
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی و رباتیک

تشخیص سرطان مبتنی بر یادگیری عمیق و تنوع تعداد کپی (CNV)

نگارنده: سعید حسنی برزادران

اساتید راهنما

دکتر محسن رضوانی و دکتر علی اکبر پویان

اساتید مشاور

دکتر حمید علی‌نژاد رکنی و دکتر منصور فاتح

بهمن ۱۳۹۷

تقدیم به

پدر و مادر عزیز و مهربانم

که در سختی‌ها و دشواری‌های زندگی، همواره یاوری دلسوز و فداکار و پشتیبانی محکم و مطمئن برایم بوده‌اند.

تشکر و قدردانی

از اساتید گرانقدر جناب آقای دکتر محسن رضوانی و جناب آقای دکتر علی اکبر پویان که راهنمایی این پایانامه را به عهده داشته‌اند و آقای دکتر حمید علی‌نژاد رکنی و آقای دکتر منصور فاتح که اساتید مشاور من در این راه بودند کمال تشکر و قدردانی را دارم.

همچنین بر خود لازم میدانم از آقای دکتر غلامرضا محتشمی برزادران به دلیل یاری ها و راهنمایی های بی چشم داشت ایشان در تمام مقاطع تحصیلی که بسیاری از سختی ها را برایم آسان نمودند تشکر نمایم .

تعهدنامه

اینجانب **سعید حسنی برزادران** دانشجوی دوره کارشناسی ارشد رشته **مهندسی کامپیوتر- هوش مصنوعی و رباتیک** دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **تشخیص سرطان مبتنی بر یادگیری عمیق و تنوع تعداد کپی (CNV)** تحت راهنمائی **جناب آقای دکتر محسن رضوانی و جناب آقای دکتر علی اکبر پویان** 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 بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- 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

سرطان به عنوان یکی از شایع‌ترین بیماری‌های چند دهه اخیر، توجه بسیاری از پژوهشگران را در حوزه‌های مختلف علوم به خود جلب نموده است. درمان این بیماری با استفاده از روش‌های درمانی رایج، اغلب یا هزینه‌بر است یا با شکست مواجه شده یا بیمار را متحمل عوارض جانبی شدید می‌کند. بدین سبب نیاز به توسعه روش‌های جدید درمانی، به خوبی احساس می‌شود.

تغییرات ژنومی در DNA باعث ایجاد انواع مختلف سرطان در انسان می‌گردد. تنوع تعداد کپی یا CNV، به عنوان یکی از انواع مختلف جهش در DNA، سبب بروز سرطان‌های مختلفی در انسان شده است. به منظور درک تفاوت بین سرطان‌های مختلف با کمک CNVها، در این مطالعه، به طبقه‌بندی شش کلاس مختلف سرطان در انسان با استفاده از مقادیر سطح CNV و با کمک یادگیری عمیق می‌پردازیم. در سال‌های اخیر از یادگیری عمیق در تشخیص انواع سرطان مانند سرطان ریه، پوست و سینه استفاده شده است. بسیاری از این روش‌ها از شبکه‌های کانولوشن برای تشخیص سرطان استفاده کرده‌اند. ما در این تحقیق از شبکه عمیق LSTM برای طبقه‌بندی استفاده می‌کنیم. برای این منظور اطلاعات CNV از ۲۴۱۷۴ ژن به عنوان ویژگی‌ها برای ساخت طبقه‌بند استفاده گردید.

نتایج آزمایش‌ها دقت ۹۲ درصدی در طبقه‌بندی را نشان می‌دهد. سپس برای تحلیل بایولوژی به شناسایی ژن‌هایی که بیشترین تاثیر را در ایجاد سرطان‌ها داشتند پرداخته شد. بدین سبب از معیار بیشترین وابستگی-کمترین افزونگی جهت شناسایی ژن‌های موثر استفاده گردید. با کمک این الگوریتم تعداد ۲۰۰ ژن برتر را شناسایی کردیم و سپس طبقه‌بندی را با استفاده از این ۲۰۰ ژن انجام دادیم. دقت بیش از ۸۱ درصدی در طبقه‌بندی نشان از اهمیت این ژن‌ها را می‌داد. در نهایت ۱۰ ژن برتر که

بیشترین تاثیر را در طبقه‌بندی داشتند شناسایی کردیم. همچنین عملکرد ژنتیکی این ژن‌ها را مورد بررسی قرار دادیم تا نتایج آماری و تجزیه و تحلیل بیولوژیکی از این ژن‌ها کمک فراوانی به شناخت انواع سرطان‌های مختلف انسان و پیشنهادهایی را برای ارائه درمان‌های مناسب بیماران ارائه دهد.

