‌های محاسباتی پایین‌تر نسبت به پوششی هستند [۱۳].

تاکنون تعداد بسیار زیادی پژوهش در زمینه‌ی انتخاب ویژگی انجام شده است. اما هنوز به‌صورت مطلق، بهترین روش برای انجام انتخاب ویژگی یافت نشده است. به عبارتی می‌توان گفت که هر یک از این روش‌ها باتوجه‌به نوع مسئله، می‌تواند عملکرد به‌خصوصی داشته باشد و از این‌رو تلاش‌های پژوهشگران بر انتخاب روشی متمرکز شده است که برای یک مسئله‌ی مشخص بهترین عملکرد را دارد.

۲-۴-۲- روش‌های انتخاب ویژگی نظارت شده^۲ و نظارت نشده^۳

برای انتخاب ویژگی برای یک مدل یادگیری نظارت شده، "روش‌های انتخاب ویژگی نظارت شده" ارائه شده است. هدف این الگوریتم‌ها، انتخاب بهترین زیرمجموعه از ویژگی‌ها برای تضمین عملکرد بهینه‌ی یک مدل نظارت شده است. این الگوریتم‌ها از داده‌های برچسب خورده شده استفاده می‌کنند. دسته‌ای دیگر از روش‌های انتخاب ویژگی نیز هستند که "روش‌های انتخاب ویژگی نظارت نشده" نام دارند. در این روش‌ها، داده‌های در دسترس برچسب ندارند؛ بنابراین پیاده‌سازی این روش‌ها بر اساس معیارهای مختلفی نظیر واریانس است. ویژگی‌های انتخاب شده توسط روش‌های انتخاب ویژگی نظارت نشده، می‌توانند در مدل‌های یادگیری نظارت شده نیز مورد استفاده قرار گیرند.

در ادامه به معرفی تعدادی از روش‌های انتخاب ویژگی نظارت شده پرداخته می‌شود.

۲-۴-۲-۱- الگوریتم Chi2

الگوریتم‌های انتخاب ویژگی را می‌توان بر اساس انواع داده‌هایی که بر روی آن‌ها کار می‌کنند، تقسیم‌بندی کرد. داده‌ها به دو نوع کمی و کیفی تقسیم می‌شوند. داده‌های کیفی به دو دسته‌ی اسمی و ترتیبی و داده‌های کمی نیز به دو دسته‌ی گسسته و پیوسته تقسیم می‌شوند. بسیاری از الگوریتم‌های انتخاب ویژگی بر روی داده‌های گسسته کار می‌کنند، به همین دلیل یکی از اقداماتی که در پیش‌پردازش انجام

^۱ Embedded

^۲ Supervised Feature Selection

^۳ Unsupervised Learning

می‌شود، گسسته کردن داده‌ها است [۱۴].

داده‌های عددی در مسائل دنیای واقعی بسیار رایج هستند. باین حال بسیاری از الگوریتم‌های طبقه‌بندی مستلزم این هستند که داده‌های آموزشی فقط دارای ویژگی‌های گسسته باشند و حتی برخی از آن‌ها بر روی داده‌های گسسته یا دودویی^۱ بهتر کار می‌کنند. اگر داده‌های کمی به طور خودکار به داده‌های گسسته تبدیل شوند، این الگوریتم‌های طبقه‌بندی به راحتی روی آن‌ها اجرا می‌شود. استفاده از الگوریتم Chi2 این هدف را برای ما محقق می‌کند [۱۴].

یک نکته‌ی قابل توجه در الگوریتم Chi2، این است که این الگوریتم می‌تواند روی داده‌های دارای چندین کلاس به کار رود، درحالی که تعدادی از الگوریتم‌های انتخاب ویژگی هستند که تنها برای داده‌های دارای دو کلاس (دوطبقه) قابل اجرا هستند [۱۴].

بنابراین می‌توان گفت که Chi2 یک الگوریتم ساده و کلی است که می‌تواند به طور خودکار ویژگی‌های مناسب را انتخاب کند. در نتیجه Chi2 یک ابزار مفید و قابل اعتماد برای گسسته‌سازی و انتخاب ویژگی است [۱۴].

۲-۴-۲- الگوریتم LASSO

تحلیل رگرسیون^۲ نقش مهمی در مدل‌سازی آماری و انجام الگوریتم‌های یادگیری ماشین دارد. روش‌های سنتی مانند رگرسیون حداقل مربعات معمولی^۳، رگرسیون گام‌به‌گام^۴ و رگرسیون حداقل مربعات جزئی^۵ به خطاهای تصادفی بسیار حساس‌اند. در دهه‌های گذشته جایگزین‌هایی برای این الگوریتم‌ها آورده شده است که عبارت‌اند از، رگرسیون خط‌الرأس^۶ و LASSO [۱۵].

روش OSL، مجموع باقی‌مانده‌ی مربعات مدل را به حداقل می‌رساند و علاوه‌بر آن پیچیدگی مدل را نیز کاهش می‌دهد؛ بنابراین، این روش که به عنوان منظم‌سازی نیز شناخته می‌شود، می‌تواند واریانس را به میزان قابل توجهی کاهش دهد و ویژگی‌های مناسب را نیز انتخاب کند. دو روش مهمی که ضرایب رگرسیون تخمین زده شده را به صفر نزدیک می‌کنند، خط‌الرأس و LASSO نام دارند. رگرسیون خط‌الرأس، مجموع مربعات ضرایب و LASSO مجموع مطلق ضرایب را به حداقل می‌رساند. در این پروژه تمرکز بر روی الگوریتم LASSO می‌باشد [۱۵].

LASSO یک روش انتخاب ویژگی آینده‌نگر برای رگرسیون است. مجموع باقی‌مانده‌ی مربعات را به

¹ Binary

² Regression

³ Ordinary Least Squares (OLS)

⁴ Stepwise

⁵ Partial Least Squares (PLS)

⁶ Ridge

شرطی کاهش می‌دهد که مجموع مقادیر مطلق ضرایب کمتر از یک عدد ثابت باشد. LASSO با ترکیب ویژگی‌های خوب رگرسیون خط‌الرأس و انتخاب زیرمجموعه، هم دقت پیش‌بینی و هم تفسیرپذیری مدل را بهبود می‌بخشد. هنگامی که همبستگی بالایی میان چند ویژگی وجود داشته باشد، LASSO فقط یکی از آن‌ها را انتخاب می‌کند و بقیه را به صفر می‌رساند. از این‌رو، روش LASSO می‌تواند به‌عنوان جایگزینی برای روش‌های انتخاب ویژگی متداول استفاده شود [۱۵].

۲-۴-۲-۳- الگوریتم درخت تصمیم^۱

یکی دیگر از الگوریتم‌های انتخاب ویژگی، الگوریتم درخت تصمیم است. یکی از مؤثرترین روش‌های اکتشافی که برای ساخت درخت تصمیم مورد استفاده قرار می‌گیرد، معیار تفکیک‌پذیری ارزش تقسیم‌شده^۲ است. مزیت اصلی SSV آن است که می‌توان آن را در هر دو ویژگی پیوسته و گسسته به‌گونه‌ای اعمال کرد که برآورد تفکیک‌پذیری بدون توجه به تفاوت قابل‌توجه آن‌ها مقایسه شود. مقدار تقسیم برای ویژگی‌های پیوسته و گسسته، هرکدام به‌صورت متفاوتی تعریف می‌شود. برای مقادیر گسسته، یک عدد واقعی و برای مقادیر پیوسته، یک زیرمجموعه از مجموعه مقادیر، جایگزین ویژگی می‌شود [۱۶].

معیار SSV نه‌تنها برای ساخت درختان و انتخاب ویژگی با موفقیت مورد استفاده قرار گرفته، بلکه در تبدیل نوع داده‌ی پیوسته به گسسته و همچنین گسسته به پیوسته نیز کارآمد بوده است. الگوریتم‌های درخت تصمیم به دلیل همین توانایی تشخیص ویژگی‌های مهم جهت انتخاب ویژگی، الگوریتم‌های شناخته شده‌ای برای انجام طبقه‌بندی هستند [۱۶].

انتخاب ویژگی بر اساس معیار SSV می‌تواند به روش‌های مختلفی طراحی شود. کارآمدترین روش، ایجاد رتبه‌بندی بر اساس حداکثر مقادیر SSV محاسبه شده برای هر یک از ویژگی‌ها است. گاهی اوقات دو یا سه تقسیم‌بندی ممکن است نیاز باشد تا ثابت شود که این ویژگی می‌تواند کلاس‌های مختلف را به طور دقیق جدا کند. داده‌های آموزشی به‌عنوان ورودی به الگوریتم درخت تصمیم وارد می‌شوند و ویژگی‌ها به ترتیب اهمیت آن‌ها در خروجی نمایش داده می‌شود [۱۶].

برای امکان تجزیه و تحلیل کامل، روش‌های انتخاب ویژگی باید بر روی مجموعه داده‌هایی که در فضایی با ابعاد نه‌چندان زیاد تعریف شده‌اند، استفاده شوند. البته باید توجه داشت که داده‌های با ابعاد بسیار کوچک نیز آزمون انتخاب ویژگی را بی‌اساس می‌کنند [۱۶].

