
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی

مقایسه مدل‌های شبکه عصبی عمیق برای زیرگروه بندی سرطان بر اساس جهش‌های نقطه‌ای سوماتیک

نگارنده: پوریا پرهامی

استاد راهنما

دکتر منصور فاتح

اساتید مشاور

دکترحمید علی‌نژاد

دکتر محسن رضوانی

ماه و سال

شهریور ۱۴۰۰

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد آقای **پوریا پرهامی** با شماره دانشجویی **9803144** رشته **مهندسی کامپیوتر** گرایش **هوش مصنوعی و رباتیکز** تحت عنوان **مقایسه مدل های شبکه عصبی عمیق برای زیرگروه بندی سرطان بر اساس جهش های نقطه ای سوماتیک** که در تاریخ ۱۴۰۰/۰۶/۱۰ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار شد به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☑ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☐

ج) درجه خوب: نمره۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐

ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐

نوع تحقیق: نظری ☐ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استاد راهنمای اول	منصور فاتح	استادیار	
۳- استاد مشاور	محسن رضوانی حمید علی نژاد رکنی	دانشیار	
۴- استاد داور اول	مرتضی زاهدی	استادیار	
۵- استاد داور دوم	علیرضا تجری	استادیار	
۶- نماینده تحصیلات تکمیلی	محسن فرهادی	مربی	

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:

تقدیم به پدر و مادرم:

خدای را بسی شاکرم که از روی کرم، پدر و مادری فداکار نسیبم ساخته تا در سایه درخت پربار وجودشان بیاسایم و از ریشه آنها شاخ و برگ گیرم و از سایه وجودشان در راه کسب علم و دانش تلاش نمایم. والدینی که بودنشان تاج افتخاری است بر سرم و نامشان دلیلی است بر بودنم، چرا که این دو وجود، پس از پروردگار، مایه هستی ام بوده اند دستم را گرفتند و راه رفتن را در این وادی زندگی پر از فراز و نشیب آموختند. آموزگارانی که برایم زندگی، بودن و انسان بودن را معنا کردند.

تشکر قلبی و لسانی خود را از استاد عالی قدر جناب آقای دکتر فلاح که زحمت راهنمایی این پایان نامه را عهده دار گردیدند و در تمامی مراحل انجام رساله از راهنمایی های مدبّرانه ایشان استفاده نمودم ابراز می دارم و توفیقات روز افزون ایشان را توأم با صحت و سعادت خواستارم.

از جناب آقای دکتر رضوانی و دکتر علی نژاد که در امر مشاوره این رساله مساعدت نمودند و در این امر نهایت مراقبت، توجه و دقت خود را مبذول فرموده اند کمال تشکر و امتنان را دارم و برای ایشان از خداوند سلامت و سعادت ابدی را خواهانم.

از داوران گرامی که زحمت داوری و تصحیح این پایانامه را به عهده داشتند کمال سپاس را دارم.

خالصانه از تمامی اساتید و معلمان و مدرسانی که در مقاطع مختلف تحصیلی به من علم آموخته و مرا از سرچشمه دانایی سیراب کرده اند متشکرم.

تعهدنامه

اینجانب **پوریا پرهامی** دانشجوی دوره کارشناسی ارشد رشته هوش مصنوعی و رباتیکز دانشکده کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **مقایسه مدل‌های شبکه عصبی عمیق برای زیر گروه‌بندی سرطان براساس جهش‌های نقطه سوماتیک** تحت راهنمائی **دکتر فاتح** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

چکیده

تشخیص سریع و صحیح نوع سرطان یا زیرگروه‌های آن در بیماران تأثیر بسزایی در روند صحیح درمان و کاهش هزینه‌های درمانی دارد. روش‌های متنوعی برای تشخیص سرطان و زیرگروه‌های آن وجود دارد که با پیشرفت‌های اخیر در یادگیری ماشین و یادگیری عمیق دچار تحول شدند. همچنین روش‌های جدیدی با بهره‌مندی از یادگیری ماشین و یادگیری عمیق معرفی شده است که نتایج خوبی در پیش‌بینی انواع سرطان یا زیرگروه‌های آن و تشخیص تومورهای خوش‌خیم و بدخیم سرطانی در بیماران داشته است. یکی از روش‌هایی که امروزه برای تشخیص سرطان و زیرگروه‌های آن موردتوجه قرار گرفته، طبقه‌بندی سرطان مبتنی بر جهش سوماتیک است. قرارگرفتن در معرض اشعه ماوراء بنفش یا برخی مواد شیمیایی می‌تواند موجب جهش در سلول‌های بدن شود. به این جهش‌های غیرطبیعی که به‌وسیله عوامل محیطی ایجاد می‌شود جهش‌های سوماتیک می‌گویند. بااین‌وجود طبقه‌بندی سرطان مبتنی بر جهش در سوماتیک با چالش‌هایی مواجه است. چالش‌هایی مانند حجم کم داده نمونه، پراکندگی زیاد داده، بیش برازش و استفاده از طبقه‌بندی‌کننده‌های ساده خطی عواملی هستند که از افزایش عملکرد طبقه‌بندی جلوگیری می‌کنند. در این رساله روش‌هایی برای حل این چالش‌ها ارائه شده است. این روش‌ها شامل، پیش‌پردازش‌های فیلتر ژن خوشه‌ای، کاهش پراکندگی نمایه شده، روش‌های قاعده گذاری، لایه Global-Max-Pooling و استفاده از لایه embedding است. همچنین در این رساله سه مدل یادگیری عمیق CNN، LSTM و ترکیب این دو مدل بر روی مجموعه‌داده TCGA-DeepGene آزمایش شده است. مدل پیشنهادی ما یک مدل CNN تک‌لایه به همراه لایه embedding است. این مدل به دقت ۶۶,۴۵ درصد accuracy دست پیدا کرد. در مقایسه با مرجع مورد استناد در این رساله دقت افزایش ۱,۴۵ درصدی داشته است.

کلمات کلیدی: طبقه بندی نوع سرطان، جهش های De novo، شبکه عصبی عمیق، CNN، LSTM

لیست مقالات مستخرج از پایان نامه

P. Parhami, M. Fateh, M. Rezvani, H. Alinejad Rokny. “ A comparison of deep neural network models for cancer subtyping based on somatic point mutations” in Springer , Journal of Ambient Intelligence and Humanized Computing. ,Under review in May ۲۰۲۱

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل۱ : مقدمه

۱-۱ مقدمه

سرطان نام دسته‌ای از بیماری‌ها است که باعث اختلال در روند تولید و مرگ سلول‌های بدن می‌شود. چرخه تولد و مرگ سلول‌ها به این شکل است که سلول‌های بدن انسان رشد کرده، تقسیم می‌شوند و سلول‌های جدیدی را تشکیل می‌دهند. این سلول‌های جدید جایگزین سلول‌های آسیب‌دیده یا قدیمی می‌شوند که در شرف مرگ هستند. اختلال گفته شده در این روند به این معنا است که سلول‌های قدیمی یا آسیب‌دیده از بین نمی‌روند اما روند تولید سلول‌های جدید بدون توقف ادامه می‌یابد که به این ترتیب بافتی به نام تومور را تشکیل خواهند داد. این تومور یا تومورها می‌توانند خوش‌خیم یا بدخیم باشند. تومورهای بدخیم به بافت‌های اطراف خود حمله کرده و گسترش می‌یابند. با رشد این تومورها قسمتی از سلول‌های آنها از تومور جدا شده و با کمک جریان خون یا سیستم لنفاوی به سایر قسمت‌های بدن منتقل می‌شوند. این سلول‌ها می‌توانند به سرعت تکثیر شده و تومورهای دیگری را در سایر نقاط بدن بیمار ایجاد کنند. به این ترتیب این تومورها می‌توانند در سرتاسر بدن پراکنده شوند و جان بیمار را به خطر اندازند. به علت وجود چنین قابلیتی در تومورهای بدخیم اگر این تومورها از بدن خارج شوند امکان دارد دوباره در همان محل رشد کنند و یا در سایر نقاط بدن دیده شوند. تومورهای خوش‌خیم بر خلاف تومورهای بدخیم به بافت‌های اطراف خود حمله نمی‌کنند و در سرتاسر بدن پراکنده نمی‌شوند. اما می‌توانند بسیار بزرگ شوند. اگر این تومورها در اندام‌های حیاطی مانند مغز وجود داشته باشند رشد بیش از حد این تومورها می‌تواند منجرب به مرگ بیمار شود. همچنین برعکس تومورهای بدخیم در صورتی که از بدن خارج شوند دوباره رشد نخواهند کرد [۱].

سرطان مانند بسیاری از بیماری‌های دیگر، یک بیماری واحد نیست و دارای گروه‌ها و زیرگروه‌های فراوانی است. امروزه دانشمندان بیش از صد نوع مختلف از سرطان را شناسایی کرده‌اند. شایع‌ترین انواع سرطان‌ها در مردان و زنان، سرطان ریه می‌باشد. بعد از آن به ترتیبت سرطان پستان، سرطان پروستات و سرطان روده بزرگ قرار دارند. از نظر آمار مرگ‌ومیر در بین بیماران نیز بعد از سرطان ریه برای هر دو جنس به

ترتیب، سرطان روده بزرگ، سرطان معده و سرطان کبد قرار دارند. شایع‌ترین علت اصلی مرگ‌ومیر در بیماران مرد سرطان ریه است. به دنبال آن به ترتیب سرطان پروستات و سرطان روده بزرگ شایع‌ترین انواع سرطان در مردان هستند. همچنین سرطان کبد و سرطان معده کشنده‌ترین انواع سرطان بعد از سرطان ریه در مردان هستند. شایع‌ترین و کشنده‌ترین نوع سرطان در بیماران زن سرطان پستان است. بعد از سرطان پستان به ترتیب سرطان روده بزرگ و سرطان ریه قرار دارد. تخمین زده شده است که تنها در سال ۲۰۱۸ سرطان علت مرگ تعداد ۹,۶ میلیون نفر در دنیا بوده است. در این بین هرساله ۳۰۰،۰۰۰ مورد جدید ابتلا به سرطان بین افراد یک تا ۱۹ سال تشخیص داده می‌شود که تا به امروز فقط ۳۰ تا ۵۰ درصد از انواع سرطان قابل پیشگیری است [۱].

در سال‌های اخیر هوش مصنوعی و به ویژه یادگیری عمیق در زمینه‌های مختلفی از زندگی انسان وارد شده و تأثیرات چشم‌گیری در بهبود کیفیت زندگی انسان‌ها داشته است. مفهوم یادگیری عمیق در اواخر قرن بیستم مطرح شد. یادگیری عمیق، زیرمجموعه‌ای از الگوریتم‌های یادگیری ماشین است. این نوع یادگیری، به معنی استفاده از شبکه‌های عصبی داده‌محور با تعداد لایه‌های مخفی زیاد است. در یادگیری عمیق، به صورت خودکار، مهندسی ویژگی انجام می‌شود. بدین معنا که شبکه، به صورت خودکار و بدون نیاز به انسان، به دنبال بهترین ویژگی‌ها برای تمیز دادن کلاس‌های مختلف داده می‌گردد. در نهایت، با استفاده از ویژگی‌های بدست آمده، می‌تواند داده‌های ورودی را با بیشترین دقت و بالاترین کیفیت طبقه‌بندی کند. این ویژگی اساسی یادگیری عمیق، باعث می‌شود وابستگی سیستم به انسان در مرحله استخراج ویژگی کاهش پیدا کند [۲]. موفقیت چشمگیر در یادگیری عمیق زمانی رخ داد، که هینتون[۱] در سال ۲۰۰۶ معماری جدیدی از یادگیری عمیق به نام شبکه‌های باور عمیق(DBN) [۲] را ارائه کرد [۳]. برخلاف رویکردهای سنتی یادگیری ماشین و هوش مصنوعی، فناوری‌های یادگیری عمیق پیشرفت‌های چشمگیری در کاربردهای تشخیص گفتار، پردازش زبان طبیعی، بینایی ماشین، تجزیه و تحلیل تصاویر، مراقبت‌های

[۱] Hinton

[۲] Deep belief network

بهداشتی و بازیابی اطلاعات داشته‌اند [۴].

باتوجه‌به تنوع زیاد انواع سرطان و زیرگروه‌های آن و روش‌های متفاوت برای درمان هر یک از انواع سرطان همچنین تعداد زیاد مبتلایان و خطر مرگ‌ومیر بالا در بیماران، تشخیص یا پیش‌بینی سریع و صحیح نوع یا انواع سرطان و زیرگروه‌های آن‌ها می‌تواند، در تعیین نوع درمان و کاهش مرگ‌ومیر مبتلایان به این بیماری مؤثر باشد. دراین‌بین تشخیص سرطان و تومورهای سرطانی با استفاده از روش‌های یادگیری عمیق در تصاویر پزشکی گامی مهم در درمان این بیماری بوده است. یکی دیگر از مهم‌ترین روش‌های طبقه‌بندی سرطان، طبقه‌بندی مبتنی بر جهش در نقطه سوماتیک (SMCC)[۱] می‌باشد [۵] که دارای چالش‌های جدیدی نسبت به روش‌های تشخیص سرطان در تصاویری پزشکی بوده و نمی‌توان به راحتی روش‌های استفاده شده در تصویر را بر روی این داده‌ها استفاده کرد.

جهش‌های ژنتیکی در بدن انسان را می‌توان به دودسته کلی به نام‌های جهش‌های ارثی و جهش‌های سوماتیک تقسیم کرد. منظور از جهش ارثی، جهش‌هایی هستند که به دلیل تولید مثل یک موجود نر و ماده در فرزند آنها رخ می‌دهد. این جهش‌ها معمولاً در تخمک یا اسپرم فرد رخ می‌دهد و به نسل‌های بعدی نیز منتقل شود. اما جهش‌های اکتسابی (سوماتیک) آن دسته از جهش‌ها هستند که بر اثر وجود عوامل زیست‌محیطی در بدن انسان‌ها رخ می‌دهند. برای مثال هنگامی که افراد در معرض اشعه ماوراء بنفش، مواد شیمیایی قوی یا مواد رادیواکتیو قرار می‌گیرند، سلول‌های بدن آنها دچار جهش می‌شوند. یا اگر خطایی در تقسیم سلولی رخ دهد می‌تواند منجرب به جهش در آن سلول شود. این جهش‌ها می‌توانند در هر قسمت از بدن رخ دهند. اگر این جهش‌ها در سلول‌هایی به غیر از تخمک و اسپرم رخ دهند به نسل بعدی منتقل نمی‌شوند. معمولا این جهش‌ها علت بروز انواع سرطان در انسان‌ها می‌باشند.

[۱] Somatic point mutation based cancer classification

۲-۱ بیان مسئله

باتوجه‌به کاهش هزینه توالی‌یابی DNA در سال‌های اخیر و دسترسی گسترده به این داده‌ها استفاده از SMCC برای تشخیص انواع سرطان گسترش یافته است. هدف اصلی روش‌های SMCC تشخیص انواع مختلف سرطان یا زیرگروه‌های آن بر اساس جهش‌های ژنی سوماتیک است. در مقایسه با روش‌های طبقه‌بندی مرسوم سرطان که بیشتر بر اساس ظاهر مورفولوژیکی[۱] یا بیان ژن[۲] تومور عمل دسته‌بندی را انجام می‌دهند. SMCC در تشخیص و تمایز تومورها با ظاهر هیستوپاتولوژیک[۳] مشابه عملکرد خوبی داشته است و در برابر تأثیرات محیطی مقاوم است. از نظر بالینی، SMCC می‌تواند درمان و تشخیص سرطان را سرعت ببخشید و ساده‌تر کند که همین امر موجب کاهش هزینه‌های تشخیص و درمان می‌شود. می‌توان از SMCC در کارهایی مانند، تومور درمانی هدفمند [۶] و ساخت داروهای ترکیبی [۷] استفاده شود و به تشخیص زود هنگام سرطان (CED) [۴] نیز کمک کند [۸, ۹].

بر اساس دلایل ذکر شده، امروزه SMCC در تحقیقات مختلف به طور گسترده‌ای مورد مطالعه و پژوهش قرار گرفته است [۱۰, ۱۱]. پیشرفت‌های اخیر در زمینه یادگیری ماشین باعث شده است که الگوریتم‌های این حوزه برای بهبود روش‌های SMCC به کار گرفته شوند. الگوریتم‌های یادگیری ماشین در زمینه‌های مربوط به تشخیص تومور روده بزرگ [۱۲] و تومور پستان [۱۳] موفقیت‌هایی را کسب کرده‌اند. با این حال هنوز تحقیقات جامعی در مورد روش‌های یادگیری عمیق و تاثیر آن در بهبود عملکرد روش‌های SMCC انجام نشده است. در این رساله قصد داریم تا تاثیر سه مدل یادگیری عمیق را برای تشخیص سرطان بر اساس جهش‌ها سوماتیک مورد بررسی قرار دهیم.

---

[۱] Morphological

[۲] Gene Expressions

[۳] Histopathological

[۴] Cancer early diagnosis

۳-۱ چالش‌های موجود در روش‌های SMCC

روش‌های پیشین در SMCC با مشکلاتی مواجه هستند. این مشکلات شامل: کارایی نامناسب بر روی مجموعه‌داده‌های دودویی[۱]، استفاده از طبقه‌بندهای ساده خطی، مانند هسته خطی ماشین بردارپشتیبان (SVM) [۲] ، بیش برازش[۳] در مدل‌های عمیق و تشخیص ژن‌های تأثیرگذار در کار دسته‌بندی است. منظور از داده‌های دودویی در این رساله داده‌های یک و صفر است. یک به معنی وقوع جهش در ژن و صفر به معنی عدم وقوع جهش در ژن است.

یک توالی DNA شامل تعداد بسیار زیادی ژن است که فقط یک زیرمجموعه کوچک از این ژن‌ها می‌تواند به ما در تشخیص و دسته‌بندی سرطان کمک کند. فعالیت‌هایی برای ساخت چنین زیرمجموعه‌ای انجام شده است مانند اعمال میانگین و انحراف معیار استاندارد بر روی فاصله هر نمونه تا مرکز کلاس یا انتخاب ژن‌های خوشه‌بندی‌شده که خوشه‌بندی را به‌وسیله کا - میانگین[۴] انجام می‌دهد و ژن‌های برتر هر خوشه که به مرکز خوشه نزدیک‌تر هستند را انتخاب می‌کند. این روش‌ها بر روی برخی داده‌های پیوسته ساده مؤثر هستند اما هنگام روبه‌رو شدن با داده‌های گسسته مخصوصاً داده‌های دودویی عملکرد مناسبی ندارند [۱۴, ۱۵]. برای رفع این مشکل ما از پیش پردازش استفاده شده در مرجع [۵] به نام فیلتر ژن خوشه بندی شده و لایه جاسازی[۵] استفاده می‌کنیم.

حتی پس از ایجاد یک زیرگروه خاص از ژن‌ها، همه اطلاعات موجود در آن به دلیل وجود تعداد زیاد صفر در آن برای کار دسته‌بندی مناسب نیست. این تعداد زیاد صفر باعث پراکندگی در داده‌های ما شده و کارایی مدل‌ها را کاهش می‌دهد. برای حل این مشکل ما از پیش‌پردازش معرفی شده در مرجع [۵] به نام کاهش پراکندگی فهرست شده و لایه جاسازی استفاده می‌کنیم.

---

[۱] Binary
[۲] Support vector machine
[۳] Overfitting
[۴] K-means
[۵] Embedding

در هر یک از انواع خاص سرطان ژن‌هایی حضور دارند که این ژن‌ها عموماً با یکدیگر ارتباط دارند و دارای فعل و انفعالات پیچیده‌ای هستند. این فعل‌وانفعالات پیچیده و ارتباطات ژن‌ها مانع از عملکرد مناسب طبقه‌بندهای ساده خطی مانند SVM می‌شود؛ بنابراین به طبقه‌بندهای پیشرفته‌ای نیاز است که بتوانند ویژگی‌های سطح بالا را از زیرمجموعه منتخب ایجاد شده استخراج کنند. در این رساله ما برای رفع این مشکل سه مدل یادگیری عمیق CNN[۱]، LSM[۲] و ترکیب این دو مدل را مورد بررسی قرار دادیم.

یکی از چالش‌های رایج در مدل‌های عمیق مشکل بیش برازش[۳] است. برای حل این مشکل در مدل‌های خود ما از روش‌های نظم‌دهی[۴] و لایه پولینگ حداکثر سراسری[۵] استفاده می‌کنیم.

۴-۱ هدف پایان‌نامه

ایجاد SMCCهمان‌طور که در بخش قبل گفته شد، روش‌های متنوعی با استفاده از یادگیری ماشین در شده است که هر یک قصد رفع یک چالش را داشتند. اما هیچ یک از این روش‌ها کامل و بی‌نقص نیستند و هنوز چالش‌های حل نشده‌ای باقی است که می‌توان برای رفع یا بهبود آنها تلاش کرد. مشکلاتی مانند استفاده از طبقه‌بندهای ساده خطی، پراکندگی داده‌ها و یافتن ویژگی‌های تأثیرگذار در کار دسته‌بندی مواردی از این چالش‌ها هستند. همچنین تأثیر مدل‌های یادگیری عمیق در این زمینه هنوز به صورت دقیق موردمطالعه قرار نگرفته است. هدف این پایان‌نامه ارائهٔ راهکاری هوشمند و دقیق برای رفع چالش‌های ذکر شده، به همراه افزایش دقت دسته‌بندی است. براین‌اساس این پایان‌نامه بر روی موارد زیر تمرکز دارد:

- یافتن راهی برای ازبین‌بردن پراکندگی زیاد داده‌ها
- آزمایش سه مدل یادگیری عمیق و تأثیر عملکرد آنها در کار دسته‌بندی

---

[۱] Convolutional Neural Network
[۲] Long Short - Term Memory
[۳] Overfitting
[۴] Regularization
[۵] Global Max Pooling

- یافتن روشی برای استخراج ویژگی‌های سطح بالاتر برای بهبود عملکرد دسته‌بندی
- استفاده از روش‌هایی کارآمد برای کاهش اثر بیش برازش در مدل پیشنهادی

۵-۱ روش تحقیق

در این تحقیق ما مرجع [۵] را مبنای کار خود قرار دادیم. به این ترتیب برای مقایسه نتایج و عملکرد مدل‌ها و برای رعایت عدالت در انجام آزمایش‌ها و مقایسه مناسب نتایج، از مجموعه‌داده‌های گرد آوری شده به نام TCGA-DeepGene استفاده کردیم. این مجموعه داده داری جهش‌های سوماتیک دوازده نوع مختلف سرطان می‌باشد.

ابتدا مراحل پیش‌پردازش را بر روی داده‌ها اعمال می‌کنیم. تغییر نوع داده‌ها به دودویی و اعمال دو پیش‌پردازش برای کاهش پراکندگی داده و یافتن زیرمجموعه‌ای از ژن‌های مؤثر در کار دسته‌بندی از جمله مراحل پیش‌پردازش می‌باشد.

سپس در مرحله بعد، تحقیق و پژوهشی دررابطه با مدل‌های یادگیری عمیق برای دسته‌بندی و تشخیص سرطان انجام شده است. بر اساس این تحقیق سه مدل که بیشترین کاربرد را در کار دسته‌بندی سرطان، تشخیص تومور و بافت‌های خوش‌خیم و بدخیم داشته‌اند. برای این تحقیق انتخاب شده‌اند. در قدم بعدی با استفاده از مجموعه‌داده‌ها و پیش‌پردازش‌های اعمال شده بر روی آن، داده‌ها را برای ورود به مدل‌های خود آماده می‌کنیم. بعد از پیاده‌سازی مدل‌ها و تنظیم پارامترهای آنها نتایج مدل‌های خود را با نتایج مرجع [۵] مقایسه می‌کنیم و بهترین مدل را برای تحقیق در آینده معرفی خواهیم کرد.

۶-۱ نوآوری مسئله

در این رساله، روش جدیدی برای دسته‌بندی سرطان بر اساس جهش‌های سوماتیک ارائه شده است که منجر به افزایش دقت دسته‌بندی به میزان ۱,۴۵ درصد شده است. این روش بر روی مجموعه‌داده‌های TCGA-DeepGene آزمایش شده و می‌توان آن را بر روی مجموعه‌داده‌هایی با مقادیر دودویی، اعمال کرد.

در این روش برای رفع پراکندگی داده‌ها از دو پیش‌پردازش فیلتر ژن خوشه‌ای، کاهش پراکندگی فهرست شده و لایه embedding استفاده شده است. روش‌های نظم‌دهی و لایه پولینگ حداکثر سراسری[۱] برای رفع بیش برازش و برای استخراج ویژگی‌ها سطح بالاتر از مدل CNN استفاده کرده‌ایم. همچنین در انتها، خروجی ایجاد شده به یک‌لایه متراکم[۲] برای دسته‌بندی وارد می‌شود.

۷-۱ ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل است. در فصل دوم تعدادی از فعالیت‌های پیشین در رابطه با تشخیص سرطان موردبحث و بررسی قرار خواهد گرفت. در فصل سوم تمامی قسمت‌های روش پیشنهادی ارائه شده معرفی می‌گردد، در فصل چهارم به بیان نتایج حاصل از روش پیشنهادی و مقایسه مدل‌ها می‌پردازیم و در فصل پایانی به بیان یک جمع‌بندی کلی از موضوع، به همراه فعالیت‌هایی که در آینده می‌توانند در این زمینه انجام گیرند، می‌پردازیم.