کلمات کلیدی: طبقه‌بندی سرطان، پیش‌بینی سرطان، یادگیری عمیق، تنوع تعداد کپی، شبکه عمیق LSTM

فهرست مطالب

فهرست اشکال

فهرست جداول

فهرست اختصارات

A ……...……..…… Adenine

AUC ……...……..……..…… Area Under Curve

BRCA ……...…….…….. Breast invasive carcinoma

C ……...……..…….…… Cytosine

CNV ……...…….. Copy Number Variation

COAD ……...……...…….…… Colon adenocarcinoma

DNA ……...…….. Deoxyribonucleic Acid

G ……...……...…...……..…… Guanine

GBM ……...…….…….. Glioblastoma multiforme

KIRC ……..…….. Kidney renal clear cell carcinoma

LSTM ……..……...……. Long Short Term Memory

MI ……...…...……...……...…...…… Mutual Information

MLP ……..……..…… Multilayer Perceptron

mRMR ……...…… maximum Relevance Minumum Redundancy

OV ……...……...…….. Ovarian serous cystadenocarcinoma

ROC ……..…… Receiver Operating Characteristic

T ……...……...…… Thymine

فصل۱: مقدمه

۱-۱ مقدمه

علم ژنتیک، نیاز به تحقیق، مطالعات و آزمایشات فراوان دارد که بشر بتواند به آن‌ها دست یابد و البته مسلط بودن بر این علم، نیاز به پشتوانه قوی علومی همچون بیولوژیِ سلولی ملکولی، بیوشیمی، فیزیولوژی و آمار و احتمالات دارد. با کمک این علم تمام بیماری‌های ژنتیکی حتی در دوره جنینی نیز قابل درمان خواهند بود. به کمک این علم از جهش‌های متوالی عوامل بیماری‌زا که عامل اصلی فناناپذیر بودن آن‌ها است، جلوگیری به عمل می‌آید و درصد بالایی از بیماری‌های شناخته شده ریشه‌کن خواهد شد. اصلاح نژادی حیوانات و نباتات که باعث بالا رفتن سطح کیفیت و کمیت فرآورده‌های غذایی از آنان شده است با کمک این علم امکان‌پذیر است. تهیه داروها و هورمون‌ها با درجه خلوص بالا و صرف هزینه‌های پایین و درمان بیماری‌های ژنتیکی با ایجاد تغییرات در سلول از فواید دیگر این علم محسوب می‌شود.

تا قبل از سال ۲۰۰۰ تصور می‌شد از هر ژن در موجودات سالم فقط دو نسخه وجود دارد که یک نسخه از پدر و نسخه دیگر از مادر به ارث می‌رسد [۱]. اما کشفیات اخیر نشان می‌دهند که قطعات بزرگی از DNA که اندازه آن‌ها از هزاران تا میلیون‌ها باز نوکلئوتیدی متغیر است می‌توانند از نظر تعداد کپی در ژنوم متغیر باشند. یعنی این‌که تعداد یک، دو، سه یا تعداد بیشتر کپی از یک ژن در DNA موجود است، که به آن‌ها تنوع تعداد کپی یا Copy Number Variation یا CNV[۱] می‌گویند [۲]. علاوه بر این، تعدادی از CNVها ممکن است نقش مهمی در ابتلا به بیماری‌های خاص مانند سرطان را بازی کند [۳]. CNV یک زمینه جدید در علم ژنتیک است که به بررسی پدیده‌ای می‌پردازد که در طی آن قسمتی از ژنوم فرد در طول DNA تکرار می‌شود و تعداد این تکرار برای اشخاص متفاوت با یکدیگر

[۱] به دلیل این که در زبان فارسی ترجمه مشخصی برای Copy Number Variation یا به اختصار CNV وجود نداشت، در این تحقیق از کلمه CNV در تمام تحقیق استفاده شد که منظور همان Copy Number Variation است.

متفاوت است [۴]. در واقع CNV یک نوع تنوع ساختاری در ژنوم موجودات محسوب می‌شود و به ویژه یک نوع مضاعف شدن[۲] و حذف[۳] در ژنوم است که تعداد زیادی از نوکلئوتیدها را درگیر می‌کند [۵]. مطالعات اخیر نشان داده است که تقریباً دو سوم کل ژنوم انسان از CNVها تشکیل شده است [۳]. در پستانداران، CNV نقش مهمی را در ایجاد تنوع خصوصیات یک جمعیت از جمله فنوتایپ‌های[۴] یک بیماری ایفا می‌کند [۶]. در این بحث، سعی داریم طبقه‌بندی طراحی کنیم که با کمک آن بتوان داده‌های CNV از شش سرطان مختلف طبقه‌بندی کرد. ما این طبقه‌بندی را با الگوریتم‌های پیشرفته در هوش مصنوعییعنی تکنیک یادگیری عمیق و با کمک شبکه LSTM انجام می‌دهیم.