۲-۴-۲-۴- الگوریتم ElasticNet

الگوریتم ElasticNet، ترکیبی از دو الگوریتم LASSO و خط‌الرأس است. یک تفاوت مفید این الگوریتم

^۱ Decision Tree

^۲ Separability of Split Value (SSV)

این است که هنگامی که تعداد پیش‌بینی‌کننده‌ها از تعداد مشاهدات بیشتر باشد، این الگوریتم عملکرد بسیار بهتری نسبت به الگوریتم LASSO دارد. ElasticNet، همانند LASSO به طور هم‌زمان انتخاب ویژگی خودکار را انجام می‌دهد. مطالعات انجام شده در زمینه‌ی داده‌های شبیه‌سازی شده و همچنین نمونه‌های واقعی، این واقعیت را نشان می‌دهد که ElasticNet در اغلب موارد، از نظر دقت پیش‌بینی، بهتر از LASSO عمل می‌کند [۱۷].

۲-۴-۵- الگوریتم XGBoost

یکی دیگر از الگوریتم‌های انتخاب ویژگی XGBoost است. این الگوریتم زیرمجموعه‌ای از الگوریتم‌های یادگیری مجموعه‌ای مبتنی بر درخت است که از آن برای تقویت درختان استفاده می‌شود. طبق مطالعات انجام شده، این الگوریتم موفقیت بسیاری در چالش‌های یادگیری ماشین و داده‌کاوی^۱ کسب کرده است [۱۸]. در الگوریتم XGBoost، اهمیت نسبی ویژگی را می‌توان با معیارهای وزن تقسیم‌شده و میانگین افزایش اندازه‌گیری کرد. وزن، تعداد دفعاتی است که از یک ویژگی برای تقسیم داده‌ها در تمام درختان استفاده می‌شود. از ویژگی‌های مهم‌تر نیز بیشتر استفاده می‌شود. زیرمجموعه‌ی ویژگی‌های بهینه را می‌توان بر مبنای تطابق بین عملکرد یادگیری و سادگی مدل (یعنی ویژگی‌های کمتر) انتخاب کرد [۱۸]. این الگوریتم می‌تواند زیرمجموعه‌ای از ویژگی‌های مفید را با نرخ مشارکت زیاد تعیین کند که این کار باعث حذف افزونگی و سیگنال‌های ناخواسته می‌شود. همچنین، پیچیدگی محاسبات را کاهش داده و کارایی عملکرد را بهبود می‌بخشد [۱۸].

۲-۵ جمع‌بندی

در این فصل به معرفی ریه، انواع عفونت‌های آن، رادیومیکس و ویژگی‌های آن و همچنین روش‌های انتخاب ویژگی پرداخته شد. همان‌طور که در قبل نیز توضیح داده شد، تمامی مطالعات پیشین درمورد استفاده از رادیومیکس و ویژگی‌هایش در ریه، پیرامون تشخیص تومور سرطانی و نوع عفونت ریه بوده است. حتی در مطالعاتی که جهت تشخیص ناحیه‌ی عفونی و اطراف آن صورت گرفته، از تکنولوژی‌های دیگر استفاده شده است و تنها در این پروژه از رادیومیکس که یکی از تکنولوژی‌های نوظهور است، استفاده می‌شود.

^۱ Data Mining

فصل سوم

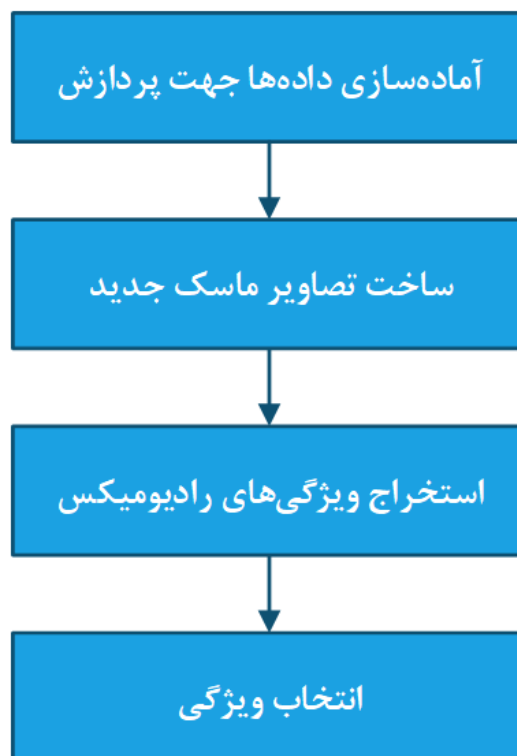
شرح پروژه

۱-۳ مقدمه

در این فصل به بررسی چهار گام روش پیشنهادی که عبارت‌اند از، آماده‌سازی داده‌ها جهت پردازش، ساخت تصاویر ماسک جدید، استخراج ویژگی‌های رادیومیکس و انتخاب ویژگی، پرداخته می‌شود.

۲-۳ روش پیشنهادی

روش پیشنهادی در این پروژه از چهار گام اصلی تشکیل شده است. این مراحل در شکل ۱-۳ نشان داده شده‌اند. در ادامه به بررسی هر یک از این گام‌ها پرداخته می‌شود.



شکل ۱-۳: چارچوب روش پیشنهادی

۳-۲-۱- آماده‌سازی داده‌ها جهت پردازش

اولین گام برای پیاده‌سازی این پروژه، آماده‌سازی داده‌ها جهت انجام پردازش بر روی آن‌ها است. داده‌های اولیه‌ی این پروژه به‌صورت تصاویر سه‌بعدی و در فرمت نیفتی^۱ است. در فرمت NIFTI، داده‌ها به دو صورت می‌توانند ذخیره شوند. در حالت اول، سرآیند^۲ در یک فایل با پسوند hdr و تصویر در یک فایل دیگر با پسوند img ذخیره می‌شود. در حالت دوم، داده‌ها به‌صورت یک فایل با پسوند nii که شامل سرآیند و تصویر است، ذخیره می‌شوند. داده‌های این پروژه نیز مانند حالت دوم و به‌صورت یک فایل با پسوند nii ذخیره شده‌اند.

۳-۲-۱-۱- بارگذاری^۳ تصاویر

داده‌های این پروژه در دو فایل تصاویر اصلی و تصاویر ماسک ذخیره شده‌اند. به‌ازای هر بیمار، یک تصویر اصلی و یک تصویر ماسک وجود دارد. تصاویر ماسک توسط رادیولوژیست برچسب‌گذاری شده‌اند. هر تصویر ماسک، تصویری به اندازه‌ی تصویر اصلی است. مقدار هر پیکسل از تصویر ماسک، صفر و یا عددی مخالف صفر است. اگر مقدار پیکسل عددی مخالف صفر باشد، نشان‌دهنده‌ی وجود عفونت در آن ناحیه و اگر برابر صفر باشد، نشان‌دهنده‌ی عدم وجود عفونت است. در این پروژه، مقادیر ذخیره شده در هر یک از پیکسل‌های تصویر ماسک صفر یا یک است که یک نشان‌دهنده‌ی وجود عفونت و صفر نشان‌دهنده‌ی عدم وجود عفونت در آن ناحیه از ریه است.

همان‌طور که بیان شد، تصاویر با فرمت NIFTI و به‌صورت فایل‌هایی با پسوند nii ذخیره شده‌اند؛ بنابراین برای بارگذاری این فایل‌ها، از کتابخانه‌ی nibabel که مخصوص تصاویر NIFTI است، استفاده می‌شود.

۳-۲-۱-۲- استخراج پیکسل‌ها

پس از بارگذاری تصاویر، نوبت به استخراج پیکسل‌ها می‌رسد. برای هر تصویر یک ماتریس در نظر گرفته و تمامی پیکسل‌ها در آن ذخیره می‌شوند. از این‌پس، برای کار با تصاویر و پردازش آن‌ها، از ماتریس حاوی پیکسل‌ها استفاده می‌شود.

۳-۲-۱-۳- جداسازی پیکسل‌های عفونی و غیرعفونی از تصاویر ماسک

طبق آنچه بیان شد، خانه‌هایی با مقدار یک در هر یک از تصاویر ماسک نشان‌دهنده‌ی نواحی عفونی و خانه‌هایی با مقدار صفر نشان‌دهنده‌ی نواحی غیرعفونی است. حال باید این دو دسته از یکدیگر جدا شده و

^۱ NIFTI

^۲ Header

^۳ Load

مختصات پیکسل‌هایشان نیز به صورت جداگانه ذخیره شوند؛ بنابراین در ماتریس حاوی پیکسل‌های هر تصویر ماسک، مختصات خانه‌های برابر با یک، در یک فایل با عنوان "عفونت" و مختصات خانه‌های برابر با صفر در فایل دیگری با عنوان "غیر عفونت" ذخیره می‌شوند.