[۱] Global Max Pooling

[۲] Dense Layer

فصل ۲ : کارهای انجام شده

۲-۱ مقدمه

در فصل قبل به صورت کلی در مورد بیماری سرطان و اهمیت تشخیص سریع و صحیح این بیماری صحبت کردیم. در سال‌های اخیر پیشرفت‌های چشمگیر در زمینه هوش مصنوعی، مخصوصاً یادگیری عمیق باعث شده است بسیاری از الگوریتم‌ها و مدل‌های این حوضه در بسیاری از فعالیت‌های مختلف مورداستفاده گسترده قرار گیرند. سیستم‌های پزشکی و بهداشتی نیز یکی از همین زمینه‌ها می‌باشد که یادگیری عمیق کمک شایانی به پزشکان در تشخیص صحیح و سریع بیماری‌ها داشته است. مدل‌های یادگیری عمیق که در تشخیص و دسته‌بندی بیماری‌ها مخصوصاً در تشخیص و دسته‌بندی سرطان مورداستفاده قرار گرفته‌اند، می‌توانند ما را در انتخاب مدل‌های خودیاری رسانند. در این فصل ابتدا به توضیح مدل‌ها پرداخته سپس به بررسی برخی از پژوهش‌های انجام شده با استفاده از این مدل‌ها در حوزه دسته‌بندی و تشخیص سرطان می‌پردازیم.

۲-۲ شبکه عصبی کانولوشنی (CNN)

یکی از زیرگروه‌های متمایز و موفق یادگیری عمیق CNN است که عملکرد خوبی روی داده‌های دوبعدی با توپولوژی شبکه‌ای[۱] از خود نشان داده است. معماری CNN از قشر بینایی حیوانات و مفهوم آن از شبکه‌های عصبی با تأخیر در زمان (TDNN) [۲] الهام گرفته شده است. یکی از ویژگی‌های CNN لایه کانولوشن می‌باشد که باعث کاهش تعداد وزن‌ها و در نتیجه کاهش پیچیدگی می‌شود. همچنین یک تصویر می‌تواند به‌عنوان یک داده خام به صورت مستقیم وارد این مدل شود [۱۶, ۱۷].

CNN به دلیل ساختار سلسله‌مراتبی خود اولین معماری یادگیری عمیق موفق بود. CNN با تقویت روابط خاص تعداد پارامترهای شبکه را کاهش می‌دهد و عملکرد خود را با استفاده از الگوریتم‌های استاندارد

[۱] Grid-like topology

[۲] Time-delay neural networks

انتشار رو به عقب[۱] بهبود می‌بخشد. یکی دیگر از ویژگی‌های CNN این است که به حداقل پیش‌پردازش نیاز دارد. این مدل توانسته است در فعالیت‌هایی نظیر، تشخیص دست خط[۲]، تشخیص چهره[۳]، تشخیص رفتار[۴]، تشخیص گفتار[۵]، سیستم‌های پیشنهاددهنده[۶]، طبقه‌بندی تصاویر[۷] و پردازش زبان طبیعی (NLP) [۸] عملکرد خوبی داشته باشد [۱۶, ۱۷].

۱-۲-۲ ساختار شبکه CNN

CNN دارای ساختاری سلسله‌مراتبی است که این ساختار شامل سه‌لایه اصلی به نام‌های، لایه کانولوشن[۹]، لایه پول[۱۰] [۱۸]، و لایه تمام متصل (FC) [۱۱] می‌باشد. ساختار قرار گیری این لایه‌ها معمولا به این شکل است، که ابتدا لایه کانولوشن قرار می‌گیرد، سپس لایه پول بعد از لایه کانولوشن قرار می‌گیرد، و در انتهای مدل یک یا چند لایه تمام متصل قرار دارد. لایه‌های کانولوشن و پول می‌توانند چندین بار تکرار شوند و به ترتیب پشت سرهم قرار گیرند، یا ابتدا چندین لایه کانولوشن وجود داشته باشد و سپس لایه پول و در آخر لایه تمام متصل قرار گیرد. شکل قرار گیری لایه‌ها می‌تواند در کاربردهای مختلف متفاوت باشد. در یک مدل استاندارد و معمولی CNN معمولا از طراحی رو به جلو[۱۲] استفاده می‌شود به این شکل که خروجی لایه قبل به عنوان ورودی به لایه بعدی داده می‌شود [۱۶, ۱۷].

---

۱ Back-propagation alghorithms
۲ Handwrite Recognition
۳ Face Detection
۴ Behavior Recognition
۵ Speech Recognition
۶ Recommender Systems
۷ Image Classification
۸ Natural Language Processing
۹ Convolutional Layer
۱۰ Pooling Layer
۱۱ Fully Connected
۱۲ Feed-Forward

۲-۲-۲ لایه کانولوشن

این لایه پرکاربردترین لایه در CNN است و بیشترین استفاده را دارد. بخش اعظم محاسبات بر عهده این لایه است و نقش اصلی آن استخراج ویژگی‌ها از تصویر یا نقشه ویژگی‌هایی است که توسط لایه یا لایه‌های قبلی ایجاد شده و به‌عنوان ورودی وارد این لایه می‌شود. هر لایه کانولوشن با استفاده از چند فیلتر یا هسته[۱] با اندازه گام S این نقشه ویژگی‌ها را به‌عنوان خروجی تولید می‌کند.

فرمول شماره ۲-۱ نحوه محاسبه عمل گفته شده را نشان می‌دهد. در این فرمول m نشان‌دهنده خروجی‌ها، K_1 و K_2 نشان دهنده ارتفاع و عرض هسته کانولوشن، S نشان دهنده اندازه گام و f تابع فعال ساز است که می‌تواند رلو[۲]، سیگموید[۳] یا تانژانت[۴] هیپربولیک باشد [۱۶, ۱۷].

$$\text{output}[m][r][c] = f\left(\sum_{n=0}^{N}\sum_{i=0}^{K_1}\sum_{j=0}^{K_2} \text{weight}[m][n][i][j] \times \text{Input}[n][S \times r + i][S \times c + j] + \text{bias}[m]\right) \quad (۱-۲)$$

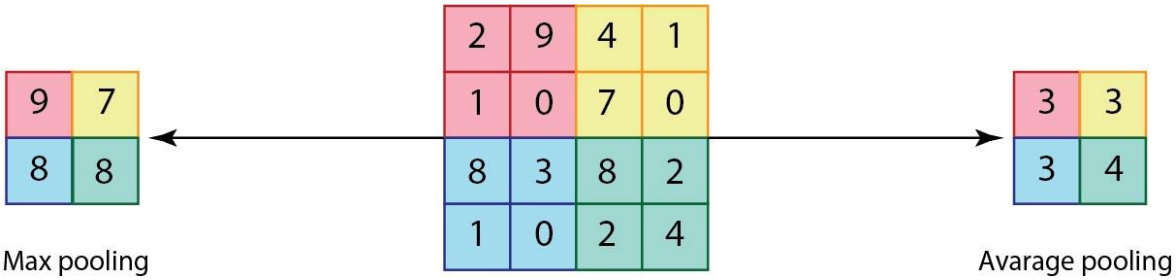
۲-۲-۳ لایه پولینگ

لایه پولینگ معمولاً بعد از یک یا چندلایه کانولوشن می‌آید و با ترکیب خروجی خوشه‌های نرون در یک‌لایه به یک نرون واحد، در لایه بعد ابعاد نقشه ویژگی را کاهش می‌دهد. همین امر از پردازش زیاد جلوگیری می‌کند، همچنین در این عمل ویژگی‌های مهم باقی می‌مانند و مابقی حذف می‌شوند. دو روش پرکاربرد از لایه پولینک، پولینگ حداکثر[۵] و پولینگ میانگین[۶] نام دارند.

Average Pooling مقدار میانگین یک محدوده و Max Pooling بیشترین مقدار یک محدوده را از یک منطقه از نقشه ویژگی محاسبه می‌کند. شکل شماره ۲-۱ این دو روش را نشان می‌دهد.

---

[۱] kernel
[۲] Relu
[۳] Sigmoid
[۴] Tangent Hyperbolic
[۵] Max Pooling Layer
[۶] Average Pooling Layer



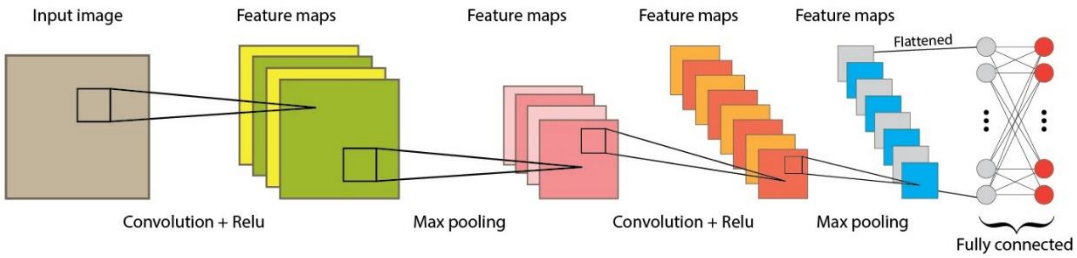
شکل ۲-۱ ماتریس سمت چپ نتیجه اعمال پولینگ حداکثر بر روی ماتریس وسط می‌باشد. ماتریس سمت راست نتیجه اعمال پولینگ میانگین بر روی ماتریس وسط می‌باشد مرجع[۱۶] .

لایه تمام متصل

معمولاً در انتهای مدل CNN یک‌لایه تمام متصل وظیفه دسته‌بندی را به عهده دارد. برای اینکه نقشه‌های ویژگی تولید شده توسط CNN وارد این لایه بشود، ابتدا باید به یک وکتور تبدیل شوند. محاسبات انجام شده در این لایه را می‌توان در فرمول ۲-۲ دید [۱۶, ۱۷].

$$Output[m] = f(\sum_{n=0}^{N} weight[m][n] \times Input[n] + bais[m]) \quad (۲-۲)$$

شکل شماره ۲-۲ مدل ساده‌ای از CNN را همراه با لایه‌های آن نشان می‌دهد.



شکل ۲-۲ ساختار مدل CNN استاندارد [۱۶, ۱۷].

۲-۳ شبکه عصبی LSTM

شبکه LSTM نوع خاصی از شبکه عصبی بازگشتی (RNN)[۱] می‌باشد که در سال ۱۹۹۷ توسط هوخرایتر[۲] و اشمیت هوپر[۳] معرفی شد. RNN تابعی است، که یک ورودی با طول غیرثابت مانند یک جمله را به صورت دنباله‌ای از n بردار d_{in} بعدی دریافت می‌کند و یک بردار خروجی y با ابعاد d_{out} را بازمی‌گرداند. فرمول ۲-۳ این رابطه را نشان می‌دهد [۱۹].

$$y_n = \mathrm{RNN}\,(x_{1:n}) \qquad (۳-۲)$$

$$x_i \in R^{d_{in}} \quad y_i \in R^{d_{out}}$$

RNN استاندارد توسط پس انتشار خطا آموزش می‌بیند، این شبکه در یادگیری دنباله‌های بزرگ دچار خطا می‌شود. این روند آموزش، موجب ایجاد مشکل ناپدید شدن[۴] یا انفجار شیب[۵] می‌شود. برای حل این مشکل سلول RNN با چند دروازه جایگزین می‌شود. LSTM ها می‌توانند وابستگی‌های طولانی‌مدت را یاد بگیرند و به صورت صریح برای همین منظور طراحی شده‌اند. به‌خاطر سپردن اطلاعات برای مدت طولانی عمل پیش‌فرض در آنها است، نه چیزی که برای یادگیری آن تلاش می‌کنند. همچنین این مدل مشکل ناپدید شدن شیب که RNN ها به آن دچار هستند را حل کرده است.

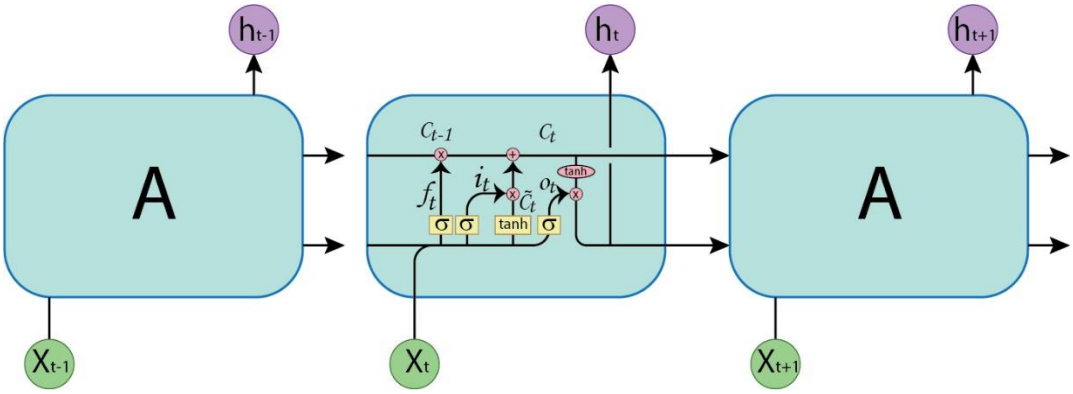
۲-۳-۱ ساختار شبکه عصبی LSTM

همه شبکه‌های عصبی به شکل دنباله‌ای (زنجیره‌ای) تکرارشونده از ماژول‌های شبکه عصبی هستند. در شبکه‌های عصبی مکرر استاندارد، این ماژول‌های تکرارشونده ساختار ساده‌ای دارند. ماژول‌های تکرارشونده می‌توانند ساختار متفاوتی داشته باشند. این ساختارها، طبق قواعد ویژه‌ای، با هم در تعامل و ارتباط هستند.

---

[۱] Recurrent Neural Network
[۲] Hochreiter
[۳] Schmidhuber
[۴] Vanishing Gradient
[۵] Exploding Gradient

مدل LSTM دارای سه دروازه به نام‌های ورودی[۱]، فراموشی[۲] و خروجی[۳] می‌باشد. این دروازه‌ها وظیفه انتخاب سیگنالی را دارند که باید به گره[۴] بعدی برود. شکل ۲-۳ شکلی از سلول LSTM را نشان می‌دهد [۲۰].



شکل ۲-۳ سلول‌های LSTM و نمای داخلی از یک سلول مرجع [۱۹].

محاسباتی که در یک سلول روی می‌دهد به این شکل است، ابتدا باید تصمیم گرفته شود کدام اطلاعات در وضعیت فعلی سلول دور انداخته شود، این تصمیم به‌وسیله لایه سیگموید که Forget Gate Layer نام دارد انجام می‌شود، فرمول شماره ۲-۴ این محاسبه را نشان می‌دهد.

$$f_t = \sigma(x_t U^f + h_{t-1} W^f) \qquad (۴-۲)$$

قدم بعدی این است که تصمیم بگیرم چه اطلاعاتی را در سلول ذخیره کنیم، این مرحله شامل دو قسمت است. قسمت اول لایه سیگموید که Input Gate Layer نام دارد و تصمیم می‌گیرد چه مقادیری به‌روز شوند و قسمت دوم لایه تانژانتی است که یک وکتور از مقادیر کاندید را ایجاد می‌کند. این وکتور می‌تواند

[۱] Input Gate
[۲] Forget Gate
[۳] Output Gate
[۴] Node

به حالت فعلی سلول اضافه شود. سپس این دو با هم ادغام می‌شوند تا یک به‌روزرسانی برای حالت فعلی ایجاد شود فرمول ۲-۵ این محاسبه را نشان می‌دهد.

$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right)$$

$$\tilde{C}_t = tanh(x_t U^g + h_{t-1} W^g) \qquad (۵-۲)$$

بعد از این محاسبات زمان آن است تا وضعیت قدیمی سلول C_{t-1} را به روز رسانی کنیم. و در آخر تصمیم بگیرم چه خروجی تولید شود، فرمول شماره ۲-۶ این محاسبات را نمایش می‌دهد. در این فرمول‌ها W

$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right)$$

$$o_t = \sigma(x_t U^o + h_{t-1} W^o)$$

$$h_t = tanh(C_t) * o_t \qquad (۶-۲)$$

نشان دهنده ارتباط بین لایه قبلی و لایه مخفی فعلی است، U ماتریس وزن‌ها، $\check{C}$ نشان‌دهنده کاندیدها و C حافظه داخلی واحد مورد نظر است.

۲-۴ کارهای انجام شده

در این بخش، به بررسی برخی از کارهای انجام شده در زمینه تشخیص و دسته‌بندی سرطان با استفاده از یادگیری عمیق می‌پردازیم.

در مرجع [۲۱] برای تشخیص و دسته بندی دو نوع از شایع ترین زیر گروه‌های سرطان ریه به نام‌های LUAD [۱] و LUSC [۲] از یک شبکه کانولوشنی به نام Inception V۳ [۲۲] استفاده شده است. برای دسته بندی این دو نوع سرطان معمولا به یک آسیب شناس حرفه‌ای نیاز است، اما مدل ساخته شده عملکردی

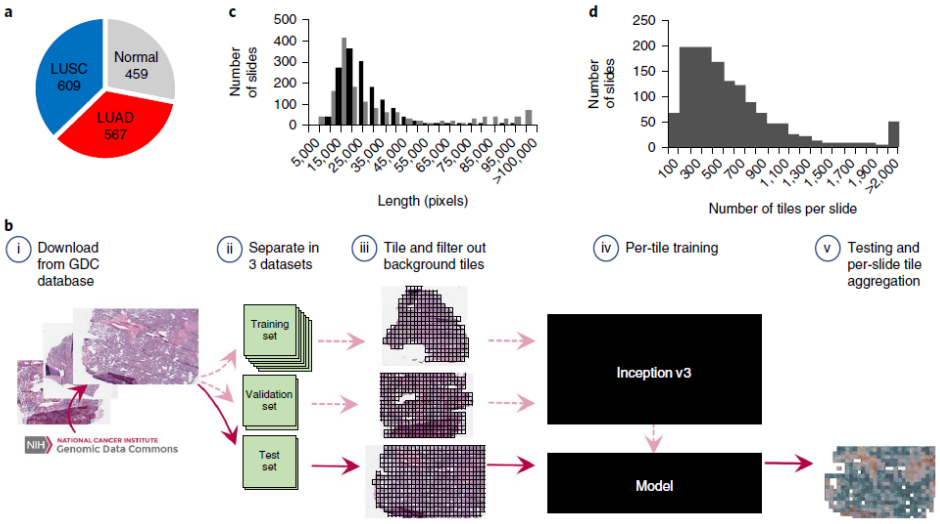
[۱] Lung Adenocarcinoma

[۲] Lung Squamous cell carcinoma

مشابه با یک آسیب شناس حرفه‌ای دارد. همچنین مدل را آموزش دادند تا رایج‌ترین ده ژنی که عموما جهش پیدا می‌کنند را پیش بینی کند.

Inception v۳ یک شبکه عصبی کانولوشنی برای کمک به تجزیه‌وتحلیل شکل و تشخیص جسم است که به‌عنوان ماژولی برای GoogleNet شروع به کارکرد. این مدل سومین نسخه از شبکه عصبی مبتنی بر Google's inception convolutional neural network است و دارای ۴۸ لایه می‌باشد.

یکی از موارد مهم در این مدل حفظ عملکرد هنگام آزمایش در مجموعه‌داده‌های مستقل است. دو مجموعه‌داده استفاده شده در این کار بافت‌های حاوی پارافین منجمد و ثابت شده با فرمالین (FFPE)[۱] است و دیگری تصاویر بدست آمده از نمونه‌برداری می‌باشد. شکل شماره ۲-۴ روند انجام استراتژی آموزش را به صورت خلاصه نمایش می‌دهد.



شکل ۲-۴ استراتژی استفاده شده در مرجع [۲۱]

به ترتیب حروف الفبا، a تعداد تصاویر کل اسلاید در هر کلاس است، b استراتژی آموزش است، b(i) تصاویر

[۱] Formalin-fixed Paraffin Embedded

بافت‌های سرطانی ریه هستند که برای اولین‌بار از پایگاه‌داده Genomic Data Commons بارگیری می-شوند. در قسمت b(ii) اسلایدها به یک مجموعه آموزش ۷۰ درصدی، یک مجموعه اعتبارسنجی ۱۵ درصدی و یک مجموعه آزمون ۱۵ درصدی تقسیم می‌شوند. در b(iii) اسلایدها توسط پنجره‌هایی با ابعاد ۵۱۲ در ۵۱۲ که با یکدیگر هم‌پوشانی ندارند کاشی‌کاری می‌شوند، سپس آن‌هایی که بیش از ۵۰ درصد پس‌زمینه دارند حذف می‌شوند. در قسمت b(iv) بعد از انجام مراحل گفته شده inception v۳ به طور جزئی یا کامل با استفاده از مجموعه آموزش و اعتبارسنجی بازآموزش می‌شود. در قسمت b(v) دسته‌بندی بر روی یک مجموعه آزمون انجام می‌شود و در نهایت نتایج برای استخراج نقشه‌های حرارتی[۱] و سطح زیر نمودار (AUC) [۲] در هر اسلاید جمع می‌شود. علامت c توزیع اندازه عرض تصاویر (خاکستری) و ارتفاع (سیاه) می‌باشد و علامت d توزیع تعداد کاشی‌ها در هر اسلاید است.

مدل مورد استفاده در این مطالعه توانست به‌دقت (۰,۹۹ AUC) برای تشخیص تومورهای عادی و سرطانی و به‌دقت (۰,۹۷ AUC) برای تشخیص چند نوع گفته شده از سرطان ریه دست پیدا کند. این مدل به دقتی بین ۰,۷۳۳ تا ۰,۸۵۶ در تشخیص ده مورد از رایج‌ترین جهش‌ها در LUAD نیز دست‌یافته است.

سلول‌های خونی عمدتاً شامل گلبول‌های قرمز، گلبول‌های سفید و پلاکت‌ها هستند. معمولاً متخصصان خون از اطلاعات دانه‌بندی‌شده و اطلاعات مربوط به شکل گلبول‌های سفید برای تقسیم گلبول‌های سفید به سلول‌های دانه‌ای و غیر دانه‌ای استفاده می‌کنند. نسبت وجود این سلول‌ها در بیماران و افراد سالم متفاوت است و پزشکان اغلب از این اطلاعات برای تعیین نوع و شدت بیماری استفاده می‌کنند. به همین دلیل مطالعه و طبقه‌بندی گلبول‌های سفید خون از اهمیت و ارزش مهمی برای تشخیص پزشکی برخوردار است.

مرجع [۲۳] برای دسته بندی سلول‌های خون در تصاویر پزشکی از ترکیب Xception[۲۴] و LSTM استفاده می‌کند. Xception یک مدل تعمیم یافته از معماری Inception است که در آن ماژول‌های

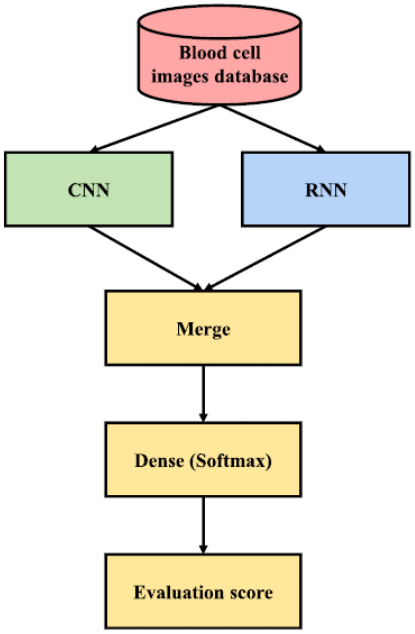
---

۱ Heat-map

۲ Area Under Curve

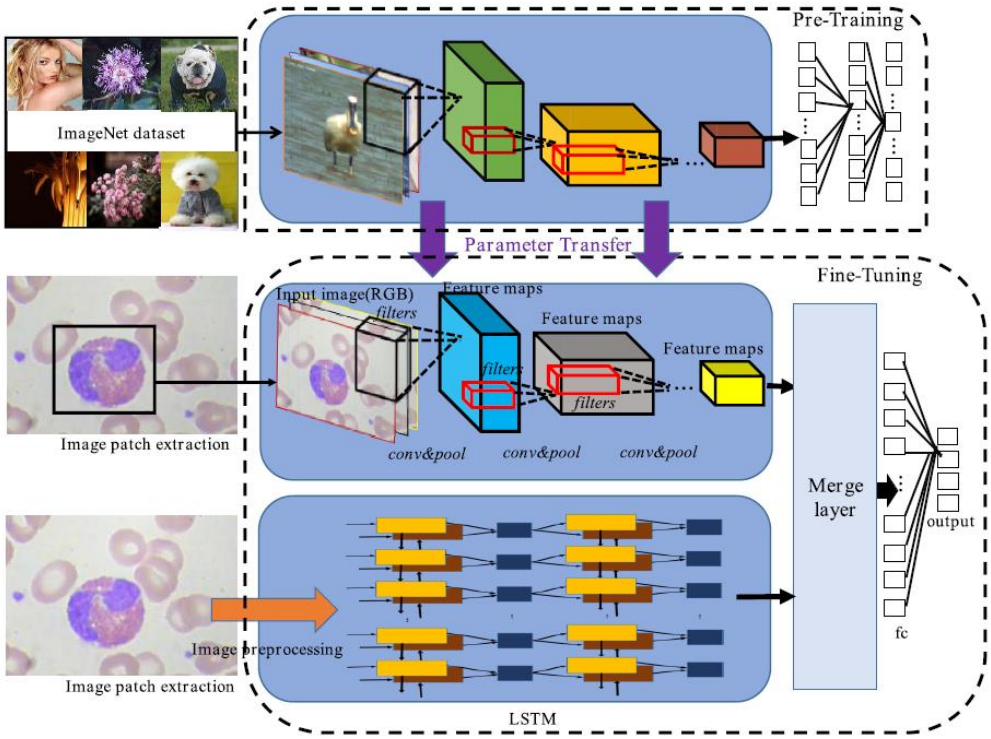
استاندارد Inception با ماژول‌های کانولوشنی قابل تفکیک عمقی جایگزین می‌شوند.