یادگیری عمیق یک اصطلاح جدید در شبکه‌های عصبی چند لایه و رویکردی بسیار قدرتمند برای یادگیری الگوهای پیچیده در چندین لایه است. این یادگیری از لایه‌هایی از نرون‌ها که اطلاعات را در مغز انسان دریافت و ترکیب می‌کنند، الهام گرفته است [۷].

مدل‌های عمیق آموزش، چالش‌های بسیار بیشتری را نسبت به آموزش مدل‌های کم عمق، برای تعیین پارامترهای مدل و سازه‌های مدل ارائه می‌دهند. الگوریتم‌های انتشار به عقب[۵] که برای شبکه‌های عصبی معمولی مورد استفاده قرار می‌گیرند، تفاوت بین خروجی‌های مشاهده شده و پیش‌بینی شده را برای تنظیم پارامترها پخش می‌کنند که در آن‌ها می‌توان عمق و عرض مدل را به صورت خودکار تنظیم کرد، در نتیجه پیچیدگی مدل کاهش می‌یابد. مدل‌های یادگیری عمیق سبب شد تا زمینه هوش مصنوعی به طور کلی متحول گردد. به طور تقریبا غیرهوشمند، هوش مصنوعی کلاسیک برای وظایفی که برای انسان مشکل است (مثلا معادلات، شطرنج) بسیار موفقیت‌آمیز بود، اما وظیفه به ظاهر ساده

---

[۲] Amplification

[۳] Deletion

[۴] به خصوصیات قابل مشاهده یا صفت یک جاندار فنوتایپ یا Phenotype گفته می‌شود.

[۵] Back-Propagation

انسان به عنوان مثال، درک صحنه، بسیار مشکل‌تر بود. یادگیری عمیق منجر به دستیابی به پیشرفت‌های فراوانی در زمینه‌های تشخیص گفتار، پردازش متن، تولید خوکار ترجمه‌های نسبتا دقیق و بسیاری از زمینه‌های دیگر شد [۸].

۲-۱ اهمیت و ضرورت انجام تحقیق

سرطان به عنوان یک بیماری شناخته شده و مرگ‌آور شناخته می‌شود که با توجه به آمار سالیانه، حدود ۸ میلیون نفر یعنی حدود ۱۴٫۶٪ از مرگ‌ها، در اثر سرطان بوده است[۹]. به تازگی تنوع تعداد کپی یا CNV با توجه به ارتباط آن‌ها با بیماری‌های پیچیده مثل سرطان توجه زیادی را به خود جلب کرده است و پیدا کردن ارتباط آن‌ها به با سرطان یک تحقیق علمی پویا برای محققان و پژوهشگران تبدیل شده است [۹-۱۳]. اما تاکنون طبقه‌بندی داده‌های ژن برای دسته‌بندی در یک کلاس سرطان با کمک یادگیری عمیق مورد توجه قرار نگرفته است.

سوالی که مطرح می‌شود این است که چرا از یادگیری عمیق استفاده کنیم و چرا این طبقه‌بندی‌ها را با الگوریتم‌های ساده‌تر، مثل SVM[۶] یا KNN[۷] یا BN[۸] انجام ندهیم. پاسخ این است که اولا در طبقه‌بندی داده‌ها با تعداد وسیعی ژن (به عنوان ویژگی) روبرو هستیم. دوم این‌که، ژن‌های متفاوت از یک نوع سرطان خاص ممکن است به یکدیگر وابستگی داشته باشند. بنابراین یک طبقه‌بند خطی مثل SVM نمی‌تواند به خوبی این طبقه‌بندی را انجام دهد [۱۴]. از طرفی تکنیک یادگیری عمیق نه تنها برای پردازش روی داده‌های بزرگ مناسب است، بلکه قدرت بالایی در استخراج ویژگی نیز دارد.