حال باتوجه به اینکه تعداد نقاط به دست آمده برای فایل‌های عفونت و غیر عفونت زیاد است، با استفاده از تابع تصادفی^۱ ۱۰۰۰ نقطه از میان نقاط عفونت انتخاب می‌شوند. در ادامه، برای انتخاب تعدادی از نقاط غیر عفونت ابتدا باید محدوده‌ی وجود عفونت در ریه مشخص شده تا تنها نقاط غیر عفونی اطراف آن‌ها به عنوان غیر عفونت در نظر گرفته شوند. برای انجام این کار، مختصات اولین و آخرین نقطه‌ی حاوی عفونت در هر سه بُعد تصویر را انتخاب کرده و پس از تشکیل یک مکعب با این نقاط، تمامی نقاط خارج از این مکعب را از فایل غیر عفونت حذف کرده و در میان نقاط باقی مانده، با استفاده از تابع تصادفی، ۱۰۰۰ نقطه انتخاب می‌شود. این کار باعث می‌شود تا از انتخاب نقاط خارج از محدوده‌ی ریه به عنوان نقاط غیر عفونی جلوگیری شده و در نتیجه، دقت انجام محاسبات افزایش یابد.

در آخر می‌توان بیان کرد که خروجی این گام برای هر تصویر، دو فایل عفونت و غیر عفونت که هر کدام حاوی ۱۰۰۰ سطر از نقاط عفونی و غیر عفونی هستند می‌باشد. لازم به ذکر است که هر سطر از این فایل بیان کننده‌ی مختصات هر پیکسل از نقاط عفونی یا غیر عفونی است.

۳-۲-۲- ساخت تصاویر ماسک جدید

هدف از انجام این پروژه، افتراق ناحیه‌ی عفونی و اطراف آن در ریه، با استفاده از ویژگی‌های رادیومیکس است. حال برای به دست آوردن ویژگی‌های رادیومیکس، ابتدا باید پیکسل‌های موجود در همسایگی هر پیکسل در نظر گرفته شود تا با بررسی تغییرات آن‌ها، عفونی بودن یا نبودن آن پیکسل مشخص شود.

۳-۲-۲-۱- ساخت مکعب اطراف هر پیکسل

در این مرحله، ابتدا باید فایل نقاط عفونت و غیر عفونت به دست آمده از گام قبل بارگذاری^۲ شود. گفته شد که هر سطر از این فایل، نشان دهنده‌ی مختصات هر پیکسل عفونی یا غیر عفونی از تصاویر ماسک است؛ بنابراین برای بررسی همسایگی هر پیکسل، ابتدا باید هر سطر از این فایل را خوانده و سپس یک مکعب با ابعاد $5 \times 5 \times 5$ در اطراف آن تشکیل داده و مقدار هر پیکسل از این مکعب برابر با عدد یک در نظر گرفته شود. دلیل تشکیل همسایگی در اطراف هر پیکسل این است که بتواند مرز میان ناحیه‌ی عفونی و غیر عفونی را تشخیص دهد؛ زیرا همان طور که بیان شد، هدف در این پروژه افتراق ناحیه‌ی عفونی و غیر عفونی است و باید بررسی شود که این کار را تا چه میزان می‌تواند انجام دهد. از این رو در همسایگی هر پیکسل، ویژگی‌های

^۱ Random

^۲ Upload

رادیومیکس محاسبه می‌شوند.

۳-۲-۲- تبدیل ماتریس به تصویر

حال که در اطراف هر یک از نقاط عفونت و غیر عفونت یک مکعب تشکیل شد، نوبت به آن رسیده است تا تصاویر ماسک جدید تولید شوند. همان‌طور که در بالا اشاره شده است، تصاویر ماسک باید هم‌اندازه با تصاویر اصلی باشند. به همین دلیل، ابتدا باید به‌ازای هر تصویر یک ماتریس با ابعاد ماتریس تصویر اصلی تشکیل شده و مقدار تمامی عناصر آن برابر با صفر در نظر گرفته شود.

هنگامی که یک تصویر ماسک به‌عنوان ورودی به یک تابع داده می‌شود، به معنای آن است که تنها نواحی مخالف صفر از این تصویر باید مورد پردازش قرار گیرند. همان‌طور که گفته شد، هدف از انجام این پروژه افتراق ناحیه‌ی عفونی و اطراف آن است؛ بنابراین تنها این نواحی باید مورد بررسی قرار گیرند و دلیل اصلی در نظر گرفتن مقدار یک برای هر پیکسل از مکعب‌های ساخته شده در اطراف هر پیکسل عفونی یا غیر عفونی، محقق کردن همین امر بوده است.

در ادامه باید هر یک از این مکعب‌ها بسط داده و هم‌اندازه‌ی تصویر اصلی شوند. برای انجام این کار، از ماتریس ساخته شده در ابتدای این مرحله استفاده می‌شود. در نتیجه، یک ماتریس با ابعادی برابر با ابعاد تصویر اصلی وجود خواهد داشت که عناصر آن یا برابر صفر و یا برابر یک خواهند بود. توجه به این نکته نیز فراموش نشود که تنها عناصر موجود در خانه‌های مکعب گذاشته شده در ماتریس برابر یک خواهند بود و بقیه‌ی عناصر حتماً باید برابر با صفر باشند.

در انتها، هر یک از ماتریس‌های ساخته شده به‌عنوان ماسک جدید، باید به تصاویری با فرمت NIFTI تبدیل شوند. به همین جهت از کتابخانه‌ی nibabel استفاده شده و برای هر تصویر اصلی، به تعداد ۱۰۰۰ تصویر ماسک عفونت و ۱۰۰۰ تصویر ماسک غیر عفونت تشکیل و ذخیره می‌شوند.

۳-۲-۳- استخراج ویژگی‌های رادیومیکس

استخراج ویژگی‌های رادیومیکس، اصلی‌ترین گام در پیاده‌سازی این پروژه است. برای کار با رادیومیکس، پیش از هر چیز باید اقدام به نصب بسته‌ی^۱ Pyradiomics کرد [۱۲].

Pyradiomics یک بسته‌ی پایتون^۲ منبع‌باز^۳ برای استخراج داده‌های رادیومیکس از تصاویر پزشکی است. بارگذاری و پیش‌پردازش تصاویر در این بسته، با استفاده از کتابخانه‌ی SimpleITK انجام می‌شود. کتابخانه‌ی SimpleITK، منبع‌باز و شیء‌گرا^۴ است. علاوه بر این، یک رابطه‌ی رویه‌ای^۱ را نیز برای اکثر

^۱ Package

^۲ Python

^۳ Open Source

^۴ Object Oriented

فیلترهای پردازش تصویر ارائه می‌دهد. Pyradiomics از تصاویر دوبعدی و سه‌بعدی پشتیبانی می‌کند؛ اما در حال حاضر استخراج از تصاویر رنگی و تصاویر با مقادیر پیچیده را ارائه نمی‌دهد [۱۲].

۳-۲-۳-۱- سفارشی‌سازی در رادیومیکس

چهار روش برای سفارشی‌سازی استخراج ویژگی‌ها در Pyradiomics وجود دارد:

- (۱) تعیین نوع تصویر برای استخراج ویژگی‌ها.
- (۲) تعیین اینکه کدام ویژگی‌ها استخراج شوند.
- (۳) تعیین تنظیماتی که پیش‌پردازش را کنترل کرده و رفتار فیلترهای فعال و کلاس‌های ویژگی را سفارشی می‌کند.
- (۴) تعیین تنظیمات خاص مبتنی بر وکسل

هر چهار دسته سفارشی‌سازی بیان شده را می‌توان در یک فایل متنی ساختاریافته‌ی YAML یا JSON ارائه کرد. YAML و JSON زبان‌های نشانه‌گذاری برای ذخیره‌ی اطلاعات هستند. در این فایل متنی ساختاریافته می‌توان به راحتی این موضوع را که کدام یک از ویژگی‌ها و انواع تصویر ورودی را می‌توان فعال کرد را نیز مشخص کرد. در مرحله‌ی بعدی به صورت کامل این پارامترها^۲ بررسی شده‌اند [۱۲].

۳-۲-۳-۲- تنظیم پارامترهای اولیه‌ی رادیومیکس

فایل متنی ساختاریافته را می‌توان با استفاده از تابع LoadParams() در رادیومیکس فراخوانی کرد. این کار انجام فرایند سفارشی‌سازی و تنظیم پارامترهای اولیه را آسان می‌کند [۱۲].
تنظیمات در این مرحله به شرح زیر است:

- ImageType: نوع تصویر برای محاسبه‌ی ویژگی‌ها مشخص می‌شود. اگر ورودی به این تابع داده نشود، هیچ تنظیمات سفارشی برای آن تصویر ورودی اضافه نمی‌شود [۱۲].
- FeatureClass: این تابع برای فعال کردن ویژگی‌ها استفاده می‌شود. اگر هیچ ورودی به این تابع داده نشود، تمام ویژگی‌های تصاویر ورودی را محاسبه می‌کند؛ اما در صورت قراردادن مقداری در ورودی این تابع، تنها همان ویژگی ذکر شده محاسبه می‌شود [۱۲].
- Setting: تنظیمات لازم جهت پیش‌پردازش و تنظیمات خاص کلاس‌ها در این تابع انجام می‌شود. یکی از موارد مهمی که باید در مرحله‌ی پیش‌پردازش در نظر گرفته شود، نرمال‌سازی^۳ است. اگر در مرحله‌ی پیش‌پردازش نرمال‌سازی انجام نشود، نتایج به دست آمده صحیح نخواهند

^۱ Procedural

^۲ Parameters

^۳ Normalization

بود. به همین دلیل در هنگام تنظیم پارامترهای اولیه باید انجام نرمال سازی را نیز فعال کرد؛

زیرا در رادیومیکس این پارامتر به صورت پیش فرض غیر فعال است [۱۲].