روش کار مدل پیشنهادی به این شکل است که از یک مدل آموزش‌دیده Xception بر روی مجموعه‌داده‌های ImageNet استفاده می‌کند تا ویژگی‌های تصاویر پزشکی را استخراج کند. LSTM را به صورت جداگانه بر روی تصاویر پزشکی خام آموزش می‌دهیم. سپس وزن‌های بدست آمده از Xception و LSTM را با یکدیگر ادغام کرده و برای دسته‌بندی استفاده می‌کنیم. شکل شماره ۲-۵ و ۲-۶ روند کار این مدل را نمایش می‌دهند.



شکل ۲-۵ ساختار مدل استفاده شده در مرجع [۲۳]

در شکل شماره ۲-۵ می‌توان ساختار کلی مدل پیشنهادی را دید که در آن Xception و LSTM جداگانه بر روی مجموعه‌داده‌ها آموزش می‌بینند.



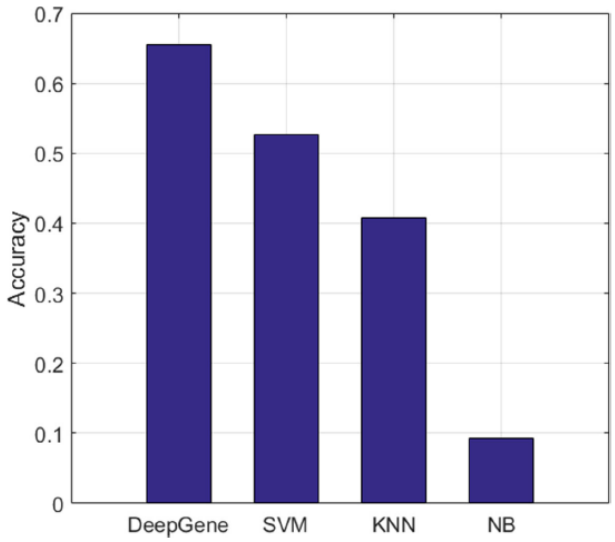
شکل ۲-۶ ساختار مدل پیشنهادی در مرجع[۲۳] برای ترکیب Xception و LSTM که در آن Xception از پیش آموزش دیده است

مدل پیشنهادی مرجع [۲۳] توانسته است به دقت ۹۰,۷۹ درصد برای دسته بندی سلول‌های خون در تصاویر دست پیدا کند. همچنین جدول ۲-۱ نشان می‌دهد که مدل پیشنهادی عملکرد بهتری نسبت به مدل‌های آزمایش شده دیگر دارد.

جدول ۲-۱ نتایج دسته بندی در چند مدل مختلف و مدل پیشنهادی توسط مرجع [۲۳]

Model	Sub-Models	Four-classification accuracy
CNN	Inception V۳	۸۴/۰۸
	ResNet۵۰	۸۶/۶۲
	Xception	۸۸/۷۰
CNN-RNN	Inception V۳-LSTM	۸۷/۴۵
	ResNet۵۰-LSTM	۸۹/۳۸
	Xception-LSTM	۹۰/۷۹
	Xception-ResNet۵۰-LSTM	۸۸/۵۸

مرجع [۵] یک روش جدید برای SMCC پیشنهاد داده است، تا با استفاده از آن بتوان انواع سرطان را دسته بندی کرد. در این روش از دو پیش پردازش منحصر به فرد به نام‌های فیلتر کردن ژن‌های خوشه بندی شده (CGF)[۱] و کاهش پراکندگی فهرست شده (ISR)[۲] استفاده شده است. عملکرد روش پیشنهادی به این شکل است که ابتدا ژن‌های که بیشترین جهش‌ها را در نمونه‌ها ما داشته‌اند با استفاده از پیش پردازش CGF تفکیک کرده و گروه بندی می‌کنیم. سپس گروه‌هایی که تعداد داده‌های درون آنها کمتر از حد آستانه مشخص شده باشند حذف می‌شوند. حد آستانه در مرحله گفته شده پنج می‌باشد. سپس تمامی داده‌های گروه‌های باقی مانده را به صورت یک ماتریس کنار هم قرار می‌دهیم. در مرحله بعد داده‌های خام به ISR داده می‌شوند، در این مرحله شماره مکان هر یک از ژن‌ها در هر نمونه که مقدار جهش بیش از صفر دارد به دست می‌آید، اگر تعداد این شماره‌ها از حد آستانه (هشتصد) مشخص شده کمتر باشد به انتهای آنها صفر اضافه می‌شود و اگر بیشتر باشد به مقدار حد آستانه داده انتخاب می‌شود. در آخر داده‌ها بدست آمده از ISR به انتهای داده‌ها بدست آمده از CGF اضافه می‌شود. مدل پیشنهادی یک شبکه تمام



شکل ۲-۷ مقایسه روش پیشنهادی در مرجع [۵] و سه روش یادگیری ماشین.

[۱] Clustered Gene Filtering

[۲] Indexed Sparsity Reduction

متصل با تعداد چهار لایه مخفی است. این مدل توانسته به مقدار دقت ۶۵,۵ درصد در تشخیص ۱۲ نوع سرطان دست پیدا کند. شکل ۲-۷ تفاوت دقت روش پیشنهادی با سه روش دیگر را نمایش می‌دهد.

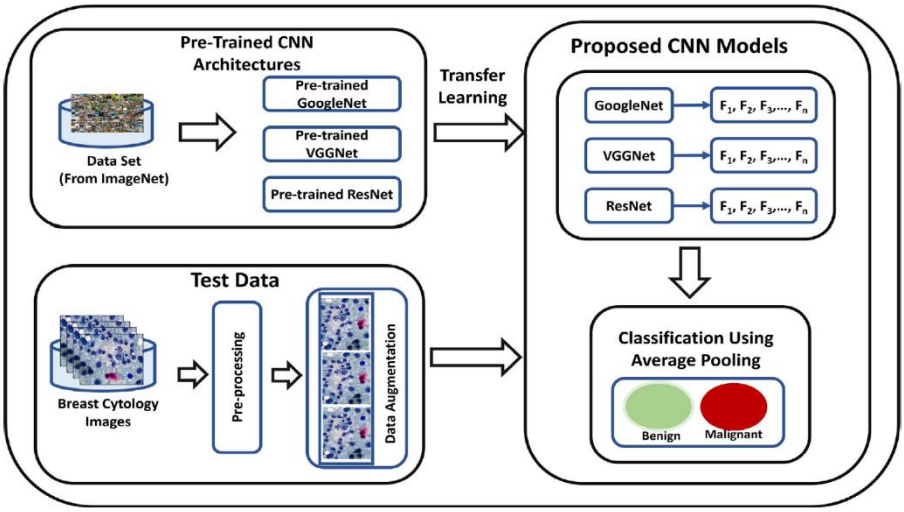
سرطان پستان یکی از شایع‌ترین دلایل مرگ‌ومیر در بین زنان ۲۰ تا ۵۹ ساله در سرتاسر جهان است. تشخیص و طبقه‌بندی آن در مراحل اولیه رشد می‌تواند به درمان بیماران کمک فراوانی کند. در مرجع [۲۵] با استفاده از معماری CNN و روش یادگیری انتقالی[۱] چارچوب[۲۶, ۲۷] جدیدی برای تشخیص و دسته بندی سرطان پستان ارائه می‌دهد.

به‌طورکلی مدل‌های یادگیری عمیق برای یک هدف خاص ایجاد و آموزش داده می‌شوند به این شکل که مدلی که برای هدف الف آموزش‌دیده است با مدلی که برای هدف ب آموزش‌دیده است از هم جدا هستند. بر خلاف این الگوی کلاسیک که به طور جداگانه توسعه و تولید می‌شود، هدف از یادگیری انتقال استفاده از دانش به‌دست‌آمده در هنگام حل یک مسئله برای حل یک مسئله مرتبط دیگر است.

چهارچوب روش پیشنهادی در این مرجع به این شکل عمل می‌کند که ابتدا ویژگی‌های مختلف سطح پایین به صورت جداگانه توسط سه معماری ResNet[۲۹], VGGNet[۲۸], GoogleNet [۲۲] که از قبل بر روی مجموعه‌داده‌های ImageNet آموزش داده شده‌اند استخراج می‌شوند. سپس ویژگی‌های استخراج شده با هم ادقام شده و به یک شبکه تمام متصل برای دسته بندی داده می‌شود. شکل ۲-۸ این سازو کار را نمایش می‌دهد.

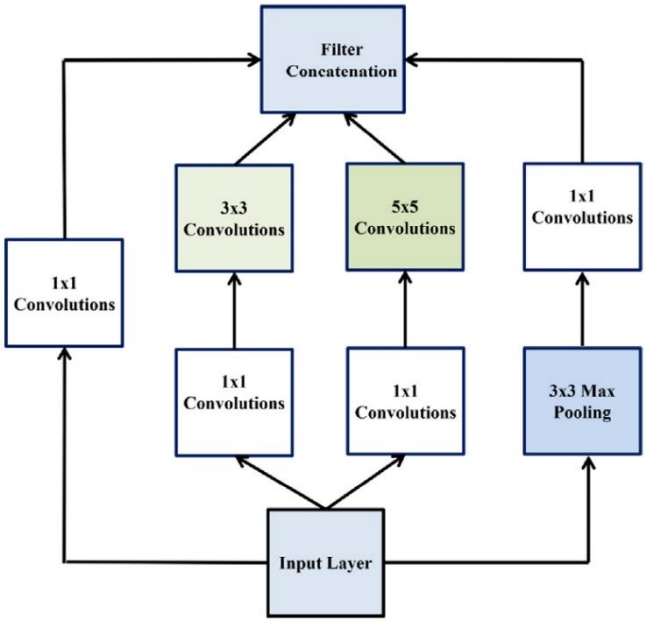
GoogleNet یک مدل کوچک است که دارای سه‌لایه کانولوشن، لایه‌های عملیاتی خطی اصلاح شده، لایه‌های جمع‌کننده و دولایه کاملاً متصل است. شکل ۲-۹ این معماری را نمایش می‌دهد. VGGNet مشابه AlexNet است که شامل ۱۳ لایه کانولوشنی و سه‌لایه تمام متصل می‌باشد، این شبکه از فیلترهایی به‌اندازه سه در سه و pooling با اندازه دو در دو استفاده می‌کند. شکل ۲-۱۰ این معماری را نمایش می‌دهد. ResNet شامل چندین ماژول رسوبی است که بر روی‌هم سوار شده‌اند، ماژول رسوبی دوراه دارد یا می‌تواند

[۱] Transfer Learning

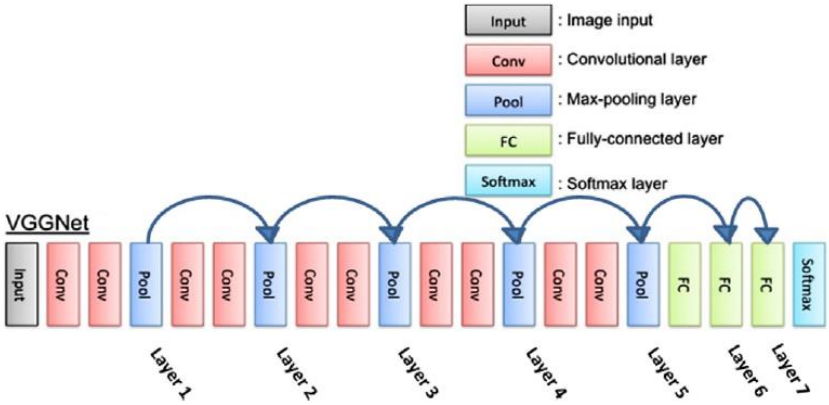


شکل ۲-۹ ساختار مدل در مرجع[۲۵] .

یک سری عملیات بر روی ورودی انجام دهد، یا تمام این مراحل را رد کند. شکل ۲-۱۱ این معماری را نمایش می‌دهد.



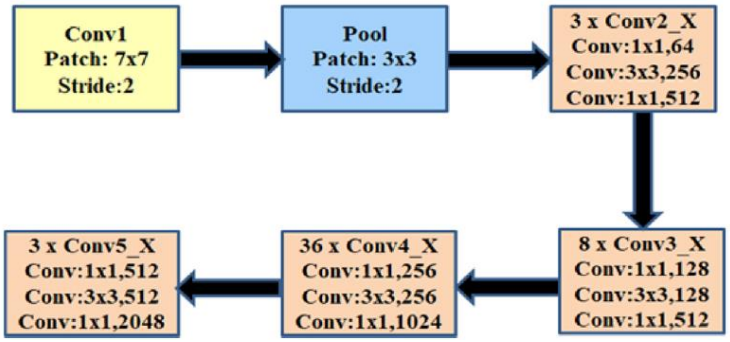
شکل ۲-۸ معماری پایه GoogleNet مرجع[۲۵]



شکل ۲-۱۰ معماری پایه VGGNet

برای افزایش تعداد داده‌ها در مجموعه‌داده از تکنیک تقویت داده‌ها (افزایش داده‌ها) [۱] [۳۰]استفاده شده و ۸۰۰۰ تصویر ایجاد شده است. میزان دقت مدل پیشنهادی برای دسته بندی تصاویر به دو دسته خوش خیم و بد خیم ۹۷,۵۲ درصد بوده است.

سرطان ریه شایع‌ترین علت مرگ ناشی از سرطان در مردان است، با تشخیص زود هنگام می‌توان میزان مرگ و میر ناشی از سرطان ریه را کاهش داد. در مرجع[۳۱] یک سیستم کاملا خودکار تشخیص سرطان

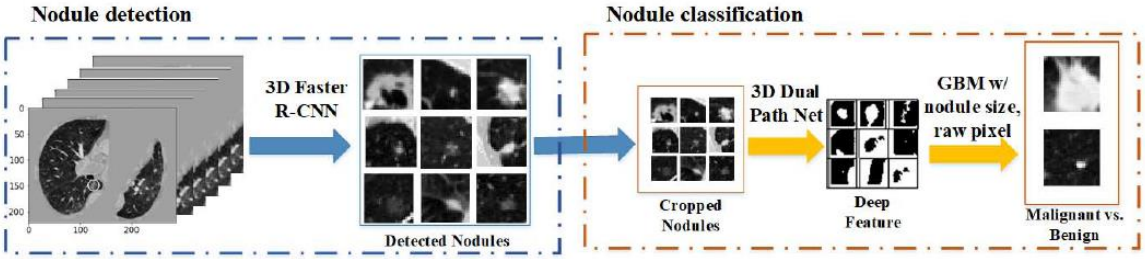


شکل ۲-۱۱ معماری پایه ResNet مرجع[۳۱].

در توموگرافی ریه (CT) [۲] رائه شده است. این سیستم دارای دو بخش تشخیص گره و طبقه بندی گره‌ها می‌باشد. روش کار سیستم به این شکل است که ابتدا تصاویر سه بعدی را به یکThree D Faster R-CNN داده می‌شود [۳۲, ۳۳]تا گره‌های کاندید را تولید کند. سپس از دو شبکه عمیق دو مسیره (DPN) [۳] برای

[۱] Data Augmentation
[۲] Computed Tomography
[۳] Dual-Path Network

استخراج ویژگی‌های عمیق در گره‌ها استفاده می‌شود، در آخر برای طبقه بندی گره از ماشین تقویت کننده شیب (GBM) [۱] استفاده خواهد شد[۳۴] . برای بهره برداری کامل از تصاویر سه بعدی CT به ترتیب دو شبکه کانولوشی سه بعدی برای تشخیص و طبقه بندی گره‌ها طراحی شده است، اما از آنجا که شبکه کانولوشنی سه بعدی دارای پارامترهای زیادی است و آموزش آن روی مجموعه‌داده‌های CT ریه که عموما کوچک است دشوار می‌نماید، شبکه‌های دو مسیره سه بعدی به کار گرفته شده است. این عمل با الهام از Faster R-CNN برای تشخیص اشیاء انجام شده است. در این مرجع شبکه Three D Faster R-CNN را برای تشخیص گره بر اساس شبکه‌های دو مسیر سه بعدی به کار برده‌اند و ساختار رمزگشای رمزگذار U-net و شبکه دو مسیره عمیق سه بعدی را برای طبقه بندی گره‌ها پیشنهاد داده‌اند. شکل ۱۲-۲ ساختار کلی سیستم را نمایش می‌دهد.



شکل ۲-۱۲ ساختار دو قسمتی پیشنهاد شده در مرجع [۳۱]

برای افزایش داده‌های مجموعه‌داده از روش‌های Data Augmentation استفاده شده است. مدل ارائه شده در مقابل تشخیص سه متخصص تست شد و میانگین دقت مدل ۹۲,۷۴ درصد و میانگین دقت متخصصین ۹۱,۲۵ درصد بوده است جدول ۲-۲ این مقایسه را نمایش می‌دهد.

جدول ۲-۲ مقایسه دقت پزشکان و مدل Deep Lung در مرجع [۳۱]

	DR ۱	DR ۲	DR ۳	DR ٤	Average
Doctors	۹۳,۴۴	۹۳,۶۹	۹۱,۸۲	۸۶,۰۳	۹۱,۲۵
DeepLung	۹۳,۵۵	۹۳,۳۰	۹۳,۱۹	۹۰,۸۹	۹۲,۷۴

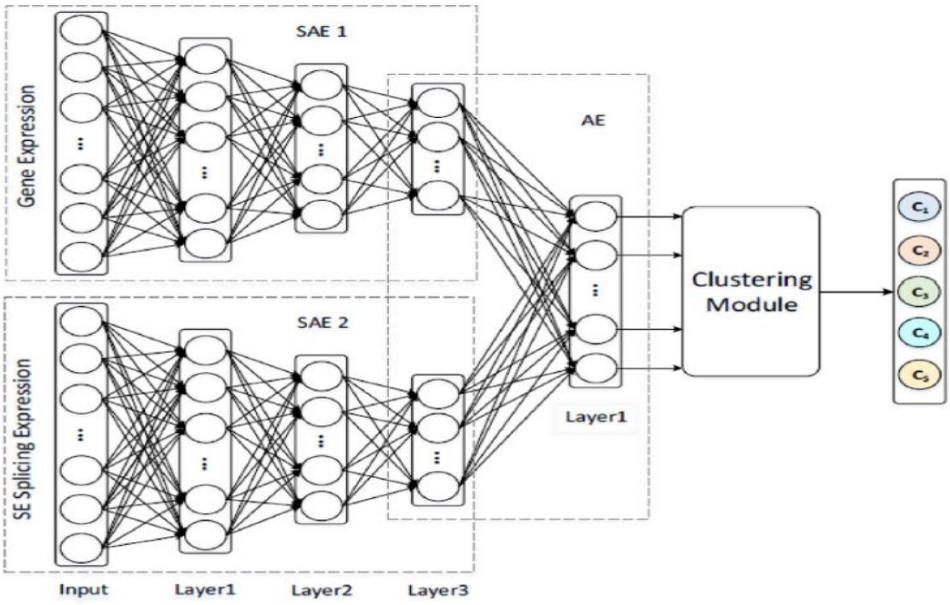
[۱] Gradient Boosting Machine

توموری که بر اثر یک سرطان شکل می‌گیرد، با پاتوژن‌ها[1] و ویژگی‌های بالینی متنوع همراه است. همین امر باعث شده است سرطان‌های مختلف روش‌های درمانی متفاوتی داشته باشند. در طی دهه‌های گذشته طبقه‌بندی زیرگروه‌های سرطان و بررسی تومورهای آنها پاتوژن‌های زیرگروه‌های خاص در تحولات سرطانی را آشکار کرده است و ثابت شده است که تومورها یا انواع مختلف سرطان غالباً در اثر جهش‌های ژنتیکی مختلف ایجاد می‌شوند. با پیشرفت تکنولوژی توالی‌یابی، انواع مختلفی از داده‌های ژنتیکی تولید شده است، بر اساس این داده‌ها روش‌های مختلفی برای شناسایی زیرگروه‌های سرطان ایجاد شده است. مرجع [۳۵] با استفاده از شبکه رمزگذار انباشته (SAE)[2] [۳۶, ۳۷] یک چهارچوب یادگیری عمیق سلسله مراتبی را برای شناسایی زیرگروه‌های سرطان با استفاده از یادگیری ویژگی‌های مهم ارائه می‌دهد.

شبکه به این صورت کار می‌کند که ابتدا نمایش مربوط به هر نوع داده را به ترتیب با استفاده از SAE در سطح پایین یاد می‌گیرد، سپس نمایش‌های آموخته شده هر نوع داده را با هم تلفیق می‌کند تا نمایش‌های یکپارچه نهایی را در سطح بالا از طریق لایه دیگری از AE یاد بگیرد، سپس از روش k-means برای خوشه‌بندی بیماران استفاده می‌شود. شکل ۲-۱۳ این مدل را نمایش می‌دهد.

[1] Pathogenesis

[2] Stacked Auto Encoder Neural Network



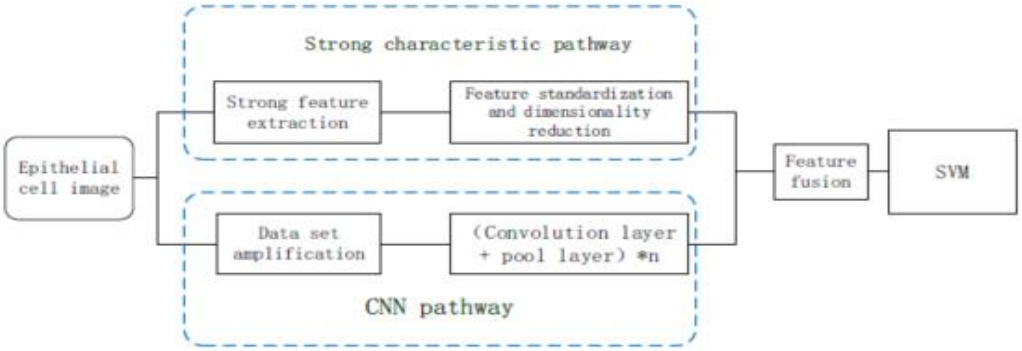
شکل ۲-۱۳ ساختار پیشنهادی مدل SAE در مرجع [۳۵]

سرطان دهانه رحم دومین علت اصلی مرگ‌ومیر در میان زنان در سراسر دنیا می‌باشد. معمولاً می‌توان این نوع سرطان را قبل از بدتر شدن با استفاده از آزمایش پاپانیکولائو (Pap) [۱] تشخیص داد و درمان کرد، به همین دلیل غربالگری منظم و درمان زودهنگام به شکل مؤثری از مرگ‌ومیر ناشی از این بیماری می‌کاهد. در آزمایش سنتی سیتولوژی[۲] سنتی مانند آزمایش پاپ‌اسمیر نیاز است تا معاینه در سطح سلول و با استفاده از میکروسکوپ توسط آسیب‌شناس انجام شود که این کار باعث خطا در آزمون می‌شود.

در مرجع [۳۸] یک رویکرد جدید برای طبقه بندی سلول‌های دهانه رحم ارائه شده است. در این رویکرد یک روش استخراج ویژگی جدید به نام GLCM[۳]+Gabor برای استخراج ویژگی‌های قوی ایجاد شده است. همچنین از LetNet-۵ به عنوان ماژول اسنخراج ویژگی CNN برای بدست آوردن سایر ویژگی‌های انتزاعی استفاده شده است. در آخر ویژگی‌های بدست آمده برای طبقه بندی به یک طبقه بندی کننده SVM وارد

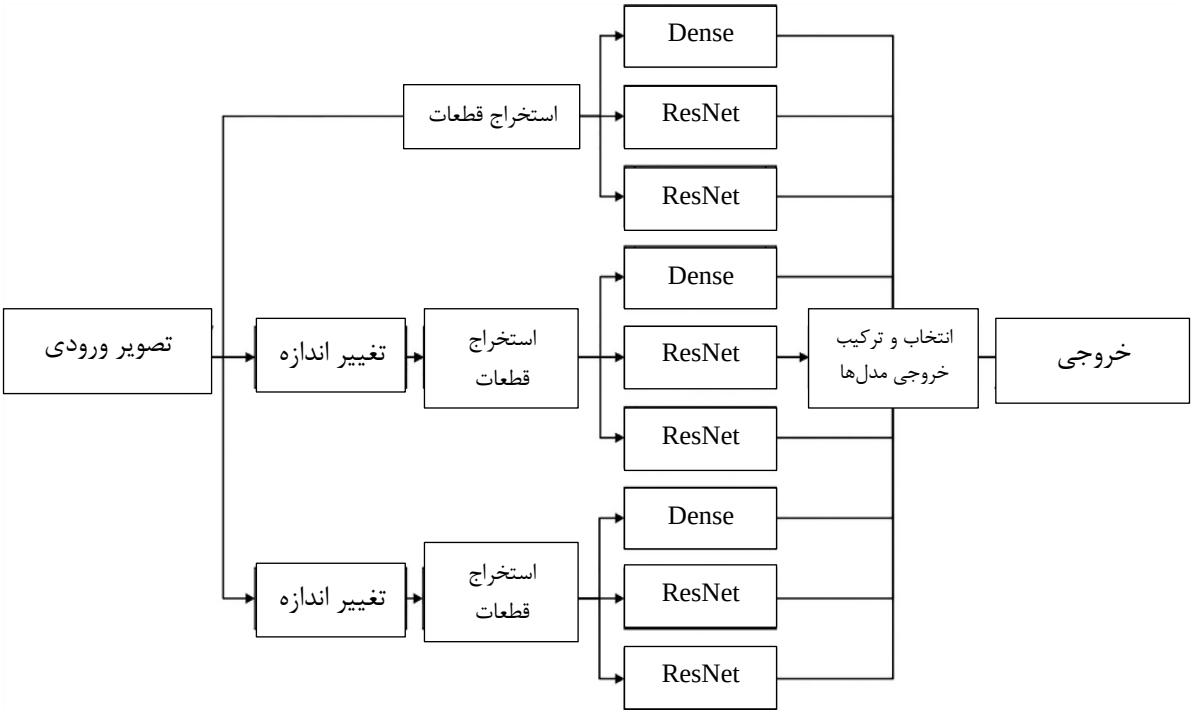
[۱] Papanicolaou
[۲] Cytology Test
[۳] Gray-Level Co-occurrence Matrix

می‌شود. شکل ۲-۱۴ ساختار ارائه شده را نمایش می‌دهد.