---

[۶] Support Vector Machine

[۷] K-Nearest Neigbor

[۸] Naïve Bayes

۱-۳ جنبه جدید بودن و نوآوری در تحقیق

تاکنون کارهای زیادی در زمینه یادگیری عمیق انجام شده است. اما برای طبقه‌بندی سرطان‌ها با کمک یادگیری عمیق و با استفاده از شبکه‌های عصبی بازگشتی LSTM و بر اساس داده‌های CNV تحقیقی به عمل نیامده است. برای اولین بار ما از شبکه عمیق LSTM برای این طبقه‌بندی استفاده کردیم و توانستیم ژن‌های مهم در ایجاد شش نوع سرطان مختلف را شناسایی کنیم. به‌طور خلاصه نوآوری‌های زیر در این تحقیق ارائه شده است:

۱- طبقه‌بندی سرطان بر اساس داده‌های CNV بیماران با کمک شبکه عمیق LSTM

۲- شناسایی ژن‌های مهم در ایجاد سرطان

۱-۴ اهداف تحقیق

اهداف این تحقیق را می‌توان به دو بخش اهداف مستقیم و اهداف غیر مستقیم طبقه‌بندی کرد. از جمله اهداف مستقیم در این تحقیق می‌توان به موارد زیر اشاره نمود:

۱- بررسی امکان استفاده از یادگیری عمیق برای بهبود دقت طبقه‌بندی داده‌های CNV

۲- مقایسه میزان دقت شبکه MLP و شبکه LSTM در طبقه‌بندی دادگان

۳- استفاده از الگوریتم SBS برای شناسایی عامل‌های مهم در طبقه‌بندی

در نهایت انجام این تحقیق می‌تواند منجر به تحقق اهداف غیرمستقیم زیر شود:

۱- کمک به پیشگیری از سرطان

۲- کاهش میزان بروز و شیوع سرطان

۳- ارائه راه‌حل‌های تئوری برای شناسایی ژن‌های مهم در ایجاد سرطان

۴- درمان مناسب بیماران سرطانی با توجه به شناسایی ژن‌های مهم در ایجاد سرطان

۵-۱ روش تحقیق

در روش پیشنهادی در این مقاله، با کمک یادگیری عمیق و همچنین داده‌های CNV در نمونه‌های ژن، به دسته‌بندی کلاس‌های سرطان پرداخته می‌شود. برای این‌کار از دادگان cBioPortal [۱۵] استفاده کردیم که جزئیات آن را در فصل چهارم به تفصیل بیان می‌کنیم. این طبقه‌بندی را با کمک شبکه MLP و شبکه عمیق LSTM انجام دادیم. سپس به مقایسه این دو روش پرداختیم. در نهایت برای استخراج ویژگی و این‌که کدام ژن‌ها بیشترین تاثیر را در ایجاد سرطان‌ها داشته‌اند، از الگوریتم SBS[۹] [۱۶] استفاده کردیم. الگوریتم SBS کارش را با مجموعه‌ای شامل تمام ویژگی‌ها شروع می‌کند و در هر بار تکرار الگوریتم، ویژگی که بوسیله تابع ارزیابی انتخاب می‌شود، را از مجموعه مورد نظر حذف می‌کند. این کار را تا زمانی ادامه می‌دهد که تعداد ویژگی‌ها برابر یک تعداد معینی شود. برای این‌که از درستی نتایج حاصل از آزمایش‌ها اطمینان حاصل کنیم به مطالعه هر کدام از ویژگی‌های استخراج شده در منابع معتبر پرداختیم و تاثیر هر کدام از آن‌ها را در سرطان‌های مختلف مورد بررسی و تحلیل قرار دادیم. شکل ۱-۱ فرآیند انجام این تحقیق شامل مراحل اصلی اجرای تحقیق را نمایش می‌دهد.

شکل ۱-۱: فرایند انجام تحقیق

[۹] Sequentional Backward Selection

۱-۶ ساختار پایان‌نامه

در فصل دوم به بیان مبانی تحقیق و اصطلاحات به کار رفته در این تحقیق پرداخته می‌شود. در فصل سوم از پایان‌نامه به مرور مطالعات گذشته در حوزه سرطان و یادگیری عمیق می‌پردازیم. عملیات بکارگرفته شده در این تحقیق و روند کلی روش ارائه شده به تفصیل در فصل چهارم توضیح داده می‌شود. همچنین معرفی دادگان مورد استفاده و جزئیات الگوریتم ارائه شده را در این فصل به طور کامل شرح می‌دهیم. در نهایت به جمع‌بندی کلی و تحلیل نتایج حاصل می‌پردازیم، مشکلات موجود را بررسی کرده و راه‌کارهای جدیدی ارائه می‌دهیم.