• VoxelSetting: تنظیمات مربوط به وکسل ها در این تابع انجام می شود [۱۲].

۳-۲-۳- به دست آوردن مقادیر ویژگی های رادیومیکس

حال در این مرحله نوبت به دست آوردن مقادیر ویژگی های رادیومیکس برای هر تصویر است. رادیومیکس، دو تصویر اصلی و ماسک را به عنوان ورودی می پذیرد و با در نظر گرفتن پارامترهای تنظیم شده در مرحله قبل، مقادیر ویژگی های رادیومیکس را برای آن ها محاسبه کرده و در یک فایل با پسوند csv ذخیره می کند.

در گام قبل برای هر تصویر اصلی ۱۰۰۰ تصویر ماسک عفونت و ۱۰۰۰ تصویر ماسک غیر عفونت به دست آمد. حال در این مرحله باید هر یک از تصاویر ماسک عفونت و هر یک از تصاویر ماسک غیر عفونت به صورت جداگانه و همراه با تصویر اصلی به عنوان ورودی رادیومیکس در نظر گرفته شود. یعنی به ازای هر تصویر اصلی ۲۰۰۰ تصویر ماسک وجود دارد که در هر بار، یکی از آن ها به همراه تصویر اصلی به عنوان ورودی به توابع رادیومیکس داده شده و ویژگی های رادیومیکس برای آن ها محاسبه می شود. در نهایت برای هر تصویر یک فایل با پسوند csv شامل ۲۰۰۰ سطر که هر سطر آن حاوی ۱۲۸ ستون (هر ستون بیانگر یک ویژگی) می باشد، به عنوان خروجی دریافت می شود.

به این نکته نیز توجه شود که در انجام این مرحله، به دلیل حجم بالای پردازش ها نیاز به استفاده از پردازش های موازی و یا استفاده از GPU است. با بهره گیری از این دو مورد می توان زمان انجام پردازش ها را کاهش داد.

بنابراین، تا این مرحله از کار، برای هر تصویر یک فایل با پسوند csv ذخیره شده است که شامل مقادیر به دست آمده برای هر ویژگی رادیومیکس است.

۳-۲-۴- انتخاب ویژگی

انتخاب ویژگی، آخرین گام در راستای پیاده سازی این پروژه است. برای اجرای الگوریتم های انتخاب ویژگی، ابتدا باید داده های ورودی برچسب گذاری شوند. در این جا دو دسته داده ی عفونت و غیر عفونت وجود دارد؛ بنابراین برای برچسب گذاری داده ها، می توان آن ها را به دو کلاس صفر و یک تقسیم کرد. به همین منظور، ابتدا باید یک ستون به عنوان "Infection" را به داده ها اضافه کرده و با توجه به اینکه سطر مورد نظر مربوط به داده های عفونی یا غیر عفونی است، مقدار آن "۱" یا "صفر" می شود. سپس تمامی فایل های به دست آمده در مرحله ی قبل را به هم چسبانده و تبدیل به یک فایل نهایی با پسوند csv کرده تا بتوان آن را به عنوان ورودی برای الگوریتم های انتخاب ویژگی در نظر گرفت. در ادامه نیز برای ترکیب شدن داده ها از تابع

بر زدن^۱ استفاده کرده تا نتایج به دست آمده در انتهای کار صحیح باشند.

در آخرین مرحله از پیاده سازی این گام، باید فایل نهایی به دست آمده را به عنوان ورودی به الگوریتم های انتخاب ویژگی داده و این الگوریتم ها را بر روی آن ها اجرا کرده تا خروجی های مورد انتظار به دست آیند.

الگوریتم های انتخاب ویژگی مورد استفاده در این پروژه، الگوریتم های LASSO، TreeClassifier، ElasticNet و XGBoost هستند. در هر مرحله یکی از این الگوریتم ها مورد استفاده قرار می گیرد و نتایج حاصل از اجرای آن ذخیره و در ادامه مورد ارزیابی قرار می گیرد.

۳-۳ جمع بندی

در این فصل به بررسی چهار گام اصلی روش پیشنهادی که عبارت اند از آماده سازی داده ها جهت پردازش، ساخت تصاویر ماسک جدید، استخراج ویژگی های رادیومیکس و انتخاب ویژگی پرداخته شد. در مسیر بررسی هر گام، مراحل پیاده سازی آن ها به طور کامل شرح داده شد. در فصل بعد نتایج به دست آمده از پیاده سازی های انجام شده، توضیح داده شده است.

^۱ Shuffle

فصل چهارم

نتایج

۴-۱ مقدمه

در این فصل، به بررسی مجموعه داده‌ها، معیارهای ارزیابی، نتایج و ارزیابی نتایج به دست آمده از اعمال روش پیشنهادی بر روی مجموعه داده‌ها پرداخته می‌شود.

۴-۲ مجموعه داده‌ها

برای ارزیابی روش پیشنهادی در این پروژه، از مجموعه داده‌های دریافتی از بیمارستان غیائی تهران استفاده شده است. این مجموعه داده شامل تصاویر CT بیماران ایرانی که ریه‌ی آن‌ها درگیر ویروس کرونا بوده است، می‌باشد.

این مجموعه داده شامل دو فایل تصاویر اصلی و تصاویر ماسک است. تصاویر ماسک توسط یک رادیولوژیست برچسب‌گذاری شده‌اند. تعداد تصاویر در اختیار این پروژه، ۶۰ تصویر CT ریه و ۶۰ تصویر برچسب‌گذاری شده‌ی ماسک است. به عبارتی، به ازای هر تصویر اصلی، یک تصویر برچسب‌گذاری شده‌ی ماسک وجود دارد.

می‌توان گفت که از تصاویر ماسک به عنوان داده‌های آموزشی استفاده می‌شود. در تصاویر ماسک، دو کلاس صفر و یک وجود دارد. کلاس "۱" نشان‌دهنده‌ی وجود عفونت و کلاس "صفر" نشان‌دهنده‌ی عدم وجود عفونت است. این تصاویر سه‌بعدی هستند و به کمک نرم‌افزار 3D Slicer برچسب‌گذاری شده‌اند که در ادامه توضیحاتی پیرامون این نرم‌افزار ارائه می‌شود.

۴-۳ معیارهای ارزیابی

برای ارزیابی روش پیشنهادی، باید معیارهایی برای سنجش دقت و صحت در نظر گرفته شود. در این بخش به معرفی و بررسی هر یک از این معیارها پرداخته می‌شود.

۴-۳-۱- ماتریس درهم‌ریختگی^۱

فرض بر این است که ۱۰۰ نمونه داده در دسترس باشد. حال این داده‌ها تک‌تک به مدل داده شده و به‌ازای هرکدام، یک کلاس به‌عنوان خروجی دریافت می‌شود. کلاس پیش‌بینی شده توسط مدل و کلاس واقعی داده‌ها را می‌توان در یک جدول نمایش داد. به این جدول Confusion Matrix گفته می‌شود.

به‌عنوان مثال در Confusion Matrix نشان داده شده در جدول ۴-۱، هدف شناسایی افراد مبتلا به یک نوع بیماری است. تعلق به دسته افراد بیمار، مثبت^۲ و عدم تعلق به این دسته منفی^۳ در نظر گرفته شده است. هر نمونه به یکی از کلاس‌های مثبت یا منفی تعلق دارد. از سوی دیگر، هر الگوریتمی که برای طبقه‌بندی استفاده می‌شود نیز هر نمونه را در یکی از دو کلاس مثبت یا منفی قرار می‌دهد؛ بنابراین برای هر نمونه داده، یکی از چهار حالتی که در ادامه بیان شده است ممکن است رخ دهد.

- منفی صحیح^۴: نمونه عضو کلاس منفی باشد و عضو همین کلاس نیز تشخیص داده شود.
- منفی کاذب^۵: نمونه عضو کلاس مثبت باشد ولی عضو کلاس منفی تشخیص داده شود.
- مثبت صحیح^۶: نمونه عضو کلاس مثبت باشد و عضو همین کلاس نیز تشخیص داده شود.
- مثبت کاذب^۷: نمونه عضو کلاس مثبت باشد ولی عضو کلاس منفی تشخیص داده شود.

جدول ۴-۱: ماتریس درهم‌ریختگی

	منفی (پیش‌بینی شده)	مثبت (پیش‌بینی شده)
منفی (واقعی)	True Positive	False Positive
مثبت (واقعی)	False Negative	True Positive

۴-۳-۲- دقت^۸

Accuracy به این معنا است که مدل تا چه اندازه خروجی را درست پیش‌بینی می‌کند. مقدار Accuracy با توجه به فرمول شماره (۱) به‌دست می‌آید [۱۹].