شکل ۲-۱۴ ساختار ارائه شده در مرجع [۳۸]

تصاویر خام به صورت جداگانه وارد LetNet-۵ و دو پیش‌پردازش GLCM و Gabor می‌شوند. GLCM یک روش تجزیه‌وتحلیل آماری است که می‌تواند اطلاعاتی جامع مانند: جهت، فاصله مجاور و دامنه تغییر را در باره سطح خاکستری تصویر در اختیار کاربر قرار دهد. Gabor برای تعیین فرکانس سینوسی و محتوای بخش‌های محلی سیگنال با تغییر در طول زمان استفاده می‌شود. بعد از استخراج ویژگی‌ها آنها را با یکدیگر ادغام می‌کنند و به SVM می‌دهند. این روش توانسته به‌دقت ۹۹,۳ درصد دست یابد.



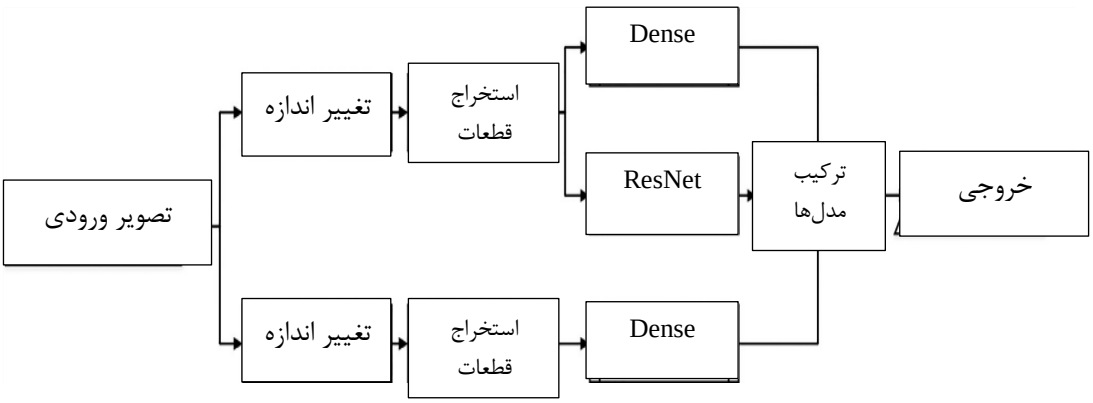
شکل ۲-۱۵ ساختار کامل مدل ارائه شده در مرجع [۳۸].

برای تشخیص سرطان پستان آزمایش‌های تشخیصی مانند: معاینه فیزیکی، ماموگرافی، تصویربرداری با تشدید مغناطیسی (MRI) [۱]، سونوگرافی و نمونه‌برداری به کار می‌رود که در میان آنها تجزیه‌وتحلیل تصویر بافت ناشی از نمونه‌برداری با سوزن نقش بسیار کلیدی ایفا می‌کند. در طی این روش تشخیص مناطق سلولی تصاویر میکروسکوپی هیستوپاتولوژیک[۲] رنگ‌آمیزی شده با هماتوکسیلین - ائوزین (H&E) [۳] توسط متخصصان ارزیابی می‌شود، تا نوع بافت پستان از جمله بافت طبیعی، ضایعه خوش‌خیم، سرطان درجا[۴]، سرطان تهاجمی[۵] مشخص شود. بااین‌وجود، به دلیل پیچیدگی زیاد تصاویر میکروسکوپی هیستوپاتولوژیک، تشخیص کارسینوم توسط بیوپسی به مهارت و تمرکز بالایی نیاز دارد. این کار زمان‌بر بوده و مستعد سوگیری متصدی می‌باشد. در مرجع [۳۹] شبکه‌های عصبی کانولوشنی چند مقیاسی

[۱] Magnetic Resonance Imaging
[۲] Histopathological
[۳] Hematoxyling-Eosin
[۴] Situ Carcinoma
[۵] Invasive Carcinoma

(EMS-Net) [۱] برای طبقه بندی خودکار تصاویر میکروسکوپی هیستوپاتولوژیک پستان آغشته به H&E پیشنهاد شده است.

روش کار به این شکل است که ابتدا هر تصویر میکروسکوپی هیستوپاتولوژیک را به مقیاس‌های متعددی تبدیل می‌کنند، سپس از تکه‌های آموزشی برش خورده و تقویت شده در هر مقیاس برای تنظیم دقیق سه



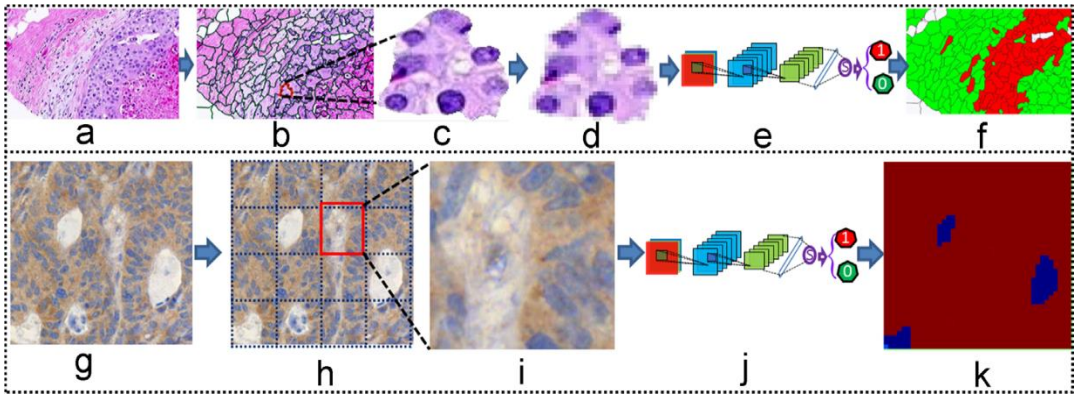
شکل ۲-۱۶ مدل نهایی ارائه شده در مرجع [۳۸].

شبکه عمیق کانولوشنی، ResNet-۱۰۱, DenseNet-۱٦۱, ResNet-۱۵۲ که از قبل آموزش‌دیده‌اند، استفاده می‌شود. در این حالت نه مدل آموزش‌دیده وجود خواهد داست که از بین آنها یک زیرمجموعه از مدل‌ها که کارایی بهتری دارند باید انتخاب شوند. برای انتخاب این زیرمجموعه تعدادی از تصاویر را به‌عنوان داده‌های اعتبارسنجی مورداستفاده قرار می‌دهند تا بهترین مدل‌ها انتخاب و با هم ترکیب شوند. شکل ۲-۱۵ الگوریتم مدل پیشنهادی و شکل ۲-۱۶ مدل نهایی را نشان می‌دهد. مدل نهایی به‌دقت ۹۱,۷۵ درصد دست‌یافته است.

اپیتلیال (EP) [۲] و استروما (ST) [۳] نام دو نوع بافت در تصاویر بافت‌شناسی[۴] هستند. ST شامل بافت‌های

[۱] Ensemble Of Multi Scale Convolutional Neural Networks
[۲] Epithelial
[۳] Stromal
[۴] Histological Images

چربی و الیافی[۱] است که مجاری و لو بول‌ها[۲]، رگ‌های خونی و عروق لنفاوی را احاطه کرده است. این بافت چارچوب‌های حمایت‌کننده یک اندام هستند. EP بافتی سلولی است، و در سیستم مجاری لوبولار[۳] مجاری شیر مادر یافت می‌شود. حدود ۸۰ درصد تومورهای پستان از سلول‌های EP پستان نشئت می‌گیرند. اگرچه به طور معمول بافت ST بخشی از بافت بدخیم محسوب نمی‌شود، اما تغییرات در استروما[۴] منجر به حمله و متاستاز[۵] تومور می‌شود؛ بنابراین نسبت تومور به استروما در بافت‌شناسی به‌عنوان یک مقیاس مهم برای پیش‌آگاهی شناخته می‌شود. زیرا رشد و پیشرفت سرطان به محیط ریز بافت‌های EP و ST وابسته است. تشخیص استروما از محفظه بافت اپیتلیال در تصاویر بافت‌شناسی دیجیتال به دلیل تراکم بالای داده، پیچیدگی ساختارهای بافتی و ناهماهنگی در آماده‌سازی بافت بسیار چالش‌برانگیز است؛ بنابراین، ایجاد الگوریتم‌های هوشمند برای تقسیم‌بندی ساختارهای مختلف بافت به روشی دقیق و سریع بسیار مهم است. مرجع [۴۰]یک مدل عمیق کانولوشنی را برای تشخیص محفظه‌های اپیتلیال و استروما در تصاویر سرطان پستان ارائه می‌دهد. هر تصویر ابتدا به صورت زیر تصویرهایی برش می‌خورد. برای ساخت این تصاویر از سوپرپیکسل (SP) [۶] و یک پنجره مربع شکل با اندازه ثابت استفاده می‌شود. سپس این تصاویر به عنوان



شکل ۲-۱۷ رویکرد روش ارائه شده در مرجع [۴۰]

۱ Fibrous
۲ Lobules
۳ Lobular
۴ Stroma
۵ Metastasis
۶ Super Pixel

ورودی به مدل داده می‌شوند. در این مرجع از دو مجموعه‌داده H&E [۱] و IHC [۲] استفاده شده است. شکل ۲-۱۷ رویکرد مدل ارائه شده بر روی این دو مجموعه‌داده را نمایش می‌دهد.

۲-۵ جمع‌بندی

در این فصل ابتدا به بیان مفهوم یادگیری عمیق و شرح کارکرد دو مدل استاندارد CNN و LSTM پرداختیم. سپس با بررسی انواع کارهای انجام شده به‌وسیله یادگیری عمیق برای تشخیص سرطان دریافتیم که ساختارهای CNN و ترکیبی از CNN با سایر روش‌ها در کارهایی مانند تقسیم‌بندی[۴۱] یا طبقه بندی تصاویر پزشکی[۴۲] مانند تشخیص سرطان پستان[۲۱ ,۴۳] یا تشخیص بافت اپیتلیال[۳] در سرطان پروستات با موفقیت استفاده شده است [۴۴]. از این رو در ادامه ما مدل‌های استاندارد و ساده CNN، LSTM و ترکیب CNN و LSMT را برای کار خود مورد بررسی قرار خواهیم داد.

جدول ۲-۱ روش‌های شرح داده شده در مقالات را به طور خلاصه نمایش می‌دهد.

جدول ۲-۳ شرح خلاصه‌ای از فعالیت‌های کارهای پیشین

مرجع	روش پیشنهادی در هر مرجع
[۲۱]	با استفاده از شبکه Inception موفق به دسته‌بندی دو نوع سرطان ریه به نام‌های LUAD وLUSC با دقت ٪۹۷ شده‌اند
[۲۳]	با ترکیب مدل‌های Xception و Inception موفق به دسته‌بندی سلول‌های خون با دقت ٪۹۰,۷۹ شدند.
[۵]	در این مرجع دو پیش‌پردازش منحصر به فرد ایجاد شده است و با ساخت یک مدل تمام متصل عمیق موفق به دسته‌بندی ۱۲ نوع سرطان مبتنی بر جهش‌های سوماتیک با دقت ٪۶۵,۵ شدند.
[۲۵]	در این مرجع با ترکیب مدل‌های از پیش آموزش دیده ResNet, VGG و GoogleNet موفق به تشخیص تومورهای خوش‌خیم و بدخیم با دقت ٪۹۷,۵۲ شدند.
[۳۱]	در این مقاله یک سیستم خودکار تشخیص سرطان بر روی تصاویر سیتی اسکن ایجاد شده است که میان

[۱] Hematoxylin and Eosin

[۲] Immunohistological

[۳] Epithelial

	گین دقت آن ۹۲,۷۴٪ است.
[۳۵]	با استفاده از مدل SAE موفق به شناسایی زیرگروه‌های سرطان با دقت ۹۰٪ شده‌اند.
[۳۸]	با استفاده از دو پیش‌پردازش به نام‌های GLCM و Gabor و یک شبکه LetNet و SVM موفق به طبقه‌بندی سلول‌های دهانه رحم شدند.
[۳۹]	برای تشخیص سرطان پستان با استفاده از تصاویر میکروسکوپی هیستوپاتولوژیک پستان و مدل EMS-Net موفق به طبقه‌بندی خودکار این تصاویر با دقت ۹۱,۷۵٪ شده‌اند.
[۴۰]	از یک کدل CNN برای تشخیص اپتلیال و استروما در تصاویر پزشکیس استفاده شده است. این مدل دقت ٪۹۰ را کسب کرده است.

فصل ۳ : روش پیشنهادی

۳-۱ مقدمه

در فصل اول مسئله را تعریف کردیم و در ادامه تعدادی از فعالیت‌های پیشین را مورد بررسی قرار دادیم. با بررسی کارهای پیشین دریافتیم که مدل‌های CNN, LSTM و ترکیب این دو مدل یا یکدیگر بیشترین استفاده را در تشخیص سرطان و تومورهای سرطانی داشته است. در این پایان‌نامه سه مدل LSTM، CNN و ترکیب این دو مدل پیاده سازی و تنظیم می‌شود. سپس، نتایج بدست آمده از این سه مدل با یکدیگر مقایسه می‌شود. بر اساس این نتایج بدست آمده مدل CNN بهترین عملکرد را در مقایسه با دو مدل دیگر دارد. بنابر این در این پایان نامه مدل CNN به عنوان روش پیشنهادی معرفی می‌شود. روش پیشنهادی ما شامل چهار مرحله اصلی است. این چهار مرحله، شامل کسب داده، پیش‌پردازش، استخراج ویژگی و طبقه‌بندی می‌باشد. در این فصل، ابتدا به شرح روش پیشنهادی برای طبقه‌بندی انواع سرطان بر اساس جهش‌های سوماتیک پرداخته شده است. در این پایان نامه تمرکز ما بر روی ساخت و تنظیم یک مدل یادگیری عمیق و پیدا کردن ارتباط جهش‌های ژن‌ها با یکدیگر است. در ادامه، مراحل پیش‌پردازش داده‌های خام شرح داده می‌شود. داده‌ها خام نشان دهنده تعداد جهش‌های هر ژن در هر نمونه می‌باشد و داده‌های پیش‌پرداز شده نشان دهنده وقوع یا عدم وقوع جهش ژن‌ها در هر نمونه می‌باشد. سپس، به ترتیب به شرح لایه جاسازی[۱]، لایه کانولوشنی[۲]، لایه پولینگ حداکثر سراسری[۳]، لایه حذف تصادفی[۴] و لایه تمام متصل[۵] برای کار دسته‌بندی و تاثیرات آن‌ها بر روی مدل پرداخته شده.

[۱] Embedding Layer
[۲] Convolutional Layer
[۳] Global Max Pooling Layer
[۴] Dropout Layer
[۵] Fully Connected Layer

۳-۲ روش پیشنهادی

روش پیشنهادی در این پایان نامه یک مدل CNN تک لایه به همراه دو پیش پردازش است. این دو پیش پردازش CGF[۱] و ISR[۲] نام دارند که که هدف این دو پیش‌پردازش ساخت یک زیرمجموعه از ژن‌های موثر در دسته‌بندی و کاهش پراکندگی در داده‌ها می‌باشد. این دو پیش‌پردازش با استفاده از زبان برنامه نویسی پایتون پیاده‌سازی شده و هر یک به صورت جداگانه بر روی داده‌های هام اعمال می‌شوند. سپس، نتیجه ISR به انتهای CGF الحاق می‌شود. بعد از آن داده‌های پیش‌پردازش شده وارد لایه جاسازی می‌شوند. این لایه تاثیر داده‌های صفر را کاهش داده و ارتباط میان ژن‌هایی را یافته و آنهایی که ارتباط معنا دارتری داشته باشند را در یک خوشه قرار می‌دهد. سپس خروجی لایه جاسازی وارد یک لایه کانولوشنی می‌شود. این لایه به ما در پیداکردن ویژگی‌های سطح بالاتر کمک می‌کند. برای رفع اثر بیش‌برازش در مدل پیشنهادی از روش‌های نظم‌دهی و لایه پولینگ حداکثر سراسری استفاده می‌شود. در انتها، خروجی لایه کانولوشنی برای دسته‌بندی به یک لایه تمام متصل فرستاده می‌شود.

۳-۳ پیش‌پردازش

داده‌های خام مورد استفاده در این پایان نامه اعداد مثبت صحیح هستند. این اعداد نشان دهنده تعداد جهش‌های هر ژن در هر نمونه است. در این بخش، ابتدا داده‌های خام به نوع دودویی تبدیل می‌شوند. در این نوع داده‌ها اطلاعات به صورت صفر و یک نشان داده می‌شوند. صفر به معنای عدم جهش در ژن و یک به معنای داشتن حداقل یک جهش در ژن می‌باشد. زمانی که داده‌های خام به داده‌های دودویی تبدیل می‌شوند یک ماتریس به نام A در ابعاد $A \in \{۰،۱\}^{m \times n}$ را تشکیل می‌دهند. m تعداد ردیف‌ها و نشان دهنده ژن‌ها و n تعداد سطون‌ها و نشان دهنده نمونه‌ها می‌باشند. همچنین به دلیل پراکندگی داده‌ها و

[۱] Clustered Gene Filtering

[۲] Indexed Sparsity Reduction

وجود ژن‌های بی‌تاثیر در کار دسته‌بندی، بعد از تبدیل داده‌ها خام به نوع دودویی، دو مرحله پیش پردازش به نام‌های ISR و CGF بر روی داده‌های خام به صورت جداگانه اعمال می‌شود [۵].

۳-۳-۱ پیش پردازش CGF

CGF تلاش دارد تا ژن‌هایی با تأثیر بیشتر در کار دسته‌بندی را در سه مرحله شناسایی کند. در مرحله اول باید تعداد کل جهش‌های هر ژن در تمامی نمونه‌ها محاسبه شود. برای این منظور، در ماتریس A برای هر ژن تمامی جهش‌های رخ داده شده در تمامی نمونه‌ها با یکدیگر جمع می‌شود. این داده‌ها به صورت نزولی مرتب می‌شوند سپس شاخص[۱] هر ژن در یک وکتور به نام A^* قرار می‌گیرد. به این ترتیب، وکتور ایجاد شده شامل شماره جایگاه هر ژن در ماتریس A می‌باشد و ژن‌هایی با بیشترین جهش، در شاخص‌های ابتدایی وکتور قرار می‌گیرند. در دومین مرحله ژن‌ها بر اساس میزان تشابه با یکدیگر گروه‌بندی می‌شوند. برای تعیین میزان تشابه دو ژن از ضریب جاکارد[۲] استفاده شده است. نحوه محاسبه ضریب جاکارد برای دو ژن به نام‌های p و q با فرمول ۳-۱ محاسبه می‌شود.

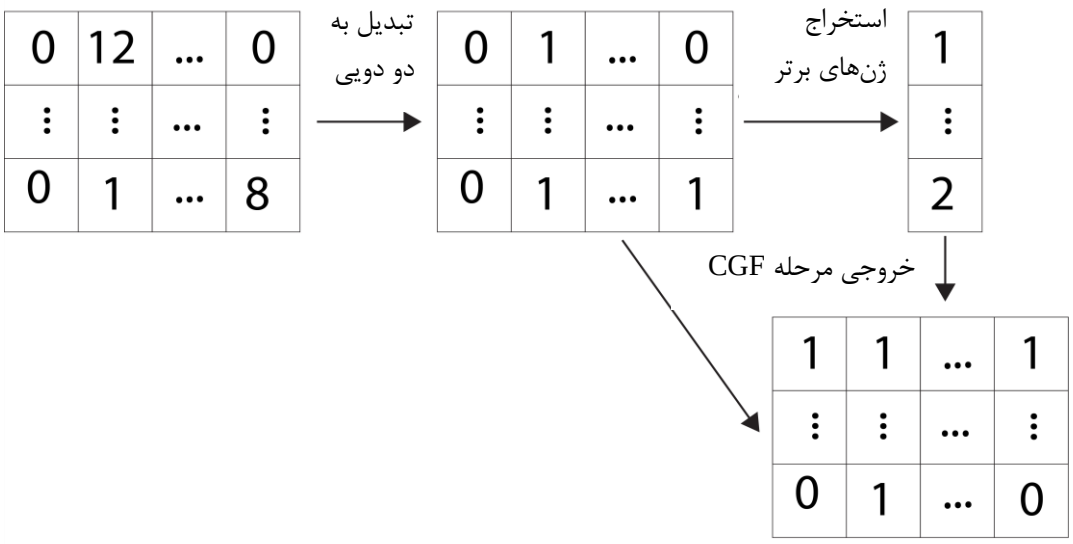
$$d(p، q) = \frac{sum\ (p\&q)}{sum\ (p|q)} \qquad (۱-۳)$$

p و q ژن‌هایی با ابعاد $n \times ۱$ هستند و علامت های | و & به ترتیب نشان دهنده (یا) منطقی[۳] و (و) منطقی[۴] می‌باشند. مراحل ساخت گروه‌هایی از ژن‌های مشابه به این ترتیب است. ابتدا براساس ترتیب قرار گیری ژن‌ها در A^* اولین ژن را انتخاب کرده و بعد تمامی اطلاعات مربوط به جهش‌های ژن از ماتریس A استخراج می‌شود. سپس دومین ژن را از A^* انتخاب کرده و اطلاعات آن را از ماتریس A استخراج می‌کنیم. حال بر اساس ضریب جاکارد میزان تشابه این دو ژن محاسبه می‌شود. اگر میزان تشابه دو ژن از ۰/۷ بیشتر باشد ژن دوم در گروه ژن اول قرار گرفته و در مقایسه‌های دیگر از آن استفاده نمی‌شود. به همین ترتیب

---

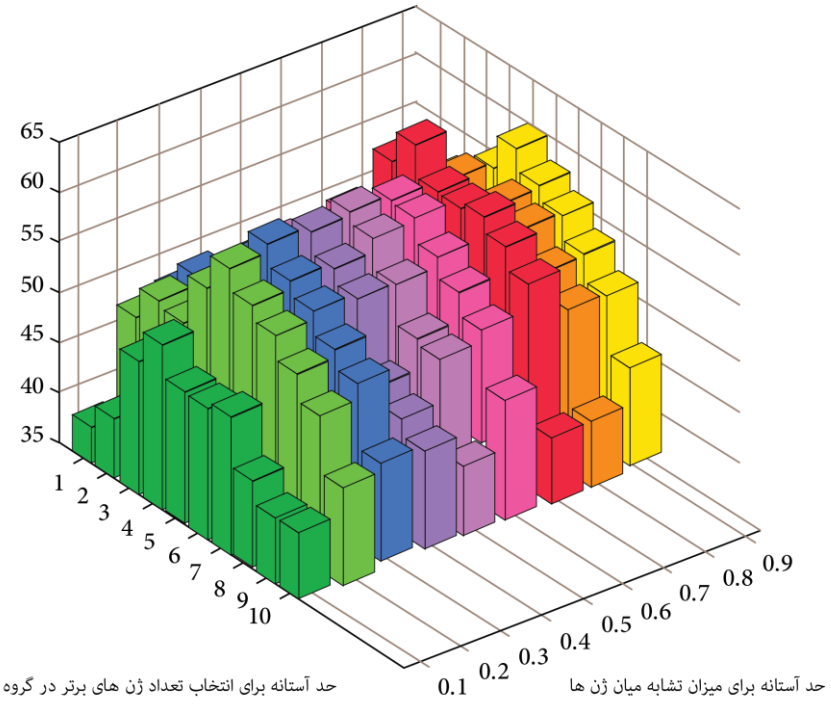
[۱] Index
[۲] Jaccard Coefficient
[۳] Logical OR
[۴] Logical AND

مقایسه را برای ژن‌های اول و سوم، اول و چهارم و ... انجام می‌دهیم. تا زمانی که ژن اول با تمامی ژن‌ها A^* مقایسه شود. به این ترتیب گروهی از ژن‌های مشابه ژن اول تشکیل می‌شود. سپس همین مراحل را برای ژن بعدی در A^* که به گروهی تعلق ندارد انجام می‌شود. تا با سایر ژن‌هایی که عضو گروهی نیستند مقایسه شود و گروه ژن‌های مشابه خود را ایجاد کند. این کار تا زمانی انجام می‌شود که تمامی ژن‌ها در یک گروه قرار بگیرند. بعد از پایان گروه‌بندی، گروه‌هایی که کمتر از ۵ عضو دادند حذف می‌شوند. اگر گروهی بیش از ۵ عضو داشته باشد با استفاده از A^* ، ۵ عضو که بیشترین جهش را در گروه دارند انتخاب شده و مابقی ژن‌ها حذف می‌شوند. سپس تمامی ژن‌های منتخب یک ماتریس مشابه ماتریس A اما با تعداد ژن‌های کمتر را تشکیل می‌دهند. این ماتریس خروجی مرحله CGF است. شکل ۳-۲ مراحل انجام CGF را نمایش می‌دهد [۵].