فصل ۲: مبانی تحقیق

۲-۱ مقدمه

در این فصل از تحقیق با مبانی و اصطلاحات به کار رفته آشنا می‌شویم. برای این‌که در ابتدا با مفهوم CNV آشنا شویم و ارتباط آن را با سرطان درک کنیم، در ادامه ساختمان DNA[۱۰] را به زبان ساده شرح داده‌ایم. ابتدا تاریخچه مختصری از DNA را بیان می‌کنیم و به بررسی ساختار DNA می‌پردازیم. در ادامه به بررسی عمکرد DNA در بدن پرداخته و نحوه ایجاد جهش در ژن‌ها را توضیح می‌دهیم و به تاثیر آنها در ایجاد سرطان‌ها خواهیم پرداخت. بعد از آشنایی با نحوه چگونگی ایجاد CNVها به معرفی شبکه‌های عصبی می‌پردازیم. در ادامه با ساختمان شبکه‌های عصبی آشنا می‌شویم و علت معرفی شبکه‌های عمیق LSTM را مورد بررسی قرار می‌دهیم.

۲-۲ معرفی DNA

بدن انسان، بیش از ۳۷ تریلیون سلول دارد، در هسته هر یک از این سلول‌ها مانند سلول پوست، مو، ناخن، دندان و غیره، ۴۶ کروموزوم (۲۳ جفت) وجود دارد [۱۷]. هر انسانی نیمی از کروموزوم‌های خود را از مادر و نیمی دیگر را از پدر دریافت می‌کند. ترکیب چیدمان DNA موجود در کروموزوم‌ها، منحصر به فرد است و فقط در دوقلوهای همسان کاملاً شبیه به هم است [۱۸].

۲-۱-۱ تاریخچه کشف DNA

دانشمندان زیادی در مورد DNA مطالعه و تحقیق کرده اند. فعالیت‌ها، آزمایشات و یافته‌های اسوالد ایوری[۱۱] و همکارانش در دانشگاه راکفلر نشان می‌داد که DNA، همان مولکولی است که حاوی اطلاعات

[۱۰] Deoxyribonucleic acid

[۱۱] Osward Avery

سلول است [۱۹]. این نقش اطلاعاتی DNA چند سال بعد و در سال ۱۹۵۲ توسط آلفرد هرشی[۱۲] و مارتا چیس[۱۳] مورد بررسی قرار گرفت. این دو دانشمند، با استفاده از باکتریوفاژها (ویروس هایی که از باکتری ها به عنوان سلول میزبان خود استفاده می‌کنند)، اثبات کردند که DNA نقش محوری نسبت به پروتئین‌ها دارد(در بخش ساختار DNA با پروتئین‌ها آشنا می‌شویم). اگرچه دانشمندان به این مسئله سال ها پیش پی برده بودند اما هیچ‌کس نمی‌دانست که این اطلاعات چگونه کدگذاری شده‌اند و چه عملکردی دارند. در نهایت کشف DNA در ساختار سلولی در سال ۱۹۵۳ و توسط جیمز واتسون[۱۴] و فرانسیس کریک[۱۵] صورت گرفت. واتسون و کریک از تکنیک مدل‌سازی مولکولی و اطلاعات دیگر دانشمندان (ماوریس ویلکینز[۱۶]، رزالیند فرانکلین[۱۷]، اروین چارگف[۱۸] و لینوس پائولینگ[۱۹]) بهره بردند تا از اسرار DNA پرده بردارند. واتسون، کریک و ویلکینز به سبب کشف DNA، جایزه نوبل پزشکی را دریافت کردند [۲۰].

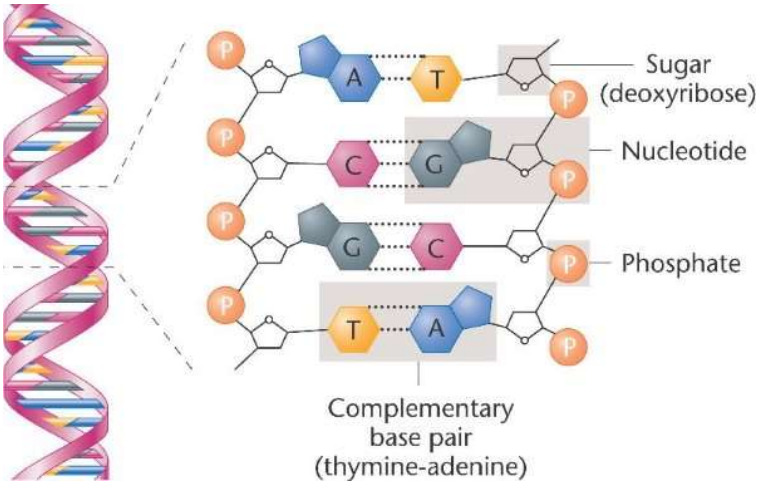
۲-۱-۲ ساختار DNA

برای اینکه ساختار DNA را بشناسیم در ابتدا باید اتم‌ها را شناخته باشیم. ساده‌ترین شکل ماده در طبیعت، اتم‌ها هستند. اتم‌ها با حالت‌ها و تعداد مختلف ترکیب می‌شوند و مولکول‌ها را تشکیل می‌دهند. به طور مثال دو اتم هیدروژن با یک اتم اکسیژن ترکیب می‌شوند و مولکول آب را تشکیل می‌دهند.