^۱ Confusion Matrix

^۲ Positive

^۳ Negative

^۴ True Negative

^۵ False Negative

^۶ True Positive

^۷ False Positive

^۸ Accuracy

$$Accuracy = \frac{true\ positives + true\ negatives}{total\ examples} \quad (۱)$$

۴-۳-۳- صحت^۱

Precision، صحت مدل را بیان می‌کند؛ یعنی هنگامی که مدل نتیجه‌ای را پیش‌بینی می‌کند، نشان می‌دهد که این نتیجه تا چه اندازه درست است. مقدار Precision باتوجه به فرمول شماره (۲) به‌دست می‌آید [۱۹].

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (۲)$$

زمانی که ارزش false positives بالا باشد، معیار Precision مناسب خواهد بود [۱۹].

۴-۳-۴- بازیابی^۲

مقدار Recall باتوجه به فرمول شماره (۳) به‌دست می‌آید [۱۹].

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (۳)$$

زمانی که ارزش false negatives بالا باشد، معیار Recall مناسب خواهد بود [۱۹].

۴-۳-۵- F1 Score

یک معیار مناسب برای ارزیابی دقت یک آزمایش است. این معیار Precision و Recall را باهم در نظر می‌گیرد و باتوجه به فرمول شماره (۴) محاسبه می‌شود [۱۹].

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (۴)$$

در این پروژه برای ارزیابی خروجی‌های روش پیشنهادی، از معیار F1 Score استفاده شده است. زیرا این معیار دو معیار ارزیابی مهم را در انجام محاسباتش در نظر گرفته و بهترین نتیجه را ارائه می‌دهد.

۴-۴ ابزارها و کتابخانه‌ها

برای انجام این پروژه از ابزارها و کتابخانه‌های مختلفی استفاده شده است که در ادامه به بررسی هر یک از آنها می‌پردازیم.

^۱ Precision

^۲ Recall

۴-۴-۱- ابزارها

ابزارهای مورد استفاده برای اجرای این پروژه عبارتند از:

- 3D Slicer: نرم‌افزاری برای پردازش، تجزیه و تحلیل و تبیین داده‌های مرتبط با تصاویر پزشکی است. این نرم‌افزار دارای ابزارهای قدرتمندی در زمینه‌ی تحقیقات بالینی و پردازش تصاویر پزشکی است و امکان استفاده در زمینه‌های آزمایشگاهی و پزشکی هسته‌ای را نیز دارد. مجموعه داده‌های این پروژه که شامل تصاویر CT هستند، با استفاده از این نرم‌افزار آماده و برچسب‌گذاری شده‌اند.
- ژوپیتِر نوت‌بوک^۱: یکی از نرم‌افزارهای بسیار کاربردی در حوزه‌ی برنامه‌نویسی پایتون در علوم داده^۲ و داده‌کاوی است. این ابزار می‌تواند به راحتی عملیات مختلفی را در یک محیط ساده همراه با خروجی‌های در لحظه انجام دهد. این نرم‌افزار جهت انجام کارهای تحقیق و توسعه در شرکت‌ها یا دانشگاه‌ها بسیار کاربرد دارد.
- گوگل کولب^۳: کوتاه شده‌ی واژه‌ی آزمایشگاه مشترک^۴ گوگل است که به اختصار به آن Colab گفته می‌شود. Colab یک میزبان^۵ برخط^۶ برای پروژه‌های Jupyter است که امکان پردازش ساده و یا پردازش با استفاده از GPU را برای آن‌ها فراهم می‌کند. Colab می‌تواند کدها را از گیت‌هاب^۷ یا گوگل درایو^۸ فراخوانی و یا به آن‌ها ارسال کند.

۴-۴-۲- کتابخانه‌ها

- کتابخانه‌ی numpy: یک کتابخانه برای زبان برنامه‌نویسی پایتون است. با استفاده از این کتابخانه امکان استفاده از آرایه‌ها و ماتریس‌های بزرگ چندبعدی فراهم می‌شود. همچنین می‌توان از توابع ریاضی سطح بالا بر روی این آرایه‌ها استفاده کرد.
- کتابخانه‌ی pandas: یک کتابخانه‌ی قدرتمند برای تحلیل، پیش‌پردازش و بصری‌سازی^۹ داده‌ها است.
- کتابخانه‌ی Matplotlib: یک کتابخانه‌ی ترسیم نمودار برای زبان برنامه‌نویسی پایتون است.
- کتابخانه‌ی Radiomics: کتابخانه‌ای برای استخراج داده‌های رادیومیکس از تصاویر پزشکی

^۱ Jupyter Notebook

^۲ Data Science

^۳ Google Colab

^۴ Colaboratory

^۵ Host

^۶ Online

^۷ GitHub

^۸ Google Drive

^۹ Visualization

است.

- کتابخانه‌ی nibabel: کتابخانه‌ای که برای کار با تصاویر NIFTI از آن استفاده می‌شود.

۴-۵ نتایج

در این بخش، نتایج حاصل از اجرای الگوریتم‌های انتخاب ویژگی بر روی ویژگی‌های رادیومیکس به‌دست‌آمده، نشان داده می‌شود. در این پروژه از تعدادی از الگوریتم‌های انتخاب ویژگی جهت مقایسه‌ی نتایج آن‌ها استفاده شده است.

در ادامه، جداول و نمودارهای نتایج به‌دست‌آمده از پیاده‌سازی هر یک از الگوریتم‌ها نشان داده شده است.

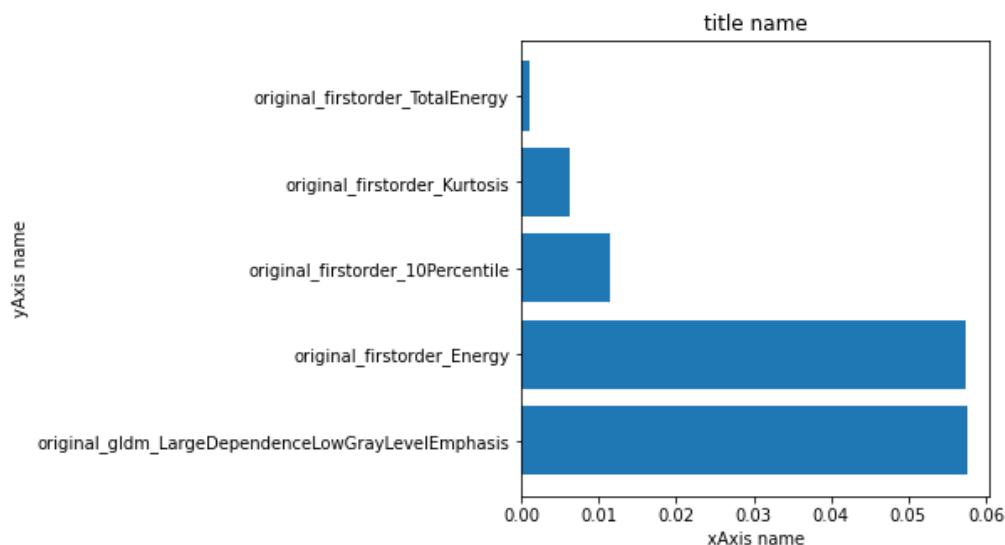
۴-۵-۱- الگوریتم LASSO

در جدول ۴-۲ نتایج حاصل از اجرای الگوریتم LASSO نشان داده شده است. این جدول ویژگی‌های انتخاب شده توسط الگوریتم LASSO را به همراه F1 Score به‌دست‌آمده برای هر یک از ویژگی‌ها را نشان می‌دهد. در ستون "Radiomics Features"، ویژگی‌های رادیومیکس انتخاب شده توسط این الگوریتم نشان داده شده است. در ستون "F1 Score" نیز، مقدار F1 Score محاسبه شده برای هر ویژگی و در ستون "F1 Score%" مقدار هر F1 Score محاسبه شده به‌صورت درصد نشان داده شده است.

جدول ۴-۲: نتایج حاصل از اجرای الگوریتم LASSO

Radiomics Features	F1 Score	F1 Score%
original_gldm_LargeDependenceLowGrayLevelEmphasis	0.05755284	57.55%
original_firstorder_Energy	0.0573402	57.34%
original_firstorder_10Percentile	0.01156015	11.56%
original_firstorder_Kurtosis	0.00635323	6.35%
original_firstorder_TotalEnergy	0.00107739	1.08%

در شکل ۴-۱ نمودار حاصل از اجرای الگوریتم LASSO نشان داده شده است. ویژگی‌های رادیومیکس انتخاب شده توسط این الگوریتم که در جدول ۴-۲ نیز مشاهده شد، در این نمودار ویژگی‌های رادیومیکس انتخاب شده به همراه F1 Score محاسبه شده برای هر یک از آن‌ها نشان داده شده است. این نمودار جهت درک بهتر میزان اهمیت ویژگی‌ها نسبت به هم رسم شده است.