شکل ۳-۱ مراحل انجام پیش‌پردازش CGF مرجع [۵].

برای تعیین دو حد آستانه ۰/۷ و ۵ تمامی پارامترهای مدل ثابت در نظر گرفته می‌شود. سپس با استفاده از صحت اعتبار سنجی متقابل ده برابر[۱] مقدار حد آستانه برای تشابه بین ژن‌ها و مقدار حد آستانه برای



شکل ۳-۲ دقت بدست آمده برای مقادیر مختلف حدآستانه میزان تشابه میان ژن‌ها و حد آستانه انتخاب ژن‌های برتر در هر گروه مرجع [۵].

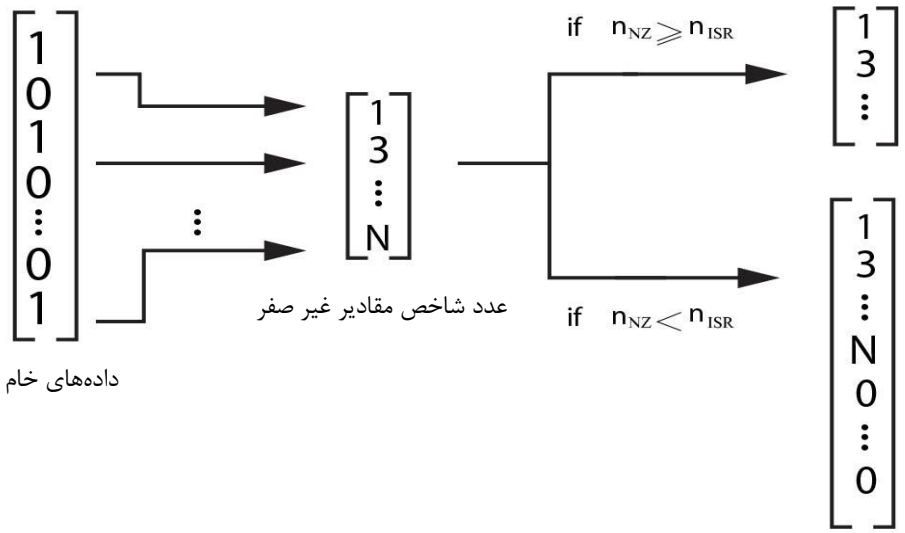
انتخاب ژن‌های برتر در هر گروه به ترتیب در بازه ۰/۱ تا ۰/۹ و ۱ تا ۱۰ مورد بررسی قرار می‌گیرند. شکل ۳-۳ دقت بدست آمده از صحت اعتبار سنجی متقابل ده برابر را برای مقادیر مختلف حدآستانه میزان تشابه بین ژن‌ها و حدآستانه برای انتخاب تعداد ژن‌های برتر در گروه نمایش می‌دهد. محور Z نشان دهنده میزان دقت است. مقادیر ۰/۷ و ۵ بیشترین میزان دقت را بین سایر مقادیر کس کرده‌اند [۵].

۳-۳-۲ پیش‌پردازش ISR

هدف پیش‌پردازش ISR کاهش میزان پراکندگی داده‌ها با جایگذینی عدد شاخص ژن‌های است که جهش داشته‌اند. این پیش‌پردازش مانند CGF بر روی ماتریس A اعمال می‌شود. مراحل اعمال این پیش‌پردازش بر روی ماتریس A به این ترتیب است. برای هر نمونه که دارای ابعاد $n \times 1$ عدد شاخص مقادیر غیر صفر

[۱] Ten Fold Cross-Validation

استخراج شده و به صورت صعودی مرتب می‌شود. اگر تعداد این مقادیر برابر یا بیشتر از ۸۰۰ عدد باشد با استفاده از A^* تعداد ۸۰۰ ژن که بیشترین جهش را در این مجموعه دارند انتخاب شده و در یک وکتور قرار می‌گیرند. اما اگر تعداد این مقادیر کمتر از ۸۰۰ عدد باشد به انتهای مجموعه به میزانی صفر اضافه می‌شود که طول آن به ۸۰۰ برسد. سپس تمامی وکتورهای استخراج شده تشکیل یک ماتریس را می‌دهند که این ماتریس خروجی پیش‌پردازش ISR می‌باشد. شکل ۳-۴ مراحل پیش‌پردازش ISR را نمایش می‌دهد.



شکل ۳-۳ مراحل پیش پردازش ISR مرجع [۵].

در پیش‌پردازش ISR عدد ۸۰۰ به عنوان یک حدآستانه معرفی شده است. با بررسی مقادیر غیر صفر در در هر نمونه ازمجموعه داده مشخص می‌شود که بیش از ۹۷٪ از نمونه‌ها کمتر از ۸۰۰ مقدار غیر صفر دارند. یعنی کمتر از ۸۰۰ ژن در هر نمونه وجود دارد که جهش کرده است. به همین دلیل عدد ۸۰۰ به عنوان حدآستانه در پیش‌پردازش ISR انتخاب شده است. بعد از اعمال پیش‌پردازش‌های CGF و ISR بر روی ماتریس A، ماتریس نتیجه ISR به انتهای ماتریس نتیجه CGF اضافه می‌شود. این ماتریس جدید به عنوان

ورودی برای مدل مورد استفاده قرار گرفته است.

۴-۳ لایه Eembedding

داده‌هایی که قرار است به‌عنوان ورودی به مدل پیشنهادی داده شوند، ترکیبی از صفر، یک و شماره فهرست ژن‌هایی است که در مرحله ISR و CGF بدست آوردیم. با وجود استفاده از دو پیش‌پردازش ISR و CGF هنوز در داده‌های ما مقادیر زیادی سفر وجود دارد، و مجموعه‌داده را پراکنده می‌کند، همچنین در انتهای بعضی از نمونه‌ها برای هماهنگ کردن طول آنها با سایر نمونه‌ها تعدادی صفر اضافه شده است که مدل در کار خود نباید آنها را در نظر بگیرد. برای رفع این مشکلات و یافتن ارتباط و روابط مخفی بین جهش‌ها ما از لایه جاسازی در مدل پیشنهادی خود استفاده کردیم.

در زمینه شبکه‌های عصبی مصنوعی، مخصوصا پردازش متن لغت نامه‌های[۱] انسانی بصورت متن ساده هستند. اما برای اینکه یک مدل یادگیری ماشینی قادر به فهم و پردازش زبان طبیعی باشد، نیازمند تبدیل کلمات از متن ساده به مقادیر عددی هستیم. یکی از ساده‌ترین روش‌های تبدیل کلمات، انجام فرایند کد گذاری وان - هات[۲] است که در آن هر کلمه دارای یک جایگاه خاص در یک بردار است و مقادیر دودویی صفر و یک نشانگر عدم وجود یا وجود یک کلمه در یک بردار می‌باشند [۴۵] . اما، کدگذاری وان – هات در زمانی که با تمام یک لغت نامه مواجه باشیم از نظر محاسباتی سربار بسیار زیادی بر ما تحمیل خواهد کرد و در نتیجه فرایندی غیرعملی خواهد بود چرا که بازنمایی حاصل در اینجا نیازمند صدها هزار بُعد خواهد بود. جاسازی کلمات[۳] یکی از روش‌های نمایش کلمات[۴] است [۴۶]. کلمات و عبارات را در قالب بردارهایی عددی غیر دودویی برخلاف آنچه در کدگذاری وان – هات انجام می‌شود با ابعادی به مراتب کوچکتر و متراکمتر ارائه می‌کند. جاسازی کلمات الگوریتم ما را قادر می‌سازد تا روابط و شباهت‌های بین

---

[۱] Dictionary
[۲] One-Hot Encoding
[۳] Word Embedding
[۴] Word Representation

کلمات را به صورت خودکار متوجه شود. به این صورت که ابتدا برای تشخیص کلمات یک لغت نامه از آن‌ها ایجاد می‌شود و هر کلمه یک عدد صحیح منحصر به فرد دریافت می‌کند. سپس به هر کلمه که اکنون دارای یک عدد منحصر به فرد است یک بردار ویژگی از اعداد اعشاری n بعدی اختصاص داده می‌شود. جدول ۳-۵ جاسازی کلمات برای مرد، زن و درخت با طول بردار ۴ را نمایش می‌دهد [۴۷, ۴۸].

شکل ۳-۴ نحوه تخصیص بردار ویژگی به کلمات موجود در لغت نامه در جاسازی کلمات

مرد (۱۲۰)	زن (۱۰)	درخت (۵۶۳)	در (۸۵)
۰/۰۱	۰/۰۰	۰/۰۴	۰/۰۶
۰/۱۲	۰/۱۴	۰/۲۶	۰/۵
۰/۰۲	۰/۰۵	۰/۰۳	۰/۰۷
۰/۲۲	۰/۱۹	۰/۲	۰/۹

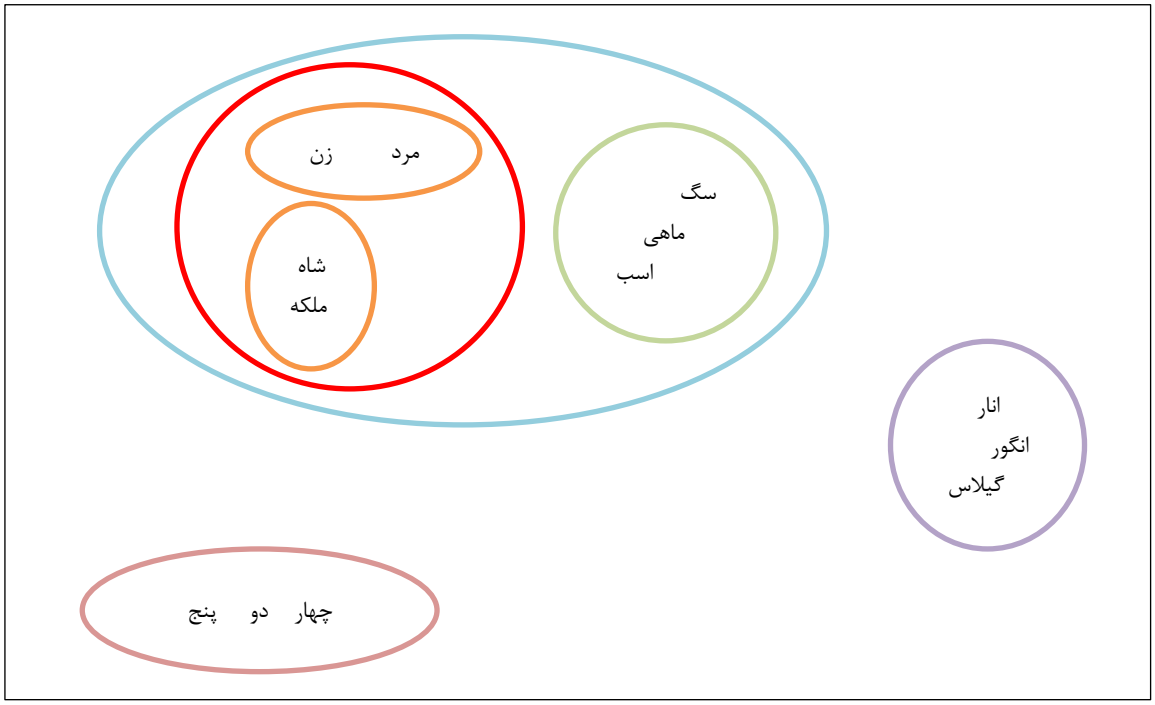
به این ترتیب جاسازی کلمات به هر کلمه یک بردار اختصاص می‌دهد. جاسازی کلمات این قدرت را دارد تا در لغت نامه‌های بزرگ شباهت و رابطه بین کلمات را یافته و آنها را دسته بندی کند. این کار موجب افزایش کارایی مدل‌ها در پردازش متن می‌شود. شکل ۳-۶ نمونه‌ای از دسته‌بندی کلمات مشابه و مرتبط را در جاسازی کلمات نمایش می‌دهد.

به دلیل قدرت لایه جاسازی کلمات در یافتن روابط بین کلمات و گروه‌بندی آنها در این پایان نامه تلاش شده است تا از ویژگی‌های لایه جاسازی برای یافتن روابط پیچیده بین جهش‌های ژنی و حل مسئله پراکندگی در یک ماتریس استفاده شود. به دلیل وجود اعداد صحیح مثبت در داده‌های پیش‌پردازش شده نیازی به مرحله ساخت لغت نامه در این مرحله نمی‌باشد. برای پیاده سازی لایه جاسازی[۱] از کتاب خانه

[۱] Embedding Layer

کراس[1] استفاده

شده است. این لایه، اعداد صحیح مثبت را به‌عنوان ورودی گرفته و یک بردار متراکم با طول ثابت برای هر



شکل ۳-۵ جاسازی کلمات تلاش کرده است رابطه میان کلمات را یافته و کلمات هم معنی و مرتبط را کنار هم قرار دهد مرجع[۴۸].

عدد ایجاد می‌کند. این بردار دارای وزن‌هایی است که در ابتدای آموزش مدل به صورت تصادفی تولید شده و در طول روند آموزش مدل به‌روزرسانی می‌شوند[۴۹, ۵۰] . برای مثال اگر وکتور [۰, ۱, ۲, ۳, ۴] به یک‌لایه جاسازی بدهیم و بخواهیم وکتورهایی با طول دو ایجاد کند. خروجی مانند جدول ۳-۱ خواهیم داشت

جدول ۳-۱ وکتورهای ایجاد شده در لایه جاسازی با طول ثابت دو.

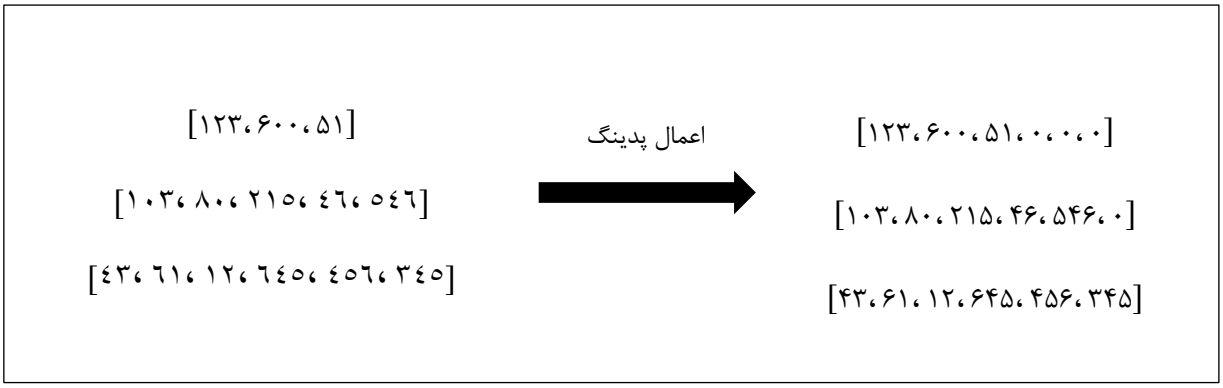
شاخص	جاسازی
۰	[۲,۱, ۳,۲]
۱	[۰,۵, ۳,۴]
۲	[۱,۲, ۱,۴]
۳	[۰,۷, ۱,۷]
۴	[۳,۰, ۷,۱]

[1] Keras

همان‌طور که در جدول بالا مشاهده می‌شود، به هر یک از اعداد صفر تا چهار که درون وکتور ما هستند یک وکتور با طول ثابت دو داده شده است. بر اساس جدول ۳-۱ خروجی لایه جاسازی وکتور [[۲,۱, ۳,۲] [۳,۰, ۷,۱] , [۰,۷, ۱,۷] , [۱,۲, ۱,۴] , [۰,۵, ۳,۴] , است.

در پردازش داده‌هایی که به صورت یک توالی هستند، بسیار متداول است که بعضی نمونه‌ها طول یکسانی نداشته باشند. بعضی به علت کم بودن داده‌ها طول کمی دارند (کوتاه هستند) و برخی دیگر ممکن است به علت زیاد بود داده‌ها طولانی باشند (بلند هستند). از آنجا که تمامی داده‌ها برای ورود به شبکه عمیق باید طول یکسانی داشته باشند باید طول آنها را یکی کنیم. برای اینکه طول این داده‌ها یکسان باشد به انتهای آن‌هایی که طول کم‌تری دارند صفر اضافه می‌کنیم. به این کار پدینگ[1] می‌گویند. ISR در بخش پیش‌پردازش برای ایجاد بردارهایی با طول یکسان از این روش استفاده کرده است. شکل ۷-۳ سه بردار با طول متفاوت را نمایش می‌دهد که با استفاده از پدینگ و اضافه کردن صفر به دو بردارد اول طول آنها را با بردار سوم یکی کرده است.

مقادیر پدینگ تاثیری در عمل دسته‌بندی ندارند به همین علت مدل باید آنها را نادیده بگیرد. به این منظور



شکل ۳-۶ اعمال پدینگ با اضافه کردن صفر به انتهای دو بردار اول برای یکی کردن طول آنها با بردار سوم مرجع[۴۸].

بعد از یکسان سازی طول داده‌ها با استفاده از تکنیک پدینگ از تکنیک ماسک زدن[2] برای آگاهی سازی

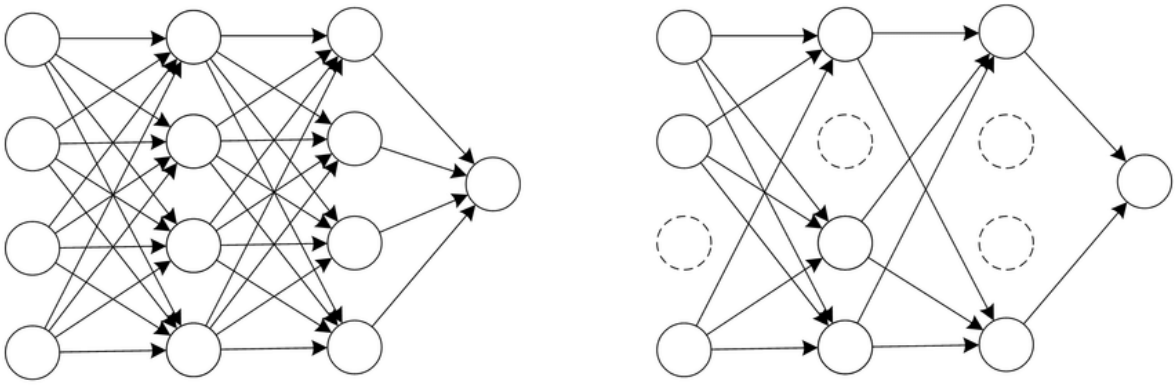
[1] Padding
[2] Masking

مدل از وجود پدینگ در داده‌ها استفاده می‌شود. این عمل با قراردادن مقدار درست[۱] برای پارامتر mask_zero در لایه جاسازی که توسط کتابخانه کراس ایجاد شده است انجام می‌شود.

۳-۵ لایه Dropout و Global max pooling

یک شبکه کانولوشنی استاندار از لایه‌های، پولینگ[۲]، کانولوشنی[۳] و تمام متصل[۴] ساخته شده است. به‌این‌ترتیب که لایه‌های کانولوشنی و پولینگ در ابتدای شبکه قرار دارند. نقش این دولایه استخراج ویژگی‌ها است و در انتها لایه‌های تمام متصل برای دسته‌بندی استفاده می‌شوند [۵۱]. از آنجا که لایه‌های تمام متصل مستعد بیش برازش هستند، می‌توانند مشکلاتی را در شبکه و عملکرد آن ایجاد کنند. برای حل این مشکل در این پایان نامه از لایه حذف تصادفی[۵] و پولینگ حداکثر[۶] تصادفی استفاده شده است [۵۲].

لایه حذف تصادفی یک روش منظم سازی است که تقریباً در بهبود آموزش تعداد زیادی از شبکه‌های عصبی با معماری‌های مختلف نقش دارد. در طول آموزش، تعدادی از نرون‌های لایه به طور تصادفی نادیده



شبکه عصبی تمام متصل قبل از اعمال حذف تصادفی | شبکه عصبی تمام متصل بعد از اعمال حذف تصادفی

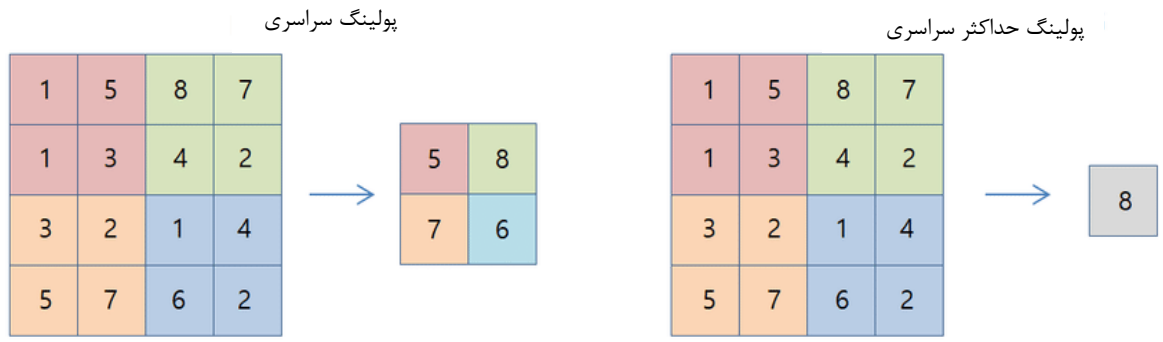
شکل ۳-۷ شکل سمت چپ شبکه عصبی تمام تصل را قبل از اعمال حذف تصادفی نمایش می‌دهد. شکل سمت راست همان شکبه را بعد از اعمال حذف تصادفی نمایش می‌دهد مرجع[۵۲].

---

[۱] True
[۲] Pooling
[۳] Convolutional
[۴] Fully-Connected
[۵] Dropout
[۶] Global Max Pooling

گرفته می‌شوند. به این ترتیب در هر بروز رسانی مدل یا پیکربندی جدیدی آموزش می‌بیند. حذف تصادفی فقط در مرحله آموزش انجام می‌شود و می‌توان آن را روی یک یا چند لایه اعمال کرد. شکل ۳-۸ یک شبکه عصبی تمام متصل را قبل و بعد از اعمال حذف تصادفی نمایش می‌دهد.

لایه‌های تمام متصل که به شکل سنتی در معماری CNN استفاده می‌شوند مستعد بیش برازش هستند.

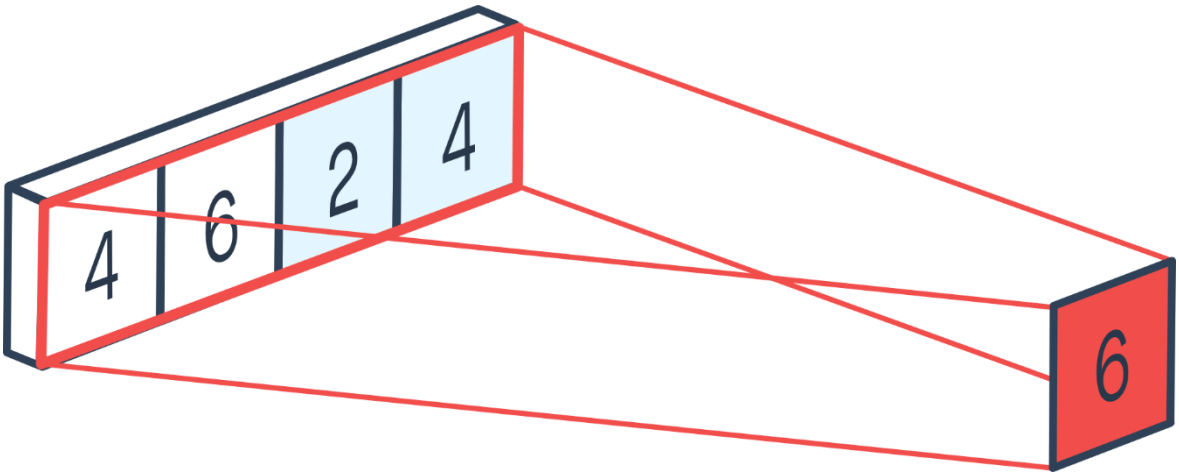


شکل ۳-۸ شکل سمت چپ نتیجه اعمال پولینگ حداکثر و شکل سمت راست نتیجه اعمال پولینگ حداکثر سراسری را نمایش می‌دهد مرجع[۵۲].

برای رفع این مشکل می‌توان آن‌ها را با لایه پولینگ حداکثر سراسری[۱] تعویض کرد. بر خلاف لایه پولینگ حدکثر[۲] که با استفاده از یک پنجره لغزان با ابعاد دلخواه بزرگ‌ترین اعداد درون پنجره را به عنوان خروجی انتخاب می‌کند. لایه پولینگ حداکثر سراسری بیشترین مقدار در هر ماتریس یا وکتور را بر می‌گرداند. شکل ۳-۹ تفاوت لایه پولینگ حداکثر سراسری و پولینگ حداکثر رانمایش می‌دهد.