---

[۱۲] Alfred Hershey
[۱۳] Martha Chase
[۱۴] James Watson
[۱۵] Francis Crick
[۱۶] Maurice Wilkins
[۱۷] Rosalind Franklin
[۱۸] Erwin Chargaff
[۱۹] Linus Carl Pauling

نوع خاصی از مولکول‌ها وجود دارند که به آنها آمینواسید[۲۰] گفته می‌شود، آمینواسیدها، با نسبت‌ها و ترتیب‌های مختلف با هم ترکیب می‌شوند و پروتئین را به وجود می‌آورند. در بدن انسان، بیش از ۲۰ نوع آمینواسید شناخته شده است و ترکیبات مختلف این آمینواسیدها، موجب شکل‌گیری هزاران نوع پروتئین می‌شود. پروتئین‌ها با مولکول‌های دیگر ترکیب می‌شوند و سلول‌ها را تشکیل می‌دهند. سلول‌ها در کنار هم قرار می‌گیرند و بافت‌ها را می‌سازند. انواع مختلف سلول‌ها، اشکال مختلفی از بافت‌ها مانند بافت پوست، بافت عضله، بافت استخوان و غیره را ایجاد می‌کنند. بنابراین بدن کلیه موجودات زنده از



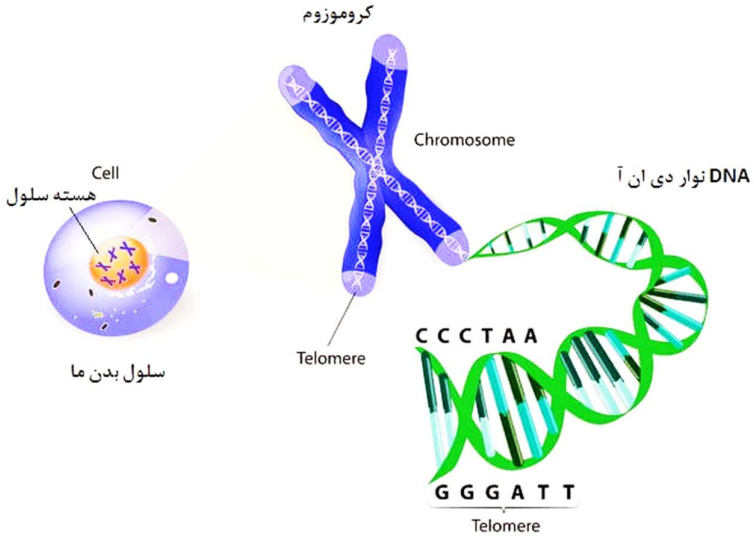
شکل ۲-۱: ساختمان DNA [۲۱]

سلول تشکیل شده است. پیکر تمام گیاهان، بدن همه جانوران و نیز بدن ما انسان‌ها از سلول تشکیل شده است. شکل ۲-۱ ساختمان DNA را نشان می‌دهد [۲۱].

ساختمان DNA از دو نوار طولانی (از جنس فسفات و شکر) تشکیل شده است و روی این نوارها

[۲۰] Amino acid

چهار نوع مولکول قرار دارد، که آدنین[۲۱]، تیامین[۲۲]، سیتوزین[۲۳] ، گوانین[۲۴] نامیده می‌شوند. مولکول‌های



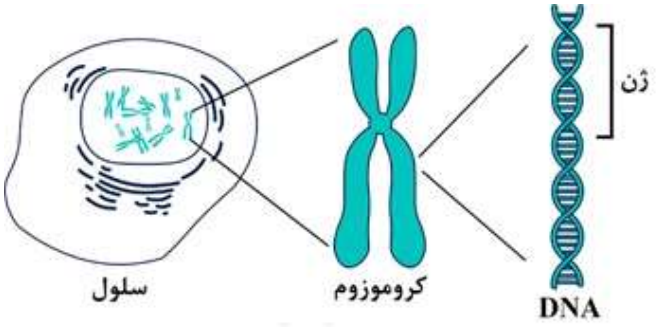
شکل ۲-۲: جایگاه DNA در بدن انسان [۲۱]

آدنین(A) فقط با مولکول های تیامین(T) ارتباط دارند و مولکول‌های سیتوزین(C) فقط با مولکول‌های گوانین(G) ارتباط دارند. جایگاه DNA در بدن ما مطابق شکل ۲-۲ است [۲۱].