شکل ۴-۱: نمودار حاصا از اجرای الگوریتم LASSO

۴-۵-۲- الگوریتم TreeClassifier

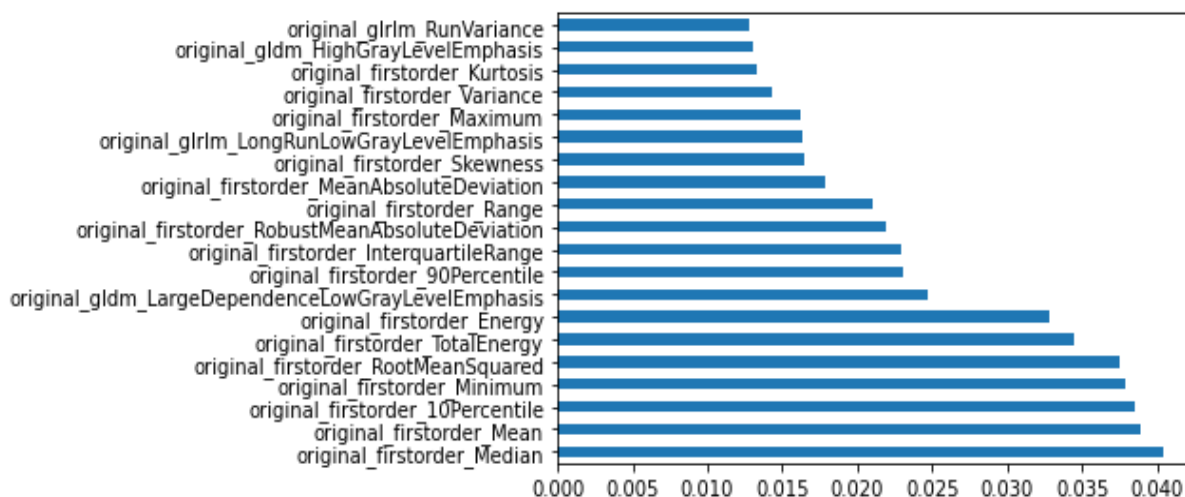
در جدول ۴-۳ نتایج حاصل از اجرای الگوریتم TreeClassifier نشان داده شده است. همان طور که بیان شد ویژگی‌های رادیومیکس انتخاب شده توسط این الگوریتم به همراه F1 Score محاسبه شده برای هر یک و درصد F1 Score آن‌ها در جدول نشان داده شده است.

در شکل ۴-۲ نیز نمودار حاصل از اجرای این الگوریتم، جهت مشاهده‌ی رابطه‌ی میان ویژگی‌های انتخاب شده و میزان اهمیت هر یک از آن‌ها نشان داده شده است.

جدول ۴-۳: نتایج حاصل از اجرای الگوریتم TreeClassifier

Radiomics Features	F1 Score	F1 Score%
original_firstorder_Median	0.040402	40.40%
original_firstorder_Mean	0.038917	38.91%
original_firstorder_10Percentile	0.038502	38.5%
original_firstorder_Minimum	0.037834	37.83%
original_firstorder_RootMeanSquared	0.037513	37.51%
original_firstorder_TotalEnergy	0.034469	34.46%
original_firstorder_Energy	0.032799	32.79%
original_gldm_LargeDependenceLowGrayLevelEmphasis	0.024682	24.68%
original_firstorder_90Percentile	0.022993	22.99%
original_firstorder_InterquartileRange	0.022918	22.91%
original_firstorder_RobustMeanAbsoluteDeviation	0.021873	21.87%
original_firstorder_Range	0.021008	21%
original_firstorder_MeanAbsoluteDeviation	0.017817	17.81%
original_firstorder_Skewness	0.016417	16.41%

original_glrml_LongRunLowGrayLevelEmphasis	0.016282	16.28%
original_firstorder_Maximum	0.016205	16.2%
original_firstorder_Variance	0.014258	14.25%
original_firstorder_Kurtosis	0.013261	13.26%
original_gldm_HighGrayLevelEmphasis	0.013069	13.06%
original_glrml_RunVariance	0.012746	12.74%



شکل ۴-۲: نمودار حاصل از اجرای الگوریتم TreeClassifier

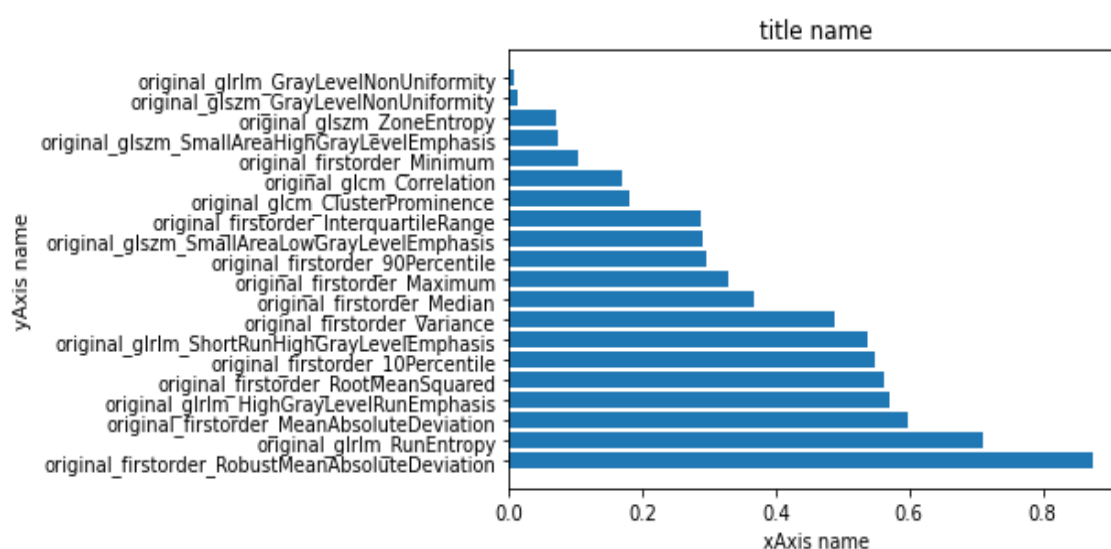
۴-۵-۳- الگوریتم ElasticNet

در جدول ۴-۴ و در شکل ۴-۳ نتایج حاصل از اجرای الگوریتم ElasticNet نشان داده شده است. نمودار شکل ۴-۳ به خوبی میزان اهمیت هر یک از ویژگی‌های به دست آمده و جدول ۴-۴ نام ویژگی‌ها و میزان F1 Score به دست آمده برای هر یک از آن‌ها را نشان می‌دهد.

جدول ۴-۴: نتایج حاصل از اجرای الگوریتم ElasticNet

Radiomics Features	F1 Score	F1 Score%
original_firstorder_RobustMeanAbsoluteDeviation	0.875197	87.52%
original_glrml_RunEntropy	0.709475	70.95%
original_firstorder_MeanAbsoluteDeviation	0.597902	59.79%
original_glrml_HighGrayLevelRunEmphasis	0.569827	56.98%
original_firstorder_RootMeanSquared	0.561298	56.13%
original_firstorder_10Percentile	0.548224	55%
original_glrml_ShortRunHighGrayLevelEmphasis	0.537584	53.76%
original_firstorder_Variance	0.488843	48.88%
original_firstorder_Median	0.367471	36.75%
original_firstorder_Maximum	0.330114	33.01%

original_firstorder_90Percentile	0.296733	29.67%
original_glszm_SmallAreaLowGrayLevelEmphasis	0.290829	29.08%
original_firstorder_InterquartileRange	0.287659	28.77%
original_glcm_ClusterProminence	0.18045	18.05%
original_glcm_Correlation	0.16997	17.00%
original_firstorder_Minimum	0.102832	10.28%
original_glszm_SmallAreaHighGrayLevelEmphasis	0.07268	7.26%
original_glszm_ZoneEntropy	0.069496	6.95%
original_glszm_GrayLevelNonUniformity	0.012872	1.29%
original_glrIm_GrayLevelNonUniformity	0.006869	0.69%



شکل ۴-۳: نمودار حاصل از اجرای الگوریتم ElasticNet

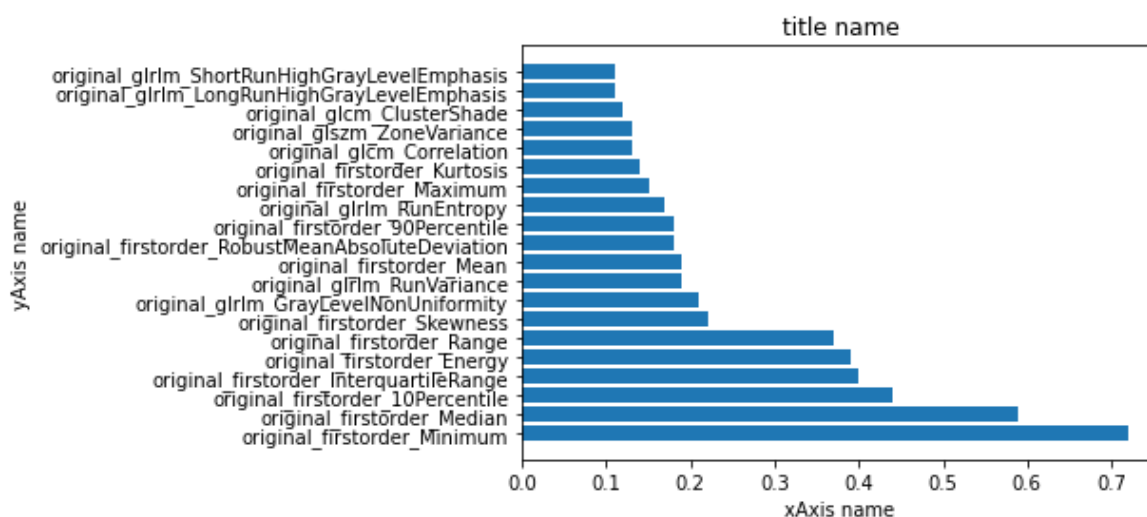
۴-۵-۴- الگوریتم XGBoost

نتایج به دست آمده از اجرای الگوریتم XGBoost، به عنوان آخرین الگوریتم انتخاب ویژگی بر روی ویژگی‌های رادیومیکس به دست آمده از مجموعه داده‌ها، در جدول ۴-۵ نشان داده شده است. در این جدول نیز مانند جداول قبلی، ویژگی‌های انتخاب شده به همراه F1 Score محاسبه شده برای هر یک از ویژگی‌ها و درصد F1 Score محاسبه شده، نشان داده می‌شود.