[۱] Global Max Pooling
[۲] Max Pooling

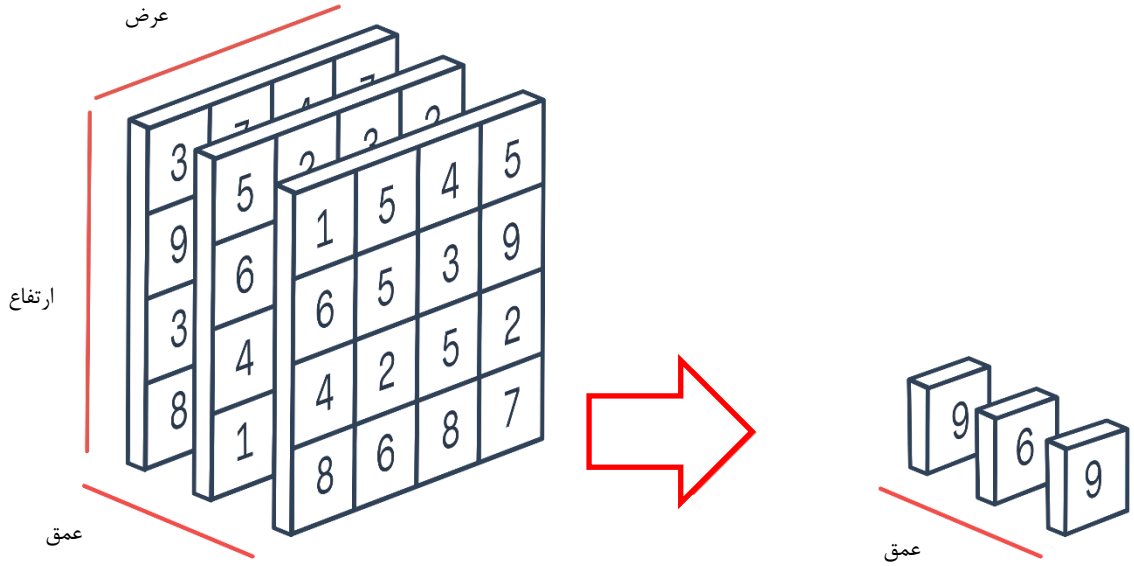
برای درک بهتر عملکرد لایه پولینگ حداکثر سراسری شکل ۳-۱۰ نیتجه اعمال این لایه را بر روی وکتوری از اعداد ۴، ۶، ۲، ۴ نشان می‌دهد که خروجی آن ۶ خواهد بود. همچنین شکل ۳-۱۱ بردار حاصل از اعمال این لایه بر روی ماتریس‌های خروجی CNN را نشان می‌دهد به این ترتیب به لایه صاف کننده[۱] نیز نیازی نیست.



شکل ۳-۹ عدد ۶ بزرگترین مقدار درون بردار سمت چپ است که با اعمال لایه پولینگ حدکثرسراسری به عنوان خروجی در سمت راست نمایش داده شده است مرجع[۵۲].

به‌این‌ترتیب ما یک نقشه ویژگی برای هر دسته ایجاد می‌کنیم و برای دسته‌بندی در لایه آخر مقدار بیشترین از هر نقشه ویژگی را گرفته و مستقیماً به تابع فعال‌ساز softmax می‌دهیم. به‌این شکل به می‌توانیم تأثیر بیش برازش که به دلیل استفاده از چند لایه تمام متصل بر روی مدل ایجاد می‌شود را کاهش دهیم.

[۱] Flattern



شکل ۳-۱۰ نتیجه اعمال لایه پولینگ حداکثر سراسری بر روی سه ماتریس خروجی CNN با عمق سه، یک وکتور به عمق سه خواهد بود که در سمت چپ شکل نمایش داده شده است مرجع[۵۲].

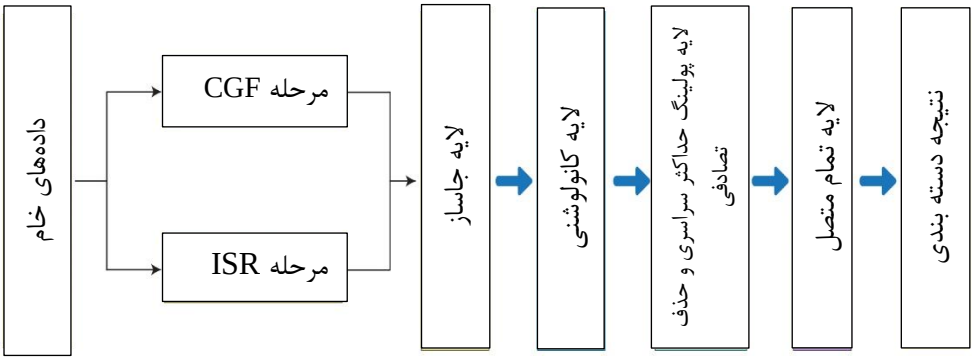
۳-۶ مدل CNN پیشنهادی

در این بخش، مدلی مبتنی بر یادگیری عمیق ارائه می‌دهیم. این مدل بر اساس روش شبکه‌های عصبی کانولوشنی برای استخراج ویژگی‌ها و شبکه عصبی تمام متصل برای طبقه‌بندی داده‌ها عمل می‌کند. از این مدل برای دسته‌بندی ۱۲ نوع سرطان بر اساس جهش‌های سوماتیک استفاده می‌کنیم و ساختار این مدل را با جزئیات شرح می‌دهیم. در این ساختار ابتدا داده‌های پیش‌پردازش شده با اندازه ۴۴۹۵ × ۳۱۲۲ وارد لایه جاسازی می‌شوند. این لایه به عنوان ورودی، یک تنسور[۱] دو بعدی دریافت کرده و یک تنسور سه بعدی به عنوان خروجی ایجاد می‌کند. لایه جاسازی در ابتدای آموزش اعداد تصادفی ایجاد می‌کند اما در طول روند آموزش این اعداد را تغییر می‌دهد تا بتواند جهش‌های مشابه را یافته و آنها را در بعدهایی نزدیک

[۱] Tensor

هم قرار دهد. همچنین این لایه پدینگ ایجاد شده در داده‌ها را نیز نادیده می‌گیرد. تنسور سه بعدی خروجی از لایه جاسازی برای استخراج ویژگی‌های مهم وارد لایه کانولوشنی یک بعدی[۱] می‌شود. تعداد ابعاد برای لایه جاسازی ۵۱۲ و طول ورودی برای این لایه ۴۴۹۵ تنظیم شده است. به دلیل خروجی سه بعدی لایه جاسازی باید از لایه کانولوشنی استفاده کرد که ورودی سه بعدی را قبول کند. در بین لایه‌های کانولوشنی ساخته شده توسط کتابخانه کراس فقط لایه کانولوشنی یک بعدی است که ورودی سه بعدی دریافت کرده و به عنوان خروجی نیز یک تنسور سه بعدی تولید می‌کند. اندازه هسته[۲] برای تک لایه کانولوشنی یک بعدی سه می‌باشد. این لایه دارای ۵۱۲ فیلتر[۳] است. خروجی این لایه برای کاهش اندازه وارد لایه پولینگ حداکثر سراسری می‌شود. بعد از کاهش اندازه این لایه یک تنسور دو بعدی به عنوان خروجی ایجاد می‌کند که می‌توان این تنسور دو بعدی را به صورت مستقیم به عنوان ورودی به یک لایه تمام متصل ارسال کرد. قبل از ارسال تنسور دو بعدی خارج شده از لایه پولینگ حداکثر سراسری این تنسور وارد یک لایه حذف تصادفی با مقدار نرخ ۵۰٪ می‌شود و در آخر یک لایه تمام متصل با ۱۲ نورون وظیفه دسته‌بندی را به عهده دارد.

شکل ۳-۱ چهارچوب روش پیشنهادی را نشان می‌دهد.



شکل ۳-۱۱ ساختار روش پیشنهادی، مدل CNN

[۱] Conv۱D
[۲] Kernel_Size
[۳] Filters

علاوه بر طراحی یک ساختار مناسب، انتخاب توابع فعال‌ساز و بهینه‌ساز تاثیر زیادی بر فعالیت مدل دارد. توابع فعال‌ساز و بهینه‌ساز زیادی موجود می‌باشد. با این حال، تابع بهینه ساز RMSprop با نرخ یادگیری ۰،۰۰۰۱ بهترین عملکرد را برای این مدل نشان داد. همچنین، برای همه لایه‌ها بجز لایه آخر از تابع فعال‌ساز Relu استفاده شده است. در لایه آخر نیز از تابع فعال‌ساز بیشینه هموار[۱] برای پیش‌بینی احتمالی نهایی هر کلاس استفاده می‌شود. فرمول ۳-۲ نحوه محاسبه تابع بیشینه هموار را نمایش می‌دهد. که در آن $\vec{z}$ وکتور ورودی، K تعداد کلاس‌ها در طبقه‌بندی چند کلاسه، e^{z_i} تابع نمایی استاندارد برای بردار ورودی و e^{z_j} تابع استاندارد نمایی برای بردار خروجی می‌باشد.

$$\sigma(\vec{z}) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad (۳-۲)$$

جدول ۳-۵ جزئیات داخلی ساختار CNN پیشنهاد شده در این پایان نامه را به صورت خلاصه بیان می‌کند.

جدول ۳-۲ جزئیات مدل CNN ساخته شده در این پایان نامه

نام پارامتر	مقدار
تعداد بعد خروجی لایه جاسازی	۵۱۲
طول ورودی به لایه جاسازی	۴۴۹۵
ماسک صفر در لایه جاسازی	TRUE
تعداد فیلتر لایه کانولوشن	۵۱۲
اندازه قدم در لایه کانولوشن	۱
اندازه هسته در لایه کانولوشن	۳
تابع فعال ساز در لایه کانولوشن	Relu
حداکثر هنجار در لایه کانولوشن	۲
پدینگ در لایه کانولوشن	valid
نرخ حذف در لایه حذف تصادفی	۵۰٪
تعداد گره‌ها در لایه تمام متصل	۱۲
تابع فعال ساز در لایه تمام متصل	Softmax
تابع خطا	Sparse Categorical Crossentropy
تابع بهینه ساز	RMSprop
اندازه بخش	۲۵۶
تعداد دور	۵۰

[۱] Softmax

۳-۷ جمع‌بندی

در این فصل، روش پیشنهادی و نوآوری ارائه شده مورد بررسی قرار گرفت. روش پیشنهادی دارای دو پیش‌پردازش و یک شبکه عصبی CNN یک لایه است. این دو پیش‌پردازش در مرجع [۵] ایجاد و استفاده شده است. هدف این دو پیش پردازش انتخاب زیرگروهی از ژن‌های موثر در دسته‌بندی و کاهش پراکندگی داده‌ها می‌باشد. برای کاهش تاثیر پراکندگی و یافتن ارتباط‌های پیچیده و مخفی بین ژن‌ها دراین پایان‌نامه از لایه جاسازی در کنار پیش‌پردازش‌های گفته شده استفاده می‌شود. سپس داده‌ها به یک مدل CNN فرستاده می‌شوند تا ویژگی‌های سطح بالاتر را استخراج کند. این ویژگی‌ها برای کار دسته‌بندی نهایی به لایه تمام متصل فرستاده می‌شود. همچنین برای کاهش تاثیر بیش‌برازش بر روی مدل پیشنهادی از روش‌های نظم دهی و لایه پولینگ حداکثر سراسری استفاده شده است.

فصل ۴ : نتایج و آزمایش‌ها

۴-۱ مقدمه

در فصل قبل روش پیشنهادی مورد بررسی قرار گرفت. در این فصل نحوه آموزش، اعتبارسنجی و آزمایش سه مدل CNN پیشنهادی، LSTM و ترکیب CNN – LSTM شرح داده شده و سپس نتایج مدل‌ها با یکدیگر مقایسه خواهد شد. به صورت کلی، روند آموزش و آزمایش مدل‌ها دارای چهار مرحله اصلی است. تمامی مراحل آموزش، آزمایش و ساخت مدل‌ها توسط زبان برنامه‌نویسی پایتون انجام شده است. برای ایجاد مجموعه‌داده‌ها و تقسیم آن‌ها از کتاب خانه‌های pandas وscikit-learn و برای ساخت مدل‌ها از Tensorflow و Keras استفاده می‌شود.

۴-۲ مجموعه دادگان

برای مقایسه عادلانه عملکرد مدل‌ها، از مجموعه‌داده ارائه شده در مرجع [۵] استفاده کرده‌ایم. این مجموعه‌داده TCGA-DeepGene نام دارد که زیرمجموعه‌ای از مجموعه‌داده TCGA[۱] [۵۳] است. ما جهش‌های ژن‌ها را با مقادیر دودویی نمایش می‌دهیم، مقدار صفر برای ژن‌های جهش نیافته و یک برای ژن‌های جهش یافته استفاده خواهند شد. TCGA-DeepGene دارای ۲۲۸۳۴ ژن و ۳۱۲۲ نمونه می‌باشد که ماتریسی با ابعاد ۲۲۸۳۴×۳۱۲۲ را ایجاد می‌کند. هر نمونه در یک ستون و هر ژن در یک سطر قرار دارد. هر یک از نمونه‌ها(ستون‌ها) دارای یک برچسب از ۱ تا ۱۲ است که علت آن ۱۲ نوع مختلف سرطانی است که در اختیار داریم. جدول ۴-۱ جزیات بیشتری از این مجموعه داده را نمایش می‌دهد.

[۱] Cancer Genome Atlas

جدول ٤-١ اطلاعات مربوط به نمونه‌های استفاده شده در مجموعه‌داده TCGA-DeepGene مرجع [٥].

نام سرطان	تعداد نمونه	جهش‌های Missense	جهش‌های Nonsense	جهش‌های Nonstop	RNA	Silent	Splice_Site	Translation start site	مجموع جهش‌ها
ACC	۹۱	۶۷۴۱	۵۰۱	۱۵	۳۶۸	۲۵۳۴	۳۴۴	۴۲	۱۰۵۴۵
BLCA	۱۳۰	۲۴۰۶۷	۲۱۴۲	۴۶	۰	۹۶۶۲	۵۲۸	۵۵	۳۶۵۰۰
BRCA	۹۹۲	۵۵۰۶۳	۴۸۴۱	۱۳۳	۳۹۹۸	۱۷۹۰۱	۱۴۲۴	۰	۸۳۳۶۰
CESC	۱۹۴	۲۶۶۰۶	۲۷۱۶	۸۴	۵۵۹۵	۹۷۶۵	۵۲۷	۰	۴۵۲۹۳
HNSC	۲۷۹	۳۱۴۱۶	۲۵۴۵	۴۴	۰	۱۲۱۴۹	۷۷۶	۰	۴۶۹۳۰
KIRP	۱۷۱	۸۹۱۰	۴۹۹	۱۷	۳۹۴	۳۴۱۱	۵۲۴	۰	۱۳۷۵۵
LGG	۲۸۴	۵۳۴۱	۳۷۸	۷	۱۰۲	۲۰۷۴	۲۹۴	۰	۸۱۹۶
LUAD	۲۳۰	۴۴۸۰۰	۳۴۷۷	۴۶	۰	۱۵۵۹۴	۱۳۷۷	۹۹	۶۵۳۹۳
PAAD	۱۴۶	۲۱۰۶۷	۱۴۹۶	۱۹	۸۵۹	۷۹۳۶	۱۰۰۵	۱۱۱	۳۲۴۹۳
PRAD	۲۶۱	۹۶۲۸	۵۶۳	۱۵	۶۵۲	۳۷۵۰	۵۱۳	۵۵	۱۵۱۷۶
STAD	۲۸۸	۸۲۲۶۵	۴۲۰۰	۹۲	۴۸	۳۳۳۴	۱۸۶۸	۲۲۷	۱۲۲۰۴۴
UCS	۵۶	۳۰۷۰	۱۸۷	۲	۲۳۴	۱۱۱۴	۱۷۱	۰	۴۷۷۸
مجموع	۳۱۲۲	۳۱۸۹۷۴	۲۳۵۴۵	۵۲۰	۱۲۲۵۰	۱۱۹۲۳۴	۹۳۵۱	۵۹۸	۴۸۴۴۶۳

۴-۳ نحوه تقسیم مجموعه‌داده و هایپرپارامترها

در فصل قبل در مورد دو پیش‌پردازش استفاده شده در این پایان‌نامه صحبت کردیم. این پیش‌پردازش‌ها بر روی داده‌های خام اعمال می‌شوند و خروجی حاصل از پیش‌پردازش‌ها را برای آموزش و آزمایش مورداستفاده قرار می‌دهیم. به این منظور ده درصد از کل داده‌ها را برای آزمایش جدا می‌کنیم و به آن مجموعه‌داده آزمایش[۱] می‌گوییم. ۹۰ درصد باقی‌مانده را برای آموزش استفاده خواهیم کرد و به آن مجموعه‌داده آموزش[۲] خواهیم گفت. این کار را با استفاده از روش train_test_split که توسط کتابخانه sklearn پیاده‌سازی شده است انجام می‌دهیم.

زمانی که در حال آموزش مدل‌های خود بر روی مجموعه‌داده آموزش هستیم، ده درصد از کل این مجموعه داده را جدا کرده و از آن برای اعتبارسنجی عملکرد مدل‌های خود استفاده می‌کنیم که به آن مجموعه‌داده اعتبارسنجی[۳] نیز گفته می‌شود. در هنگام استفاده از اعتبار متقابل ۱۰ برابر، ابتدا مدل را بر روی

[۱] Test Data Set

[۲] Train Data Set

[۳] Validation Set/Cross Validation Set

مجموعه‌داده آموزش، آموزش می‌دهیم. در هر مرتبه از روند آموزش، قسمتی از داده به‌عنوان مجموعه اعتبارسنجی کنار گذاشته می‌شود و وزن‌هایی با بیشترین مقدار دقت در اعتبارسنجی ذخیره خواهند شد. در انتها، وزن‌ها بر روی مدل بارگذاری می‌شوند و با استفاده از مجموعه‌داده آزمون، مورد آزمایش قرار می‌گیرند. بیشترین مقدار دقت بر روی مجموعه‌داده‌های آزمون گزارش می‌شوند. در این آزمایش‌ها تمرکز ما بر روی ابرپارامترهای زیر می‌باشد: تعداد بعد خروجی برای لایه جاسازی بین مقادیر [۱, ۲, ۴, ۸ ,۱۶, ۳۲, ۶۴, ۱۲۸, ۲۵۶, ۵۱۲, ۱۰۲۴]، مقدار نرخ حذف برای لایه حذف تصادفی بین مقادیر [۰,۵, ۰,۶, ۰,۷, ۰,۸, ۰,۹]، مقدار پارامتر، حداکثر هنجار در بین مقادیر [۱, ۲, ۳, ۴]، مقدار نرخ یادگیری بین مقادیر [۰,۱, ۰,۰۱, ۰,۰۰۱, ۰,۰۰۰۱]، تعداد سلول‌های لایه LSTM بین مقادیر [۱, ۵, ۱۰, ۱۵, ۲۰, ۲۵, ۳۰, ۳۵, ۴۰, ۴۵, ۵۰] و تعداد فیلترهای لایه CNN بین مقادیر [۲, ۴, ۸, ۱۶, ۳۲, ۶۴, ۱۲۸, ۵۱۲] هستند.

برای تنظیم ابرپارامترها[۱]، روش‌هایی مانند تنظیم دستی، جستجوی شبکه[۲] و جستجوی تصادفی[۳] وجود دارد. ما در آزمایش‌های خود برای تنظیم ابرپارامترها از کتابخانه تنظیم‌کننده کراس[۴] و جستجوی تصادفی استفاده کرده‌ایم. این روش جستجو در مقایسه با جستجوی شبکه به طور شگفت انگیزی در فضای جستجو با ابعاد بالا کارآمدتر است. در جستجوی شبکه اگر مقادیر پارامترهای بهینه در فضای جستجوی ما موجود باشند در نهایت مقدار بهینه پارامترها پیدا می‌شود. اما ممکن است از جستجوی شبکه استفاده کنیم و هرگز مقدار بهینه نهایی یافت نشود به این دلیل که مقادیر در فضای جستجوی ما وجود نداشته باشند. به این ترتیب زمان زیادی برای آزمایش تمامی مقادیر به هدر رفته است. همچنین در جستجوی شبکه از آنجا که تمامی ترکیب‌های ممکن از مقادیر ابرپارامترها باید چک شود. تعداد آزمایش‌هایی که باید انجام شود از فرمول ۴-۱ بدست می‌آید. در این فرمول K تعداد کل ابرپارامترها و L^k تعداد کل مقادیر در هر ابرپارامتر است. [۵۴]

---

[۱] Heyperparameter
[۲] Grid Search
[۳] Random Search
[۴] Keras - Tuner

$$S = \Pi_{k=1}^{K} |L^{(k)}| \qquad (۱-۴)$$

این حاصل ضرب بر روی مجموعه‌های K باعث می‌شود جستجوی شبکه از نفرین ابعاد[۱] رنج ببرد زیرا تعداد مقادیر مشترک با تعداد ابرپارامترها به طور تصاعدی افزایش می‌یابد. این موضوع باعث افزایش هزینه محاسباتی در جستوجوی شبکه می‌شود. اما جستجوی تصادفی معمولاً در تکرارهای بسیار کمتر، مقدار به اندازه کافی نزدیک، به مقدار بهینه را پیدا می‌کند. در هنگام تنظیم یک مدل یادگیری ماشین یا مدل یادگیری عمیق تعداد کمی از ابرپارامترها هستند که تنظیم آنها باعث بهبود عملکرد مدل بر روی مجموعه داده‌ها می‌شود و این ابرپارامترها در مجموعه داده‌های متفاوت یکی نیستند. جستجوی شبکه زمان زیادی را صرف ارزیابی مناطق غیرمعمول فضای جستجوی ابرپارامترها می‌کند زیرا باید هر ترکیب در شبکه را ارزیابی کند. در مقابل، جستجوی تصادفی با این فرضیه که تمامی پارامترها به یک اندازه مهم نیستند با ترکیب تصادفی مقادیر ابرپارامترها شانس بیشتری دارد تا در زمان کمتر مقدار نزدیک به بهینه نهایی را بیابد حتی اگر مقدار بهینه نهایی در فضای جستجوی ما وجود نداشته باشد. [۵۴]

برای جستجوی تصادفی مقدار حداکثر آزمایش‌ها[۲] را بر روی ۲۰ و تعداد اجرا در هر آزمایش[۳] را دو قرار دادیم. همچنین مقدار ده درصد از مجموعه آموزش را برای اعتبارسنجی در این مرحله جدا کردیم. به دلیل عدم تعادل در مجموعه‌داده که در جدول ۴-۱ قابل‌مشاهده است، برای مرحله اعتبار متقابل ۱۰ برابر از روش طبقه‌بندی کا - تائی[۴] استفاده کرده‌ایم. این روش به ما اطمینان می‌دهد که به مقدار مساوی از هر نمونه در هر یک از مجموعه‌داده‌ها وجود خواهد داشت.

---

۱ Curse Of Dimensionality

۲ Max Trials

۳ Executions Per Trial

۴ Stratified-K-Fold

۴-۴ ساخت و تنظیم مدل CNN پیشنهادی

در مرحله اول، ابتدا با استفاده از تنظیم‌کننده کراس[۱] ابرپارامترهای مربوط به مقدار نرخ یادگیری، تعداد فیلترها، مقدار خروجی - بعد[۲] را برای CNN و لایه جاسازی تنظیم می‌کنیم. جدول ۴-۲ ده مورد از بهترین مقادیر مربوط به ابرپارامترها را در مرحله اول که بیشترین مقدار دقت اعتبارسنجی[۳] را داشته‌اند نمایش می‌دهد. بر اساس این جدول بهترین مقادیر برای این سه ابرپارامتر به ترتیب ۵۱۲، ۵۱۲ و ۰/۰۰۱ می‌باشد.

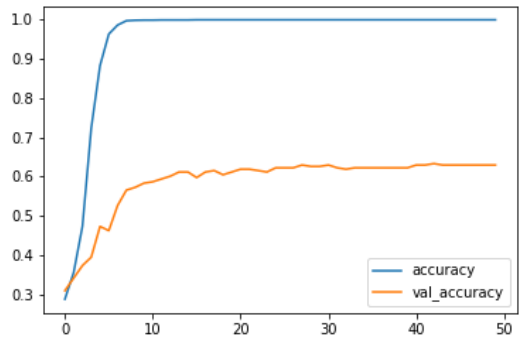
جدول ۴-۲ ده مورد از بهترین مقادیر ابرپارامترها در مرحله اول ساخت مدل CNN

دقت اعتبارسنجی	نرخ یادگیری	ابعاد خروجی	تعداد فیلتر
۰/۶۶۰۱	۰/۰۰۱	۵۱۲	۵۱۲
۰/۶۲۴۵	۰/۰۱	۶۴	۵۱۲
۰/۵۹۰۷	۰/۰۰۰۱	۱۰۲۴	۲۵۶
۰/۵۶۵۸	۰/۰۱	۱۰۲۴	۱۲۸
۰/۴۱۶۳	۰/۰۱	۳۲	۱۶
۰/۴۱۱	۰/۰۰۰۱	۱۲۸	۱۶
۰/۴۰۹۲	۰/۰۰۱	۴	۵۱۲
۰/۴۰	۰/۰۱	۴	۲۵۶
۰/۳۹۶۷	۰/۰۰۱	۱۰۲۴	۱۶
۰/۳۸۲۵	۰/۰۰۱	۴	۶۴

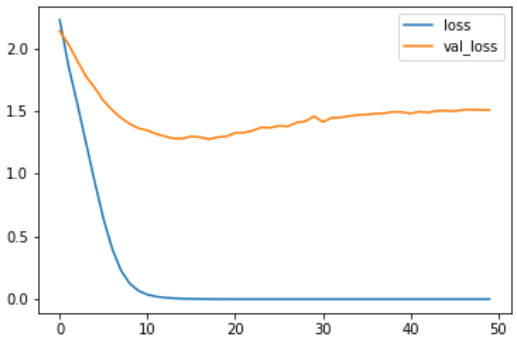
در بخش قبل توضیح داده شد که برای کاهش پراکندگی داده‌ها و یافتن جهش‌های مشابه از لایه جاسازی استفاده می‌کنیم. برای نمایش تأثیر این لایه بر روی مدل CNN و همین‌طور بررسی عملکرد ابر پارامترها پیشنهاد شده توسط تنظیم‌کننده کراس مدل CNN ساخته شده در مرحله قبل را بدون لایه جاسازی آزمایش می‌کنیم. سپس عملکرد مدلی که از لایه جاسازی استفاده کرده است و مدلی که از این لایه استفاده

[۱] Keras-Tuner
[۲] Output-Dim
[۳] Validation Accuracy

نکرده است را مورد بررسی قرار خواهیم داد. تصاویر ۴-۱ و ۴-۲ روند مقادیر دقت و خطا را در مدل دارای لایه جاسازی و تصاویر ۴-۴ و۴-۳ همین مقادیر را در مدلی که از لایه جاسازی استفاده نمی‌کند نشان می‌دهد.