طول نوار DNA می‌تواند بیش از چند میلیون مولکول باشد. میلیاردها حالت برای نحوه چیدمان مولکول‌ها در این نوار (DNA) وجود دارد. تنوع انواع چیدمان در مولکول DNA، باعث کثرت تنوع خلق و خو، اندام، رنگ پوست و مو، قد و غیره در انسان‌ها و سایر جانداران می‌گردد. همچنین مولکول DNA، به پروتئین می‌چسبد و کروموزم را تشکیل می‌دهد، کروموزوم شکلی شبیه شکل ۲-۳ دارد [۲۲].

---

[۲۱] Adenine
[۲۲] Thymine
[۲۳] Cytosine
[۲۴] Guanine



شکل ۲-۳: شکل و محل قرار گرفتن کروموزوم در سلول [۲۲]

این‌که افراد نسبت به داروها و محرک‌ها واکنش‌های متفاوتی نشان می‌دهند، به دلیل تفاوت چیدمان مولکول DNA در هسته سلول بدن آن‌ها است. چیدمان مولکول DNA در هسته سلول بدن دو انسان تا حدود ۹۰ درصد شبیه به هم است، ساختار و وظیفه سلول‌ها تفاوت عمده‌ای ندارند، بلکه رفتار سلول‌ها در شرایط مختلف تغییر می‌کند [۲۲].

۲-۱-۳ عملکرد DNA

برای این‌که مشخص شود هر سلول چه وظیفه‌ای را انجام داده و چه رفتاری از خود نشان دهد، توسط یک کد ژنتیکی مشخص می‌شود. هر سلول با تولید مثل، کد DNA را به سلول جدید منتقل می‌کند و به این ترتیب، سلول جدید دستورالعمل و وظایف خود را می‌داند و با سایر سلول‌ها همکاری می‌کند. بنابراین فاکتورهای اصلی که موجودات زنده از نسل‌های قبل از خود به ارث می‌برند، را باید در ساختار DNA آن‌ها، به عنوان ماده اصلی ژنتیک جستجو کرد.

یکی از دستاوردهای کشف رمز سلول جانداران را می‌توان انجام آزمایش و ریشه‌یابی و درک علیت بسیاری از بیماری‌ها، مانند سرطان نام برد و با کمک آن به دنبال راه‌حلی برای درمان آن بود.

۲-۱-۴ جهش

جهش یک تغییر ژنتیکی است که صفات زیستی بعضی از افراد یک گونه را تغییر می‌دهد. به عبارت

دقیق‌تر، جهش‌ها تغییرات در توالی DNA می‌باشند. جهش‌ها می‌توانند در هر ناحیه‌ای از DNA رخ دهند. در هر یک از فعالیت‌های سلولی نظیر فرایندهای همانندسازی[۲۵]، رونویسی[۲۶]، ترجمه[۲۷]، ترکیب مجدد یا نوترکیبی کروموزوم‌ها و بروز و ظهور اطلاعات ژنتیکی احتمال خطا و اشتباه وجود دارد. در موارد نادر ممکن است تغییر خود بخودی در قسمتی از DNA رخ دهد [۲۳]. این تغییر که جهش نامیده می‌شود، ممکن است تغییر در رمز ایجاد نموده و به تولید یک پروتئین ناقص منجر شود. گاهی نتیجه خالص به صورت تغییری در ظاهر فرد یا تغییری در یک شاخص قابل اندازه‌گیری موجود زنده، به نام ویژگی یا صفت مشاهده می‌گردد. طی فرایند جهش، یک ژن ممکن است به دو یا چند شکل متفاوت به نام آلل[۲۸] تغییر یابد [۲۴]. شکل ۲-۴ مثالی از جهش در توالی از یک ژن را نشان می‌دهد.

شکل ۲-۴: نمونه‌ای از جهش در توالی از ژن

۲-۱-۵ انواع مختلف جهش

می‌توان جهش‌ها را بسته به این که چه تاثیری بر فنوتیپ موجود زنده وارد می‌کنند، به سه دسته تقسیم نمود [۲۵]:

۱- جهش‌های مضر: به جهش‌هایی گفته می‌شود که ویژگی‌هایی از فرد را کاهش می‌دهد. جهش‌های

---

[۲۵] Replication
[۲۶] Transcription
[۲۷] Translation
[۲۸] Allele

مضر غالبا با حذف بخشی از ژن همراه می‌باشد. در واقع می‌توان گفت که تغییر نادرست در ژن باعث مضر شدن جهش می‌شود.