در شکل ۴-۴ نیز نمودار حاصل از اجرای این الگوریتم، جهت مشاهده‌ی آسان‌تر رابطه و میزان اهمیت ویژگی‌های به دست آمده، نشان داده شده است.

جدول ۴-۵: نتایج حاصل از اجرای الگوریتم XGBoost

Radiomics Features	F1 Score	F1 Score%
original_firstorder_Minimum	0.72	72%
original_firstorder_Median	0.59	59%
original_firstorder_10Percentile	0.44	44%
original_firstorder_InterquartileRange	0.4	40%
original_firstorder_Energy	0.39	39%
original_firstorder_Range	0.37	37%
original_firstorder_Skewness	0.22	22%
original_glrlm_GrayLevelNonUniformity	0.21	21%
original_glrlm_RunVariance	0.19	19%
original_firstorder_Mean	0.19	19%
original_firstorder_RobustMeanAbsoluteDeviation	0.18	18%
original_firstorder_90Percentile	0.18	18%
original_glrlm_RunEntropy	0.17	17%
original_firstorder_Maximum	0.15	15%
original_firstorder_Kurtosis	0.14	14%
original_glcmm_Correlation	0.13	13%
original_glszm_ZoneVariance	0.13	13%
original_glcmm_ClusterShade	0.12	12%
original_glrlm_LongRunHighGrayLevelEmphasis	0.11	11%
original_glrlm_ShortRunHighGrayLevelEmphasis	0.11	11%



شکل ۴-۴: نمودار حاصل از اجرای الگوریتم XGBoost

۴-۶ ارزیابی نتایج

جداول و نمودارهای نشان داده شده، نتایج حاصل از اجرای الگوریتم‌های انتخاب ویژگی بر روی ویژگی‌های رادیومیکس به دست آمده از تصاویر را نشان می‌دهند. همان‌طور که بیان شد، معیار ارزیابی در این پروژه F1 Score است. هر چه مقدار F1 Score به دست آمده برای هر ویژگی در میان ویژگی‌های انتخاب شده بیشتر باشد، به این معنا است که آن ویژگی تأثیر بیشتری در جدا کردن نواحی عفونی و غیر عفونی در ریه را دارد.

طبق نتایج به دست آمده از پژوهش‌های پیشین، الگوریتم‌های XGBoost و LASSO قدرتمندترین الگوریتم‌ها در انتخاب ویژگی‌های رادیومیکس در بحث تشخیص تومور ریه بودند [۲۰].

یک نقص اصلی در ویژگی‌های رادیومیکس، تکرارپذیری کم ویژگی‌ها است که مقایسه و تفسیر مطالعات رادیومیکس را دشوار می‌سازد. به طور معمول، ویژگی‌ها به صورت ریاضی و با استفاده از عوامل مؤثر ناشی از پارامترهای تصویرسازی مانند وضوح و کسل‌ها و روش‌های بازسازی تعریف شده‌اند. در یکی از پژوهش‌های انجام شده در این زمینه، نشان داده شده است که الگوریتم TreeClassifier عملکرد خوبی را در زمینه‌ی شناسایی ویژگی‌هایی با قابلیت تکرارپذیری بالا دارد [۲۱].

به عبارتی می‌توان گفت که با استفاده از ویژگی‌های به دست آمده از الگوریتم TreeClassifier، نقص اصلی موجود در ویژگی‌های رادیومیکس نیز برطرف می‌شود؛ زیرا تکرارپذیری ویژگی‌های رادیومیکس یکی از موضوعات تحقیقاتی مهم در آینده است.

میانگین F1 Score محاسبه شده برای هر یک از الگوریتم‌ها در جدول ۴-۶ نشان داده شده است. همان‌طور که مشاهده می‌شود، میانگین F1 Score محاسبه شده برای الگوریتم ElasticNet، LASSO و XGBoost بیشترین مقدار را دارد و گفته شد که هر چه مقدار F1 Score محاسبه شده بیشتر باشد بهتر است.

جدول ۴-۶: میانگین F1 Score محاسبه شده در الگوریتم‌های انتخاب ویژگی

نام الگوریتم	میانگین F1 Score
ElasticNet	0.38
XGBoost	0.26
LASSO	0.26
TreeClassifier	0.24

حال باید به این نکته توجه داشت که به دلیل تفاوت زیاد در ریه‌ی افراد مختلف، تمامی پژوهش‌های انجام شده در این زمینه، کارایی^۱ پایینی داشته است که دلیل این امر همین بحث تفاوت بسیار زیاد در ریه‌ی

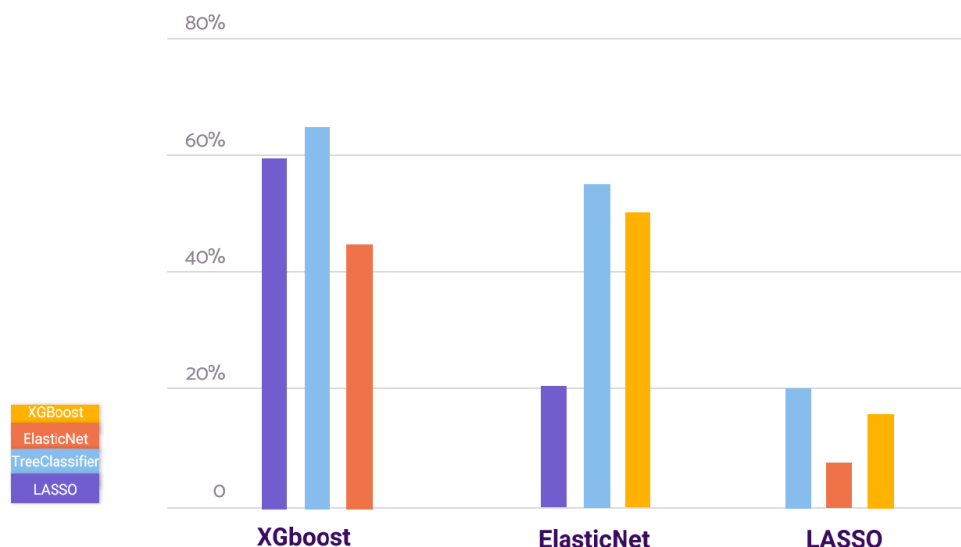
^۱ Performance

افراد است. به عنوان مثال در پژوهش‌های انجام شده در زمینه‌ی بیماری‌ها و تومورهای مغزی، کارایی‌های به‌دست‌آمده در این زمینه بسیار بالا بوده و در برخی از موارد به کارایی ۹۸٪ نیز رسیده است اما در ریه اینگونه نیست؛ به همین دلیل این میزان F1 Score به‌دست‌آمده در این پروژه قابل قبول و تفاوت پائین میان آن‌ها مورد توجه است.

باتوجه به اینکه الگوریتم TreeClassifier کمترین F1 Score را در میان الگوریتم‌های دیگر دارد، از کاندید شدن برای بهترین عملکرد در این زمینه حذف می‌شود و تنها میزان هم‌پوشانی میان سه الگوریتم ElasticNet، XGBoost و LASSO بررسی می‌شود.

با مقایسه و ارزیابی نتایج به‌دست‌آمده و همچنین در نظر گرفتن مطالعات انجام شده در زمینه‌های مشابه، همان‌طور که در شکل ۴-۵ نیز مشاهده می‌شود که الگوریتم XGBoost می‌تواند ۶۰٪ از ویژگی‌های به‌دست‌آمده از طریق اجرای الگوریتم LASSO، ۶۵٪ از ویژگی‌های الگوریتم TreeClassifier و ۴۵٪ از ویژگی‌های الگوریتم ElasticNet را پوشش دهد. الگوریتم ElasticNet نیز می‌تواند ۵۵٪ از ویژگی‌های به‌دست‌آمده از طریق اجرای الگوریتم TreeClassifier، ۴۹٪ از ویژگی‌های الگوریتم XGBoost و ۲۰٪ از ویژگی‌های الگوریتم LASSO را پوشش دهد. در این میان الگوریتم LASSO تنها ۱۵٪ از ویژگی‌های به‌دست‌آمده از طریق اجرای الگوریتم XGBoost، ۵٪ از ویژگی‌های الگوریتم ElasticNet و ۲۰٪ از ویژگی‌های الگوریتم TreeClassifier را پوشش می‌دهد؛ بنابراین میزان هم‌پوشانی الگوریتم LASSO از دو الگوریتم ElasticNet و XGBoost پایین‌تر است.