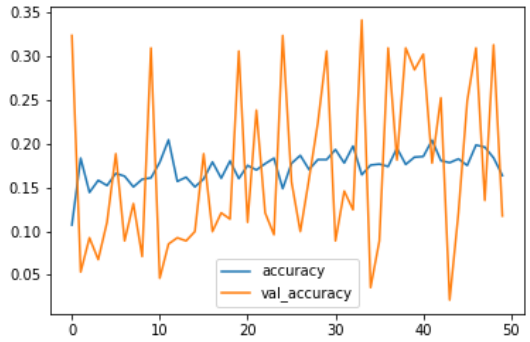


شکل ۴-۱ روند دقت در آموزش و دقت در اعتبار سنجی در ساخت مدل CNN دارای لایه جاسازی

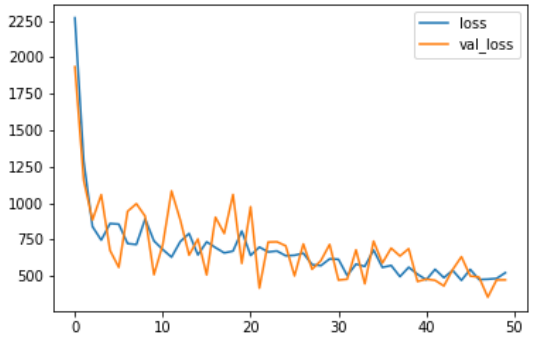


شکل ۴-۲ روند خطا در آموزش و خطا در اعتبار سنجی در ساخت CNN دارای لایه جاسازی

با مقایسه تصاویر ۴-۱ و ۴-۴ می‌توان مشاهده کرد که مقدار دقت در مدل دارای لایه جاسازی ۶۰٪ است و همین مورد در مدلی که از لایه جاسازی استفاده نمی‌کند ۳۵٪ می‌باشد. همچنین با مقایسه تصاویر ۴-۲ و ۴-۳ بیشترین مقدار خطا در مدلی که دارای لایه جاسازی است ۱,۵ و کمترین مقدار خطا در مدلی



شکل ۴-۴ روند دقت در آموزش و دقت در اعتبار سنجی در ساخت CNN بدون لایه جاسازی



شکل ۴-۳ روند خطا در آموزش و خطا در اعتبار سنجی در ساخت CNN بدون لایه جاسازی

که از لایه جاسازی استفاده نمی‌کند ۵۰۰ است. باتوجه‌به مقادیر بدست آمده در این مرحله همواره در مدل‌های خود از لایه جاسازی استفاده می‌کنیم. در شکل ۴-۲ مشاهده می‌کنیم که روند کاهش مقدار خطا

در دوره پانزده شروع به افزایش کرده و این افزایش ادامه می‌یابد تا جایی که تفاوت بسیار چشم‌گیری بین مقدار خطا در آموزش و مقدار خطا در اعتبارسنجی دیده می‌شود. این تصویر بیانگر مشکل بیش برازش است. در مرحله دوم برای رفع این مشکل ابر پارامترهای مربوط به روش‌های نظم‌دهی را توسط تنظیم‌کننده کراس بدست می‌آوریم. جدول ۴-۳ ده مورد از بهترین مقادیر برای مقدار حذف تصادفی و حداکثر هنجار[۱] را نمایش می‌دهد که به ترتیب مقادیر ۰/۵ و ۲ است. در مرحله سوم با استفاده از مقادیر بدست آمده از جدول ۴-۲ و ۴-۳ مدلی اصلی را می‌سازیم تا عملکرد مدل قبل از اعمال روش اعتبار متقابل ده برابر[۲] مورد برسی قرار دهیم. با مقایسه شکل ۴-۲ و ۴-۵ می‌توان مشاهده کرد که روند خطا در اعتبارسنجی در ۴-۵ تا دوره ۲۰ کاهش می‌یابد و بعد از آن به‌آرامی دوباره اوج می‌گیرد، اما در شکل ۴-۲ روند کاهش خطای اعتبارسنجی فقط تا دوره ۱۵ است و بعد از آن افزایش می‌یابد. همچنین با مقایسه شکل ۴-۱ و ۴-۶ می‌توان افزایش دقت اعتبار سنجی را مشاهده نمود که به دلیل استفاده از روش‌های نظم دهی می‌باشد.

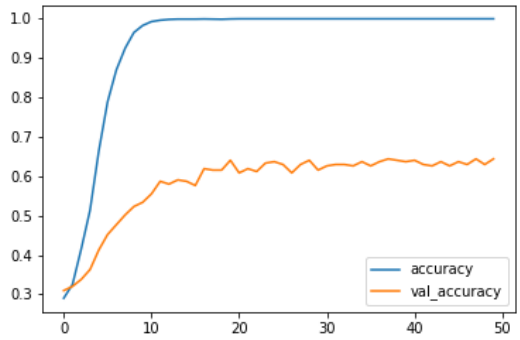
جدول ۴-۳ ده مورد از بهترین مقادیر برای هایپرپارامترها در مرحله دوم ساخت مدل CNN

دقت اعتبارسنجی	حداکثر هنجار	حذف تصادفی دوم	حذف تصادفی اول
۰/۷۱۳۵	۲	۰/۵	۰
۰/٦۹۳۹	۲	۰/٦	۰
۰/٦۹۲۱	۱	۰/۸	۰
۰/٦۹۲۱	٤	۰/۵	۰
۰/٦۹۲۱	۳	۰/۷	۰
۰/٦۹۲۱	٤	۰/٦	۰
۰/٦۸٦۸	۱	۰/۵	۰/۵
۰/٦۸۵۰	۲	۰/۸	۰
۰/٦۸۳۲	٤	۰/۷	۰
۰/٦۷۹۷	۳	۰/۸	۰/۵

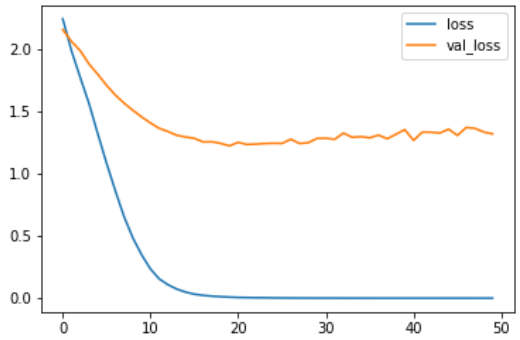
[۱] Max-Norm

[۲] Ten Fold Cross-Validation

همچنین با مقایسه جداول ۴-۲ و ۴-۳ می‌توان دریافت که مقدار دقت در اعتبارسنجی از ۰,۶۶۰۱ به مقدار ۰,۷۱۳۵ افزایش‌یافته است. با تنظیم مقادیر حذف تصادفی و حداکثر هنجار توانستیم عملکرد مدل CNN خود را بهبود ببخشیم. اکنون با استفاده از بهترین مقادیر معرفی شده در جدول‌های ۴-۲ و ۴-۳ مدل را با استفاده از تکنیک اعتبار متقابل ده برابر آموزش داده، اعتبارسنجی کرده و آزمایش می‌کنیم. به این صورت



شکل ۴-۵ روند دقت در آموزش و دقت در اعتبار سنجی در مرحله سوم ساخت مدل CNN



شکل ۴-۶ روند خطا در آموزش و خطا در اعتبار سنجی در مرحله دوم ساخت سوم CNN

که در هر یک از ۱۰ مرحله بهترین وزن‌ها را که بیشترین مقدار دقت در اعتبارسنجی را داشته باشد ذخیره کرده و بعد از تمام شدن آموزش هر یک از این وزن‌ها را با استفاده از مجموعه‌داده آزمایش، آزمایش می‌کنیم. سپس بیشترین مقدار دقت در آزمایش را به‌عنوان نتیجه نهایی اعلام می‌کنیم. این مرحله چهارم و آخرین مرحله در روند آموزش CNN است که نتیجه آن در جدول ۴-۴ قابل‌مشاهده می‌باشد.

جدول ۴-۴ بهترین نتیجه در اعتبارمتقابل ده برابر، آخرین مرحله از ساخت CNN

انحراف معیار دقت آزمون	متوسط دقت آزمون	خطای آزمون	دقت آزمون
-/+ ۱/۲٤٦	٦۳/۳٥	۱/۲۲۰۷	۰/٦٦٤٥

۴-۵ ساخت و تنظیم مدل LSTM

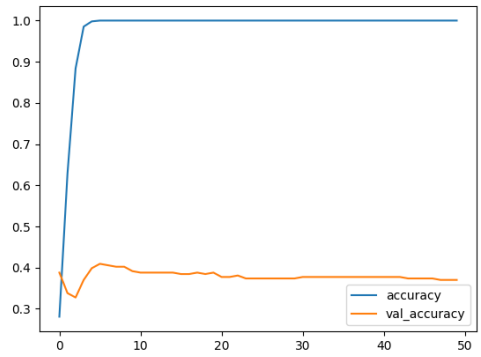
همانند ساخت مدل CNN ساخت و تنظیم مدل LSTM دارای چهار قسمت می‌باشد. تمامی ابر پارامترها با استفاده از تنظیم‌کننده کراس برای هر مدل انتخاب می‌شود. ابتدا با استفاده از تنظیم‌کننده کراس مقادیر تعداد واحدهای LSTM، تعداد بعد خروجی لایه جاسازی و نرخ یادگیری را تنظیم می‌کنیم. جدول ۴-۵ ده مورد از بهترین مقادیر برای این ابر پارامترها را به ترتیب نمایش می‌دهد. بهترین ابر پارامترها آن‌هایی هستند که بیشترین دقت در اعتبارسنجی را به ما می‌دهد. بهترین مقادیر برای هایپرپارامترهای ذکر شده به ترتیب ۶۴، ۱۰۲۴ و ۰،۰۱ می‌باشند.

جدول ۴-۵ ده مورد از بهترین مقادیر هایپرپارامترها در مرحله اول ساخت مدل LSTM

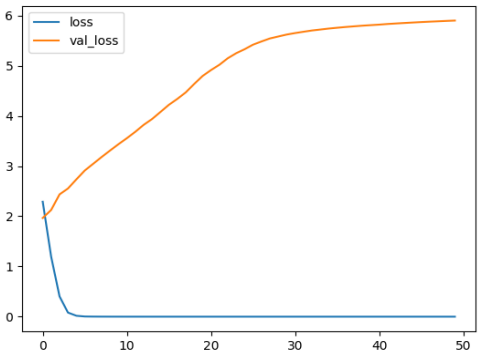
دقت اعتبارسنجی	نرخ یادگیری	بعد خروجی	سلول
۰/٤٤٤٨	۰/۰۱	۱۰۲٤	٦٤
۰/٤۲۷	۰/۰۱	۳۲	۳۲
۰/٤۲۳٤	۰/۰۱	٦٤	۳۲
۰/٤۱٦۳	۰/۰۰۱	٥۱۲	۳۲
۰/۳۹٥	۰/۰۰۱	٤	۳۲
۰/۳۸٤۳	۰/۰۰۰۱	۲٥٦	۳۲
۰/۳۸۰۷	۰/۰۰۰۱	۱۰۲٤	۱٦
۰/۳۸۰۷	۰/۰۱	۳۲	۲
۰/۳۷٥٤	۰/۰۰۱	۱۰۲٤	٤
۳	۰/۰۱	٤	۱٦

در فصل قبل گفته شد که از لایه جاسازی برای بهبود مدل‌های ساخته شده در این پایان‌نامه استفاده شده است. برای بررسی تاثیر این لایه بر روی مدل LSTM با استفاده از بهترین مقادیر ابر پارامترهای بدست آمده در جدول ۴-۵ مدل را یک بار بدون لایه جاسازی آموزش داده و اعتبار سنجی می‌کنیم. در این آزمایش مدل دارای لایه جاسازی دارای دقت اعتبار سنجی ٪۴۴/۴۸ و مدل بدون لایه جاسازی دارای دقت اعتبار سنجی ٪۳۷/۹۰ می‌باشد. اختلاف این دو مقدار برابر ۰/۱۷۳۶ است. این اختلاف نشان دهنده

رشد۱۷,۳۶٪ دقت اعتبار سنجی از ۳۷/۹۰٪ به ۴۴/۴۸٪ است. بر همین اساس از لایه جاسازی در مدل LSTM در تمامی مراحل پیش رو استفاده شده است. در مرحله دوم، روند دقت و خطا را برای مدلی که در مرحله یک تعلیم دیده و اعتبارسنجی شده است بررسی می‌کنیم. شکل ۴-۸ روند دقت را نمایش



شکل ۴-۸ روند دقت در آموزش و دقت در اعتبار سنجی در مرحله دوم ساخت مدل LSTM



شکل ۴-۷ روند خطا در آموزش و خطا در اعتبار سنجی در مرحله دوم ساخت مدل LSTM

می‌دهد. همان طور که در شکل مشاهده می‌شود، دقت در آموزش در چند دوره اول به ۱۰۰٪ و دقت در اعتبار سنجی با وجود پنجاه دوره در حدود ۴۰٪ است. همچنین با بررسی شکل ۴-۷ می‌توان مشاهده کرد که مقدار خطا در آموزش صفر و خطا در اعتبار سنجی شش است. به دلیل وجود تعداد زیاد ویژگی‌ها نسبت به تعداد نمونه‌ها در مجموعه‌داده‌ها این اختلاف شدید بین دقت در آموزش و دقت در اعبارسنجی و همچنین خطا در آموزش و خطا در اعتبار سنجی رخ می‌دهد که نشان دهنده مشکل بیش برازش در مدل LSTM می‌باشد. برای رفع این مشکل، در مرحله سوم، هایپرپارامترهای مربوط به روش‌های نظم‌دهی را توسط تنظیم‌کننده کراس تنظیم می‌کنیم. جدول ۴-۶ ده مورد از بهترین مقادیر را برای حذف تصادفی و حداکثر هنجار نمایش می‌دهد که به ترتیب شش دهم و دو می‌باشند. همان‌طور که در جدول ۴-۶ مشاهده می‌شود، دقت اعتبارسنجی به ۴۶/۴۴٪ رسیده است. با مقایسه جدول ۴-۵ و ۴-۶ می‌توان نتیجه گرفت که با استفاده از روش‌های نظم‌دهی مقدار دقت در اعتبارسنجی از ۴۴/۴۸٪ به ۴۶/۴۴٪ افزایش‌یافته است.

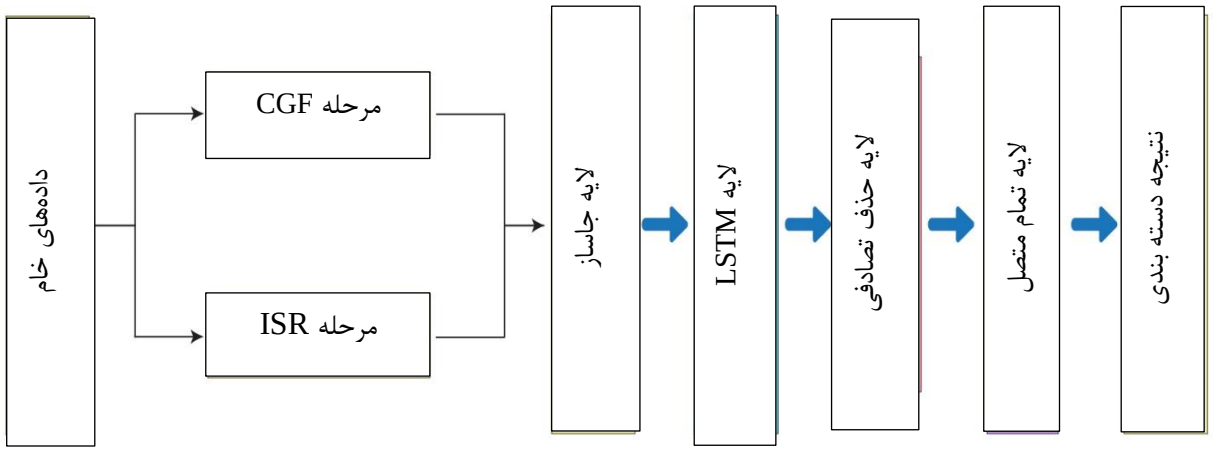
. در مرحله چهارم با استفاده از تکنیک اعتبار متقابل ۱۰ برابر و بهترین مقادیر بدست آمده در مراحل قبل برای هایپرپارامترهای استفاده شده، مدل خود را آموزش داده، اعتبارسنجی کرده و سپس آزمایش می‌کنیم. در هر دوره، بهترین وزن‌ها با بهترین دقت در اعتبارسنجی را نگه می‌داریم. سپس این وزن‌ها را بر روی داده‌های آزمون، آزمایش می‌کنیم و بالاترین دقت آزمون را گزارش می‌کنیم. جدول ۴-۷ نتایج مربوط به این مرحله را نمایش داده و شکل ۴-۹ ساختار مدل نهایی مدل LSTM است.

جدول ٤-٦ ده مورد از بهترین مقادیر هایپرپارامترها در مرحله سوم ساخت مدل LSTM

دقت اعتبارسنجی	حداکثر هنجار	حذف تصادفی
۰/۴۶۴۴	۲	۰/۶
۰/۴۶۲۶	۴	۰/۵
۰/۴۵۷۲	۳	۰/۵
۰/۴۵۵۱	۴	۰
۰/۴۵۳۷	۲	۰/۵
۰/۴۵۳۷	۳	۰
۰/۴۵۱۹	۳	۰/۶
۰/۴۵۱۹	۲	۰/۷
۰/۴۴۸۳	۴	۰/۸
۰/۴۴۸۳	۱	۰

جدول ۴-۷ نتیجه بهترین مدل در اعتبار متقابل ۱۰ برابر آخرین مرحله از ساخت مدل LSTM

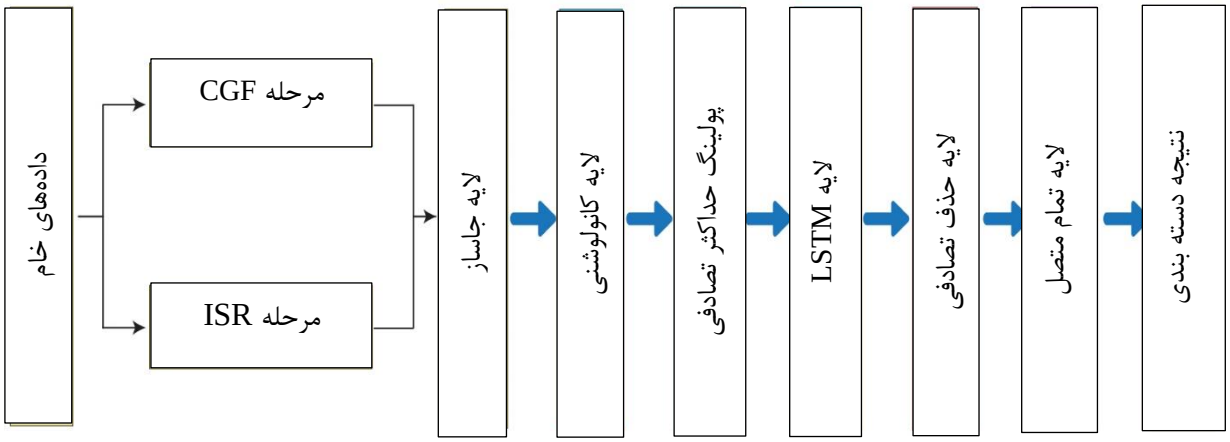
انحراف معیار دقت آزمون	متوسط دقت آزمون	خطای آزمون	دقت آزمون
+/-۲/۲۹۵۲	۳۶/۶۴۵۳	۲/۱۳۸۶	۰/۴۰۸۹



شکل ۴-۹ ساختار نهایی مدل LSTM

۴-۶ ساخت و تنظیم مدل CNN + LSTM

در بخش‌های ۴-۳ و ۴-۴، بهترین پارامترها را برای ساخت مدل‌های CNN و LSTM بدست آوردیم. با استفاده از این مقادیر، مدل ترکیبی CNN و LSTM را می‌سازیم. در شکل ۴-۱۰، ساختار این مدل



شکل ۴-۱۰ ساختار نهایی مدل CNN + LSTM

ترکیبی نمایش‌داده‌شده است. برای ساخت این مدل، داده‌ها بعد از دو پیش‌پردازش ISR و CGF وارد لایه جاسازی می‌شوند. سپس خروجی این لایه به‌عنوان ورودی به لایه CNN داده می‌شود. سومین لایه

LSTM است که خروجی CNN وارد آن می‌شود. در نهایت داده‌ها برای دسته‌بندی به یک‌لایه تمام متصل ارسال می‌شوند. با استفاده از روش اعتبار متقابل ۱۰ برابر ما بهترین وزن‌ها را در هر دوره ذخیره می‌کنیم. سپس این وزن‌ها را بر روی مجموعه‌داده آزمون، آزمایش می‌کنیم و بهترین دقت آزمون را گزارش می‌دهیم. مدل ساخته شده به‌دقت آزمون ۴۱,۲۰٪ دست‌یافته است.

۴-۷ ارزیابی عملکرد مدل پیشنهادی

در ادامه این بخش، به بررسی دقت و ارزیابی عملکرد مدل پیشنهادی (CNN) در مقایسه با دو مدل LSTM و ترکیب CNN – LSTM ساخته شده در این پایان‌نامه و مدل ارائه شده در مرجع [۵] می‌پردازیم. از آنجا که برای مقایسه DeepGene با SVM, KNN و NB از معیار دقت در آزمون استفاده شده است. برای رعایت عدالت در مقایسه نتایج ما از همین معیار برای مقایسه عملکرد روش‌ها استفاده می‌کنیم. بعد از پیاده سازی الگوریتم‌های مربوطه بر روی مجموعه داده‌ها و پیش‌پردازش‌های یکسان نتایج حاصل شده در جدول ۴-۸ قابل مشاده می‌باشد.

جدول ۴-۸ مقایسه دقت در روش پیشنهادی و سایر روش‌ها

نام مدل	دقت آزمون
CNN	۶۶/۴۵
LSTM	۴۰/۸۹
CNN + LSTM	۴۱/۲۰
DeepGene	۶۵/۵۰
SVM	۵۲/۷۰
KNN	۴۰/۸۰
NB	۹/۲۳

با توجه به جدول ۴-۸ مدل CNN پیشنهادی بیشترین مقدار دقت آزمون در مقایسه با مدلهای دیگر را بدست آورده است. مدل CNN پیشنهادی با استفاده از لایه جاساز برای یافتن ارتباط‌های پیچیده بین ژن‌ها و قرار دادن ژن‌های مشابه در دسته‌های یکسان به شبکه کانولوشنی در یافتن و استخراج ویژگی‌های

سطح بالاتر کمک می‌کند. این همکاری موجب افزایش دقت مدل پیشنهادی نسبت به سایر روش‌های نام برده شده می‌باشد.

۸-۴ نتیجه‌گیری

در این فصل نحوه پیاده‌سازی روش پیشنهادی، مدل LSTM و مدل CNN + LSTM را شرح داده و خروجی‌های حاصل از آنها را در مراحل مختلف نشان دادیم. همچنین باتوجه‌به جدول ۴-۸ می‌توان مشاهده کرد که دقت حاصل از روش پیشنهادی بیشتر از روش ارائه شده در مرجع [۵]، روش‌های یادگیری ماشین و دو مدل عمیق آزمایش شده دیگر می‌باشد. به همین دلیل مدل CNN به عنوان روش پیشنهادی ارائه شده است.

فصل ۵ جمع‌بندی و پیشنهادها

۱-۵ جمع‌بندی و نتیجه‌گیری

در این پایان‌نامه، سه مدل یادگیری عمیق CNN, LSTM و ترکیب این دو برای دسته‌بندی سرطان بر اساس جهش‌های سوماتیک مورد آزمایش و بررسی قرار گرفته است. از بین این سه مدل، مدل CNN پیشنهادی بهترین عملکرد را از خود نشان داد. در این پایان‌نامه تمرکز ما بر روی یافتن ارتباطات پیچیده جهش‌های ژنی برای مشخص‌کردن جهش‌های مشابه، ساخت و تنظیم یک مدل یادگیری عمیق برای کار دسته‌بندی و رفع بیش برازش موجود در مدل پیشنهادی بود. این سیستم با استفاده از لایه جاسازی جهش‌های مشابه را یافته و آنها را در ابعاد نزدیک به هم قرار می‌دهد. همچنین لایه پولینگ حداکثر سراسری و روش‌های نظم‌دهی باعث بهبود عملکرد مدل پیشنهادی در مواجهه با مشکل بیش برازش شدند. این سیستم ابتدا با استفاده از لایه جاسازی داده‌های ورودی را در بعدهای تعیین شده قرار می‌دهد. سپس، در طول مدت یادگیری مدل این لایه داده‌های مشابه را نزدیک به یکدیگر قرار می‌دهد. به‌این‌ترتیب جهش‌های مشابه در دسته‌هایی نزدیک به یکدیگر در ابعاد مختلف قرار می‌گیرند. همچنین این لایه موجب کاهش اثر پراکندگی بیش از حد داده‌ها و تأثیر آن بر روی مدل پیشنهادی شده و پدینگ موجود در داده‌ها را که برای یکسان‌سازی طول آن‌ها به‌کاررفته را خنثی می‌کند تا در عملکرد مدل خللی ایجاد نشود. به دلیل مستعد بودن لایه‌های تمام متصل برای ایجاد بیش برازش از لایه پولینگ حداکثر سراسری استفاده شد تا خروجی لایه کانولوشنی مستقیماً به لایه آخر برای دسته‌بندی ارسال شود. همچنین از روش‌های نظم‌دهی مانند لایه حذف تصادفی برای کاهش اثر بیش برازش استفاده شد.