در نتیجه دو الگوریتم ElasticNet و XGBoost باتوجه به اینکه بیشترین میزان F1 Score را دارند و همچنین بیشترین هم‌پوشانی را با دیگر الگوریتم‌ها دارند، به عنوان قدرتمندترین الگوریتم‌ها در این زمینه انتخاب می‌شوند.



شکل ۴-۵: مقایسه‌ی میزان هم‌پوشانی الگوریتم‌های انتخاب ویژگی

ویژگی‌هایی که در سه الگوریتم XGboost, ElasticNet و TreeClassifier مشترک هستند، عبارت‌اند از:

- original_firstorder_Minimum
- original_firstorder_Median
- original_firstorder_InterquartileRange
- original_firstorder_10Percentile
- original_firstorder_90Percentile
- original_firstorder_Maximum
- original_firstorder_RobustMeanAbsoluteDeviation

ویژگی original_firstorder_10Percentile نیز در هر چهار الگوریتم اجرا شده مشترک است.

در نتیجه، طبق ارزیابی و مقایسه‌ی نتایج به‌دست‌آمده در این پروژه، الگوریتم‌های ElasticNet و XGBoost که بیشترین F1 Score و بیشترین میزان هم‌پوشانی را داشتند، به‌عنوان قدرتمندترین الگوریتم‌ها در این زمینه انتخاب می‌شوند.

۷-۴ جمع‌بندی

در این فصل به بررسی نتایج حاصل از اعمال روش پیشنهادی بر روی مجموعه‌داده‌ها پرداخته شد. همان‌طور که در قسمت ارزیابی نتایج نیز ذکر شد، الگوریتم‌های ElasticNet و XGBoost بهترین عملکرد در زمینه‌ی انتخاب ویژگی از میان ویژگی‌های رادیومیکس، جهت افتراق میان ناحیه‌ی عفونی و اطراف آن در ریه را دارند.

فصل پنجم

نتیجه‌گیری و پیشنهادها

هدف از انجام این پروژه، افتراق ناحیه‌ی عفونی و اطراف آن با استفاده از ویژگی‌های رادیومیکس است. در روش پیشنهادی ابتدا ویژگی‌های رادیومیکس را برای مجموعه داده‌ها به دست آورده و در ادامه با استفاده از الگوریتم‌های انتخاب ویژگی، بهترین ویژگی‌ها جهت تمیز دادن نواحی عفونی و غیرعفونی انتخاب می‌شوند. نتایج به دست آمده نشان می‌دهند که الگوریتم‌های ElasticNet و XGBoost دارای بهترین عملکرد هستند. ویژگی‌های رادیومیکس به عنوان ورودی الگوریتم‌های یادگیری ماشین در نظر گرفته می‌شوند؛ بنابراین از ویژگی‌های به دست آمده از طریق اجرای الگوریتم‌های ElasticNet و XGBoost می‌توان به عنوان ورودی الگوریتم‌های طبقه‌بندی استفاده کرد.

اگر برای اجرای الگوریتم‌های یادگیری ماشین سیستم‌های پردازش قدرتمندی در اختیار باشد، می‌توان تمامی ویژگی‌های رادیومیکس را به عنوان ورودی برای آن‌ها اعمال کرد؛ اما اگر سیستم‌های پردازش موجود، قدرت بالایی نداشته و همچنین زمان مورد نیاز برای انجام پردازش محدود باشد، با استفاده از ویژگی‌های به دست آمده در این پروژه، از طریق الگوریتم‌های ElasticNet و XGBoost، می‌توان بهترین عملکرد را دریافت کرد. قدم بعدی مقایسه‌ی دقت به دست آمده از اجرای الگوریتم‌های طبقه‌بندی¹ در یادگیری ماشین است. الگوریتم‌های طبقه‌بندی برای یادگیری چگونگی تخصیص برچسب کلاس به یک نمونه ورودی استفاده می‌شوند. برخی از این الگوریتم‌ها عبارت‌اند از، الگوریتم Logistic Regression، Support Vector Machine، Decision Tree، Random Forest و غیره. حال اگر دقت به دست آمده از اجرای این الگوریتم‌ها مقدار مورد قبولی باشد، دیگر نیاز به پردازش‌های سنگین با استفاده از الگوریتم‌های یادگیری عمیق نیست؛ بنابراین به سادگی و در زمان کوتاهی می‌توان نواحی عفونی و غیرعفونی ریه را تشخیص داد. یک پیشنهاد جدیدی که می‌توان برای بهتر شدن پردازش تصاویر پزشکی در این زمینه ارائه کرد،

¹ Classification

اضافه کردن هسته‌ی^۱ رادیومیکس به یادگیری عمیق است. به این صورت که ویژگی‌هایی که قدرت بیشتری در تمییز دادن نواحی عفونی و غیرعفونی در ریه دارند، به عنوان هسته‌ی رادیومیکس به الگوریتم‌های یادگیری عمیق وارد شوند. هسته، روشی برای استفاده از طبقه‌بندی‌کننده‌ی خطی برای حل یک مسئله‌ی غیرخطی است.

^۱ Kernel

منابع

- [۱] تیم تولید محتوای پزشکی (۱۴۰۰). درمان عفونت ریه و انواع آن، بررسی علل ابتدا و راههای پیشگیری. [/https://www.paziresh24.com/blog/15348/hc-lung-infections](https://www.paziresh24.com/blog/15348/hc-lung-infections). [۲] جعفری، کسری (۱۳۹۸). ویکی آناتومی.
- [3] Park, Jongha, Jihye Yun, Namkug Kim, Beomhee Park, Yongwon Cho, Hee Jun Park, Mijeong Song, Minho Lee, and Joon Beom Seo. "Fully automated lung lobe segmentation in volumetric chest CT with 3D U-Net: validation with intra-and extra-datasets." *Journal of digital imaging* 33, no. 1 (2020): 221-230.
- [4] Singhal, Tanu. "A review of coronavirus disease-2019 (COVID-19)." *The indian journal of pediatrics* 87, no. 4 (2020): 281-286.
- [5] <https://www.polytarget.uni-jena.de/en/news/the+covid-19+puzzle>, Accessed August 1st, 2021.
- [6] van Timmeren, Janita E., Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. "Radiomics in medical imaging—"How-to" guide and critical reflection." *Insights into Imaging* 11, no. 1 (2020): 1-16.
- [7] Gillies, Robert J., Paul E. Kinahan, and Hedvig Hricak. "Radiomics: images are more than pictures, they are data." *Radiology* 278, no. 2 (2016): 563-577.
- [8] Tunali, Ilke, Robert J. Gillies, and Matthew B. Schabath. "Application of radiomics and artificial intelligence for lung cancer precision medicine." *Cold Spring Harbor perspectives in medicine* 11, no. 8 (2021): a039537.
- [9] <https://www.siemens-healthineers.com/de-ch/computed-tomography>, Accessed August 1st, 2021.
- [10] <https://podcastaddict.com/episode/120151362>, Accessed August 1st, 2021.
- [11] <http://www.futura-sciences.us/dico/d/medical-science-pet-50001141/>, Accessed August 1st, 2021.
- [12] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21), e104–e107. [‘https://doi.org/10.1158/0008-5472.CAN-17-0339](https://doi.org/10.1158/0008-5472.CAN-17-0339) <<https://doi.org/10.1158/0008-5472.CAN-170339>>’ _
- [13] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods."

Computers & Electrical Engineering 40, no. 1 (2014): 16-28.

- [14] Liu, Huan, and Rudy Setiono. "Chi2: Feature selection and discretization of numeric attributes." In Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388-391. IEEE, 1995.
- [15] Muthukrishnan, R., and R. Rohini. "LASSO: A feature selection technique in predictive modeling for machine learning." In 2016 IEEE international conference on advances in computer applications (ICACA), pp. 18-20. IEEE, 2016.
- [16] Grabczewski, Krzysztof, and Norbert Jankowski. "Feature selection with decision tree criterion. In null." (2005): 212-217.
- [17] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology) 67, no. 2 (2005): 301-320.
- [18] Chen, Cheng, Han Shi, Zhiwen Jiang, Adil Salhi, Ruixin Chen, Xuefeng Cui, and Bin Yu. "DNN-DTIs: improved drug-target interactions prediction using XGBoost feature selection and deep neural network." Computers in Biology and Medicine (2021): 104676.
- [19] Huang, Yuanpeng J., Robert Powers, and Gaetano T. Montelione. "Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics." Journal of the American Chemical Society 127, no. 6 (2005): 1665-1674.
- [20] Sun, Pan, Defeng Wang, Vincent CT Mok, and Lin Shi. "Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading." IEEE Access 7 (2019): 102010-102020.
- [21] Lee, Seung-Hak, Hwan-ho Cho, Ho Yun Lee, and Hyunjin Park. "Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer." Cancer Imaging 19, no. 1 (2019): 1-12.