در ادامه مدلی مبتنی بر یادگیری عمیق جهت طبقه‌بندی داده‌ها و استخراج ویژگی‌های سطح بالا و مؤثر در کار دسته‌بندی ارائه شد. این مدل، یک شبکه CNN تک‌لایه برای استخراج ویژگی‌های سطح بالا و یک‌لایه تمام متصل برای انجام دسته‌بندی است. طبقه‌بند استفاده شده در این مدل Softmax بود.

روش پیشنهادی به‌خوبی قادر به رفع چالش‌های ذکر شده در بخش ۱-۱ بود. این روش در مقایسه با روش‌های دیگر، دقت بالاتری را بدست آورد. ازآنجاکه هدف ما مقایسه عملکرد مدل‌های ارائه شده با مرجع

[۵] است. برای حفظ عدالت در مقایسه از پیش پردازش‌ها و دادگان معرفی شده در این مرجع استفاده شده است. نتایج نشان داد که استفاده از لایه جاسازی دقت تشخیص را به صورت قابل توجی در مدل پیشنهادی افزایش داد.

۵-۲ پیشنهادها برای ادامه فعالیت

در این پایان‌نامه، هیچ الگوریتمی جهت انتخاب جهش‌های ژنی و ژن‌های مؤثر در ایجاد سرطان ارائه نشده است. یافتن ژن و جهش‌های مؤثر در ایجاد انواع سرطان می‌تواند کمک بزرگی به دسته‌بندی دقیق‌تر انواع سرطان و سرعت بخشیدن به آموزش مدل‌های عمیق کند. در کارهای آتی می‌توان با استفاده از الگوریتم‌های مختلف انتخاب ویژگی و تئوری‌های نظریه بازی‌ها برای کشف جهش‌های ژنی و ژن‌های مؤثر در ایجاد انواع سرطان دقت را بهبود داد. همچنین به دلیل کمبود داده‌ها در صورت موفقیت در ایجاد الگوریتمی مؤثر در انتخاب ویژگی‌ها می‌توان به‌جای لایه تمام متصل در مدل پیشنهادی از لایه SVM نیز استفاده کرد.

لغت‌نامه

مرتب بر اساس حروف الفبای فارسی

عنوان فارسی	عنوان انگلیسی
ابر پارامتر	Hyperparameter
ابعاد خروجی	Output-Dim
اجرا در هر آزمایش	Executions per trial
اشمیت هوپر	Schmidhuber
اطلس ژنوم سرطان	Cancer Genome Atlas
اعتبار متقابل ده برابر	Ten Fold Cross Validation
الگوریتم‌های انتشار رو به عقب	Back-propagation alghorithms
الیافی	Fibrous
اندازه بخش	Batch Size
اندازه هسته	Kernel Size
انفجار شیب	Exploding Gradient
بردار	Vector
پارافین منجمد و ثابت شده با فرمالین	Formalin-fixed Paraffin Embedded
پدینگ	Padding
پردازش زبان طبیعی	Natural Language Processing
پولینگ حد اکثر	Max Pooling
پولینگ حداکثر سراسری	Global Max Pooling
تابع بهینه سازی	Optimizer Function
تابع خطا	Loss Function
تابع فعال ساز	Activation Function
تانژانت هیپربولیک	Tangent Hyperbolic
تشخیص چهره	Face Detection
تشخیص دست خط	Handwrite Recognition
تشخیص رفتار	Behavior Recognition
تشخیص زودهنگام سرطان	Cancer early diagnosis
تشخیص گفتار	Speech Recognition
تصاویر بافت‌شناسی	Histological Images
تصویربرداری با تشدید مغناطیسی	Magnetic Resonance Imaging

Data Augmentation	تقویت داده‌ها
Masking	تکنیک ماسک کردن
Fully Connected	تمام متصل
Keras Tuner	تنظیم‌کننده کراس
Grid-like topology	توپولوژی‌های شبکه‌ای
Embedding	جاسازی
Word Embedding	جاسازی کلمات
Max Trials	حداکثر آزمایش
Max-Norm	حداکثر هنجار
Dropout	حذف تصادفی
Feed-Forward	حرکت روبه‌جلو
Output Gate	دروازه خروجی
Forget Gate	دروازه فراموشی
Input Gate	دروازه ورودی
Validation Accuracy	دقت اعتبار سنجی
Epoch	دور
Relu	رلو
Invasive Carcinoma	سرطان تهاجمی
Situ Carcinoma	سرطان درجا
Recommender Systems	سیستم‌های پیشنهاددهنده
Sigmoid	سیگموید
Time-delay neural networks	شبکه‌های عصبی با تأخیر در زمان
Deep belief network	شبکه باور عمیق
Dual-Path Network	شبکه دومسیره
Recurrent Neural Network	شبکه عصبی بازگشتی
Stacked Auto Encoder Neural Network	شبکه عصبی رمزگذار انباشته
Ensemble Of Multi Scale Convolutional Neural Networks	شبکه عصبی کانولوشنی چند مقیاسی
Image Classification	طبقه‌بندی تصاویر
Somatic point mutation based cancer classification	طبقه‌بندی سرطان مبتنی بر جهش در نقطه سوماتیک
Input Lenght	طول ورودی

Filter	فیلتر
Clustered Gene Filtering	فیلترکردن ژن‌های خوشه‌بندی‌شده
Stride	قدم
Convolutional	کانالوشنی
Indexed Sparsity Reduction	کاهش پراکندگی فهرست شده
Keras	کراس
Node	گره
Max Pooling Layer	لایه پولینگ حداکثر
Average Pooling Layer	لایه پولینگ میانگین
Flatten	لایه صاف
Dense Layer	لایه متراکم
Dictionary	لغت نامه
Support vector machine	ماشین بردار پشتیبان
Gradient Boosting Machine	ماشین تقویت‌کننده شیب
Validation Set	مجموعه ارزیابی
Vanishing Gradient	ناپدید شدن شیب
Area Under Curve	ناحیه زیر منحنی
Dropout Rate	نرخ حذف تصادفی
Heat-map	نقشه حرارتی
kernel	هسته
Hochreiter	هوخرایتر
Hinton	هینتون
One-Hot	وان – هات
Transfer Learning	یادگیری انتقالی

مرتب بر اساس حروف الفبای انگلیسی

عنوان فارسی	عنوان انگلیسی
تابع فعال ساز	Activation Function
ناحیه زیر منحنی	Area Under Curve
لایه پولینگ میانگین	Average Pooling Layer
الگوریتم‌های انتشار رو به عقب	Back-propagation alghorithms
اندازه بخش	Batch Size
تشخیص رفتار	Behavior Recognition
تشخیص زودهنگام سرطان	Cancer early diagnosis
اطلس ژنوم سرطان	Cancer Genome Atlas
فیلترکردن ژن‌های خوشه‌بندی‌شده	Clustered Gene Filtering
کانالوشنی	Convolutional
تقویت داده‌ها	Data Augmentation
شبکه باور عمیق	Deep belief network
لایه متراکم	Dense Layer
لغت نامه	Dictionary
حذف تصادفی	Dropout
نرخ حذف تصادفی	Dropout Rate
شبکه دومسیره	Dual-Path Network
جاسازی	Embedding
شبکه عصبی کانولوشنی چند مقیاسی	Ensemble Of Multi Scale Convolutional Neural Networks
دور	Epoch
اجرا در هر آزمایش	Executions per trial
انفجار شیب	Exploding Gradient
تشخیص چهره	Face Detection
حرکت روبه‌جلو	Feed-Forward
الیافی	Fibrous
فیلتر	Filter
لایه صاف	Flatten
دروازه فراموشی	Forget Gate

Formalin-fixed Paraffin Embedded	پارافین منجمد و ثابت شده با فرمالین
Fully Connected	تمام متصل
Global Max Pooling	پولینگ حداکثر سراسری
Gradient Boosting Machine	ماشین تقویت‌کننده شیب
Grid-like topology	توپولوژی‌های شبکه‌ای
Handwrite Recognition	تشخیص دست خط
Heat-map	نقشه حرارتی
Hinton	هینتون
Histological Images	تصاویر بافت‌شناسی
Hochreiter	هوخرایتر
Hyperparameter	ابرپارامتر
Image Classification	طبقه‌بندی تصاویر
Indexed Sparsity Reduction	کاهش پراکندگی فهرست شده
Input Gate	دروازه ورودی
Input Lenght	طول ورودی
Invasive Carcinoma	سرطان تهاجمی
Keras	کراس
Keras Tuner	تنظیم‌کننده کراس
kernel	هسته
Kernel Size	اندازه هسته
Loss Function	تابع خطا
Magnetic Resonance Imaging	تصویربرداری با تشدید مغناطیسی
Masking	تکنیک ماسک کردن
Max Pooling	پولینگ حد اکثر
Max Pooling Layer	لایه پولینگ حداکثر
Max Trials	حداکثر آزمایش
Max-Norm	حداکثر هنجار
Natural Language Processing	پردازش زبان طبیعی
Node	گره
One-Hot	وان – هات
Optimizer Function	تابع بهینه سازی

Output Gate	دروازه خروجی
Output-Dim	ابعاد خروجی
Padding	پدینگ
Recommender Systems	سیستم‌های پیشنهاددهنده
Recurrent Neural Network	شبکه عصبی بازگشتی
Relu	رلو
Schmidhuber	اشمیت هوپر
Sigmoid	سیگموید
Situ Carcinoma	سرطان درجا
Somatic point mutation based cancer classification	طبقه‌بندی سرطان مبتنی بر جهش در نقطه سوماتیک
Speech Recognition	تشخیص گفتار
Stacked Auto Encoder Neural Network	شبکه عصبی رمزگذار انباشته
Stride	قدم
Support vector machine	ماشین بردار پشتیبان
Tangent Hyperbolic	تانژانت هیپربولیک
Ten Fold Cross Validation	اعتبار متقابل ده برابر
Time-delay neural networks	شبکه‌های عصبی با تأخیر در زمان
Transfer Learning	یادگیری انتقالی
Validation Accuracy	دقت اعتبار سنجی
Validation Set	مجموعه ارزیابی
Vanishing Gradient	ناپدید شدن شیب
Vector	بردار
Word Embedding	جاسازی کلمات

مراجع

[١] F. Bray, J. Ferlay, I. Soerjomataram, L. Siegel, A. Torre و A. Jemal. (٢٠١٨) "Global cancer statistics: GLOBOCAN estimates of incidence and mortality worldwide for ٣٦ cancers in ١٨٥ countries ",*CA: a cancer journal for clinicians* ,Vol ٦٨, pp ٣٩٤–٤٢٤

[٢] S. Hochreiter, J. Schmidhuber. (١٩٩٧) "Long short-term memory ",*Neural computation*, Vol ٩ , pp ١٧٣٥ –١٧٨٠.

[٣] G. E. Hinton, S. Osindero, Y. Whye Teh. (٢٠٠٦) "A fast learning algorithm for deep belief nets ",*Neural computation* ,Vol١٨ ,pp ١٥٢٧–١٥٥٤.

[٤] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu و F. E. Alsaadi. (٢٠١٧) "A survey of deep neural network architectures and their applications ",*Neurocomputing*,Vol ٢٣٤, pp ١١– ٢٦.

[٥] Y. Yuan, Y. Shi, C. Li, J. Kim, W. Cai, Z. Han و D. D. Feng. (.٢٠١٦) "DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations ",*BMC bioinformatics* ,Vol١٧ ,pp ٢٤٣–٢٥٦.

[٦] G. W. Sledge Jr. (٢٠٠٥), *What is targeted therapy* ,American Society of Clinical Oncology.

[٧] J. Gudeman, M. Jozwiakowski, J. Chollet, M. Randell. (٢٠١٣) "Potential risks of pharmacy compounding ",*Drugs in R&d* ,Vol١٣,pp ١– ٨.

[٨] B. Franken, M. R. De Groot, W. J. B. Mastboom, I. Vermes, J. van der Palen, A. G. Tibbe, L. W. Terstappen.(٢٠١٢) "Circulating tumor cells, disease recurrence and survival in newly diagnosed breast cancer ",*Breast Cancer Research*, Vol١٤ ,pp ١– ٨.

[٩] D. F. Hayes, J. Smerage. (٢٠٠٨), "Is there a role for circulating tumor cells in the management of breast cancer ",*Clinical Cancer Research* ,Vol١٤ ,pp ٣٦٤٦–٣٦٥٠.

[١٠] I. R. Watson, K. Takahashi, P. A. Futreal, L. Chin. (٢٠١٣) "Emerging patterns of somatic mutations in cancer ",*Nature reviews Genetics* ,Vol١٤ ,pp ٧٠٣– ٧١٨.

[١١] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, R. K. Wilson. (٢٠١٢) "VarScan ٢: somatic mutation and copy number alteration discovery in cancer by exome sequencing ",*Genome research*,Vol٢٢ ,pp. ٥٦٨– ٥٧٦.

[١٢] Z. Huang, D. Huang, S. Ni, Z. Peng, W. Sheng ,X. Du. (٢٠١٠) "Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer ", *International journal of cancer* ,Vol١٢٧ ,pp ١١٨–١٢٦.

[١٣] J. Aarøe, T. Lindahl, V. Dumeaux, S. Sæbø, D. Tobin, N. Hagen, P. Skaane, A. Lönneborg, P. Sharma , A. L. Børresen-Dale. (٢٠١٠), "Gene expression profiling of peripheral blood cells for early detection of breast cancer ",*Breast Cancer Research*, Vol١٢ ,pp ١– ١١.

[١٤] Cho, J. Hoon, D. Lee, J. H. Park, I. B. Lee, (٢٠٠٣) "New gene selection method for classification of cancer subtypes considering within-class variation ", Elsevier Vol١, pp.٣-٢٣.

[١٥] Z. Cai, L. Xu, Y. Shi, M. R. Salavatipour, R. Goebel, G. Lin, (٢٠٠٦) "Using gene clustering to identify discriminatory genes with higher classification accuracy ", *Sixth IEEE Symposium on BioInformatics and BioEngineering (BIBE'٠٦)*.

[١٦] J. Wang, J. Lin و Z. Wang, (٢٠١٧) "Efficient hardware architectures for deep convolutional neural network ",*IEEE Transactions on Circuits and Systems I: Regular Papers*,Vol٦٥ ,pp ١٩٤١– ١٩٥٣.

[١٧] L. Gong, C. Wang, X. Li, H. Chen و X. Zhou, (٢٠١٨) "MALOC: A fully pipelined FPGA accelerator for convolutional neural networks with all layers mapped on chip ",*IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*,Vol٣٧ ,pp ٢٦٠١–٢٦١٢.

[١٨] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn و D. Yu, (٢٠١٤) "Convolutional neural networks for speech recognition ",*IEEE/ACM Transactions on audio, speech, and language processing* ,Vol٢٢ ,pp ١٥٣٣–١٥٤٥.

[١٩] S. Varsamopoulos, K. Bertels, C. G. Almudever (٢٠١٨), "Designing neural network based decoders for surface codes ",*IEEE*, Vol٦٩, pp ٣٠٠-٣١١

[٢٠] F. A. Gers, J. Schmidhuber , F. Cummins, (١٩٩٩) “Learning to forget: Continual prediction with LSTM .”,*Nerual Computation*, Vol ١٢, pp ٢٤٥١-٢٤٧١

[٢١] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. C. Shih, J. Tomaszewski, F. A. González و A. Madabhushi, (٢٠١٧) “Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent ”,*Scientific reports* ,Vol٧ ,pp ١–١٤.

[٢٢] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, (٢٠١٦) “Rethinking the inception architecture for computer vision ”,*Proceedings of the IEEE conference on computer vision and pattern recognition*, pp ٢٨١٨--٢٨٢٦

[٢٣] G. Liang, H. Hong, W. Xie و L. Zheng, (٢٠١٨) “Combining convolutional neural network with recursive neural network for blood cell image classification ”,*IEEE Access* ,Vol٦ ,pp ٣٦١٨٨– .٣٦١٩٧.

[٢٤] F. Chollet, (٢٠١٧) “Xception: Deep learning with depthwise separable convolutions در ”,*Proceedings of the IEEE conference on computer vision and pattern recognition*, pp ١٢٥١ -- ١٢٥٨

[٢٥] S. Khan, N. Islam, Z. Jan, I. U. Din, J. J. C. Rodrigues, (٢٠١٩) “A novel deep learning based framework for the detection and classification of breast cancer using transfer learning ”,*Pattern Recognition Letters*,Vol٢٥٦ ,pp ١– .٦.

[٢٦] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, (٢٠١٤) “CNN features off-the-shelf: an astounding baseline for recognition ”,*Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp ٨٠٦--٨١٣

[٢٧] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng و T. Darrell,(٢٠١٤) “Decaf: A deep convolutional activation feature for generic visual recognition ”, *International conference on machine learning*, pp ٦٤٧-٦٥٥

[٢٨] K. Simonyan و A. Zisserman, (٢٠١٤) “Very deep convolutional networks for large-scale image recognition ”,*arXiv preprint arXiv:١٤٠٩,١٥٥٦*

[٢٩] K. He, X. Zhang, S. Ren, J. Sun, (٢٠١٦) “Deep residual learning for image recognition ”,*Proceedings of the IEEE conference on computer vision and pattern recognition*, pp ٧٧٠-٧٧٨

[٣٠] F. A. Spanhol, L. S. Oliveira, C. Petitjean, L. Heutte, (٢٠١٥) “A dataset for breast cancer histopathological image classification ”,*Ieee transactions on biomedical engineering*,Vol٦٣ ,pp ١٤٥٥– ١٤٦٢.

[٣١] W. Zhu, C. Liu, W. Fan, X. Xie, (٢٠١٨) "Deeplung: Deep ٣d dual path nets for automated pulmonary nodule detection and classification ",*IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp ٦٧٣--٦٨١

[٣٢] F. Liao, M. Liang, Z. Li, X. Hu, S. Song, (٢٠١٩) "Evaluate the malignancy of pulmonary nodules using the ٣-d deep leaky noisy-or network ",*IEEE transactions on neural networks and learning systems*,Vol٣٠ ,pp ٣٤٨٤– ٣٤٩٥.

[٣٣] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, (٢٠١٧), "Speed/accuracy trade-offs for modern convolutional object detectors ", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp ٧٣١٠--٧٣١١

[٣٤] J. H. Friedman, (٢٠٠١) "Greedy function approximation: a gradient boosting machine ",*Annals of statistics* ,pp ١١٨٩–١٢٣٢, .

[٣٥] Y. Guo, X. Shang, Z. Li, (٢٠١٩) "Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer ", *Neurocomputing*,Vol٣٢٤ ,pp ٢٠–٣٠.

[٣٦] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, L. Bottou,(٢٠١٠) "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion ",*Journal of machine learning research*,Vol١١ .,

[٣٧] A. Ng, (٢٠١١), "Sparse autoencoder ",*CS٢٩٤A Lecture notes* ,Vol٢٧ ,pp ١–١٩.

[٣٨] A. D. Jia, B. Z. Li و C. C. Zhang, (٢٠٢٠) "Detection of cervical cancer cells based on strong feature CNN-SVM network ",*Neurocomputing* ,Vol٤١١ ,pp ١١٢–١٢٧.

[٣٩] Z. Yang, L. Ran, S. Zhang, Y. Xia, Y. Zhang, (٢٠١٩) "EMS-Net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images ",*Neurocomputing* ,Vol٣٦٦ ,pp ٤٦–٥٣.

[٤٠] J. Xu, X. Luo, G. Wang, H. Gilmore, A. Madabhushi, (٢٠١٦) "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images ",*Neurocomputing*,Vol١٩١ ,pp ٢١٤–٢٢٣.

[٤١] T. Qaiser, Y.-W. Tsang, D. Epstein, N. Rajpoot, (٢٠١٧) "Tumor segmentation in whole slide images using persistent homology and deep convolutional features ", *Annual Conference on Medical Image Understanding and Analysis*, pp ٣٢٠--٣٢٩

[٤٢] D. Shen, G. Wu, H. I. Suk, (٢٠١٧) "Deep learning in medical image analysis ", *Annual review of biomedical engineering* ,Vol ١٩ ,pp ٢٢١–٢٤٨.

[۴۳] J. Z. Cheng, D. Ni, Y. H. Chou, J. Qin, C. M. Tiu, Y. C. Chang, C. S. Huang, D. Shen, C. M. Chen, (۲۰۱۶) "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans",*Scientific reports* ,Vol٦ ,pp ١–١٣.

[۴۴] W. Bulten, C. A. Hulsbergen-van de Kaa, J. van der Laak, G. J. S. Litjens (*۲۰۱۸*), "Automated segmentation of epithelial tissue in prostatectomy slides using deep learning",*Medical Imaging Digital Pathology,* Vol ١٠٥٨١, pp ١٠٥٨١٠

[۴۵] B. Wang, A. Wang, F. Chen, Y. Wang, C. C. Jay-Kuo, (۲۰۱۹) "Evaluating word embedding models: Methods and experimental results",*APSIPA transactions on signal and information processing* Cambridge University Pres, Vol٨.

[۴۶] Y. Zi, Y. Shen, (۲۰۱۸) "On the dimensionality of word embedding",*arXiv preprint arXiv: ١٨١٢,٠۴٢٢۴.*

[۴۷] E. Kawin, D. Duvenaud, G. Hirst, (۲۰۲۱) "Understanding undesirable word embedding associations",*arXiv preprint arXiv:١٩٠٨,٠۶٣۶١*

[۴۸] N. Swinger, M. De-Arteaga, N. T. Heffernan, M. D. Leiserson , (۲۰۱۹) "What are the biases in my word embedding",*Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* ,pp ٣١١--٣٠٥.

[۴۹] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, L. Carin, (۲۰۱۸) "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms",*arXiv preprint arXiv: ١٨٠٥,٠٩٨۴٣*

[۵۰] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, (۲۰۱۷) "Deep learning with word embeddings improves biomedical named entity recognition", *Bioinformatics* ,Vol٣٣ ,pp ٣٧– ۴٨.

[۵۱] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, (١٩٩٨) "Gradient-based learning applied to document recognition",*Proceedings of the IEEE* ,Vol٨٦ ,pp ٢٢٧٨– ٢٣٢۴.

[۵۲] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov,(٢٠١٢) "Improving neural networks by preventing co-adaptation of feature detectors",*arXiv preprint arXiv: ١٢٠٧,٠۵٨٠*

[۵۳] K. Tomczak, P. Czerwińska, M. Wiznerowicz, (۲۰۱۵) "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge",*Contemporary oncology* ,Vol١٩ , pp ۶٨.

[۵۴] J. Bergstra, Y. Bengio (۲۰۱۲) “Random search for hyper-parameter optimization ”,. *Journal of machine learning research* ,pp ۲.

[۵۵] C. Cortes, V. Vapnik, (۱۹۹۵) “Support-vector networks ”,*Machine learning* ,Vol ۲۰p. ۲۷۳–۲۹۷.

Abstract

Prompt and accurate diagnosis of cancer or its subtypes in patients has a significant impact on the correct treatment process and reduction of treatment costs. There are a variety of methods for diagnosing cancer and its subtypes that have evolved with recent advances in machine learning and deep learning. Also, new methods have been introduced using machine learning and deep learning that have good results in predicting different types of cancer or its subgroups and diagnosing benign and malignant cancerous tumors in patients. One of the methods that is considered today for the diagnosis of cancer and its subtypes is the classification of cancer based on somatic mutations. Exposure to UV rays or certain chemicals can cause mutations in the body's cells. These abnormal mutations caused by environmental factors are called somatic mutations. However, the classification of somatic mutation-based cancers is challenging. Challenges such as low sample data volume, high data scatter, overfitting, and the use of simple linear classifiers are factors that prevent increased classification performance. This paper presents ways to solve these challenges. These methods include clustering gene filter preprocessing, indexed scatter reduction, regulatory methods, the Global-Max-Pooling layer, and the use of the embedding layer. Also in this dissertation, three deep learning models CNN, LSTM and a combination of these two models are tested on the TCGA-DeepGene data set. Our proposed model is a single-layer CNN model with an embedding layer. This model achieved ٦٦,٤٥٪ accuracy. Compared to the reference cited in this dissertation, the accuracy has increased by ١,٤٥٪.

Keywords: Cancer classification, De novo mutations, Deep neural network, CNN, LSTM

Shahrood University of Thchnology

Faculty of Computer Engineering and Information Technology

M.Sc. Thesis in Artificial Intelligence Engineering

A comparison of deep neural network models for cancer subtyping Based on somatic point mutations

By: Pouria Parhami

Supervisors:

Dr. Mansoor Fateh

Advisors:

Dr. Mohsen Rezvani

Dr. Hamid Alinejad Rokny

September ۲۰۲۱