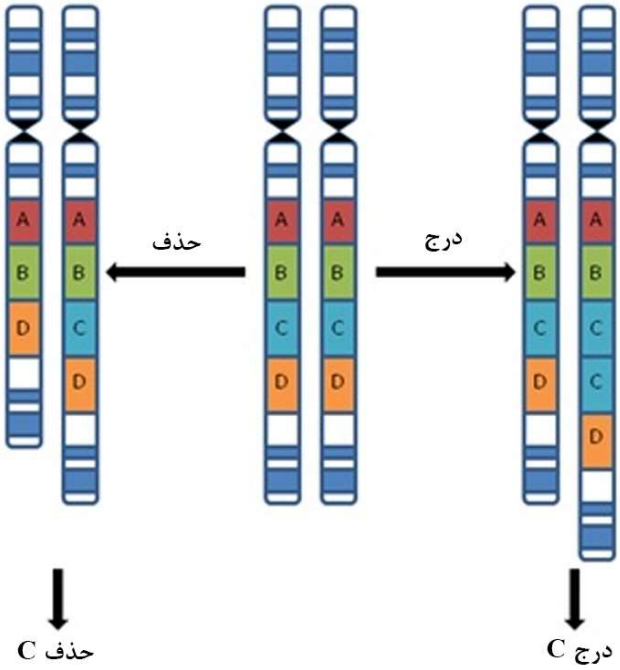
۲- جهش‌های خنثی: آن‌هایی هستند که حذف یا اضافه شدن آن تغییری در ویژگی‌های افراد ایجاد نمی‌کند. در واقع این جهش نه سودمند می‌باشد و نه برای فرد ضرری دارد.

۳- جهش‌های مفید: جهش‌های سودمند آن‌هایی هستند که آلل‌های حاصل از آن به دلیل اینکه سازگاری فرد حامل جهش را افزایش می‌دهند، باقی می‌مانند. در نهایت این جهش‌ها تمایل دارند که در جمعیت ثابت شوند. طی فرایند تثبیت، یک آلل جایگزین آلل دیگری می‌شود.

بیشتر جهش‌ها از نظر ارزش بقاء زیان‌آور یا بی‌فایده‌اند. در موارد به نسبت نادری جهش‌ها سودمند هستند. برخی از این جهش‌ها به جانوران در رقابت با دیگران برتری می‌دهند. در این گونه موارد، افرادی که این ژن‌های مفید را دارا می‌باشند صفات برتری نسبت به دیگران دارند. از آن‌جایی که ما و دیگر موجودات زنده فراورده انتخاب طبیعی هستیم، مجموعه ژن‌های موجود ما باید نزدیک به آن چیزی باشد که محیطی که ما و اجداد ما در آن زیسته‌ایم آن را بیشتر ایجاب می‌کرده است. علاوه بر این، مجموعه کامل ژن‌های موجود در یک فرد، باید ترکیب متوازنی را بوجود آورده باشند. بنابراین اگر بسیاری از این ژن‌ها جهش پیدا کنند، احتمال اینکه توازن موجود برقرار بماند بسیار اندک خواهد بود. از این رو به ندرت، جهش‌ها باعث ایجاد تغییرات مفید در ژن‌ها می‌شوند. فرایند جهش هم علت تغییرات جزئی در درون یک گونه معین و هم بسیاری از تغییرات عمده‌ای را که منجر به از بین رفتن کامل یک گونه می‌شود توضیح می‌دهد. بسیاری از بیماری‌های ژنتیکی در اثر وقوع انواع مختلف جهش‌ها (جهش‌های نقطه‌ای، حذف، تعویض، مضاعف شدن، جابجایی و معکوس شدن)، در ژن‌ها ایجاد می‌شوند [۲۵].

۲-۳ تنوع تعداد کپی (CNV)

تنوع تعداد کپی یا CNV از انواع خانواده جهش‌های مضر می‌باشد. CNV، یک زمینه جدید در علم ژنتیک است که به بررسی پدیده‌ای می‌پردازد که در طی آن قسمتی از ژنوم فرد در طول DNA تکرار می‌شود و تعداد این تکرار برای اشخاص متفاوت با یکدیگر فرق می‌کند. همچنین CNV یک نوع تنوع ساختاری در ژنوم موجودات محسوب می‌شود و بخصوص یک نوع مضاعف شدن و حذف در ژنوم است که تعداد زیادی از نوکلئوتیدها را درگیر می‌کند [۲۶].



شکل ۲-۵: نمونه‌ای از حذف و درج (مضاعف شدن) در ژن [۲۶]

۲-۴ شبکه عصبی

پیاده‌سازی ویژگی‌های مغز در یک سیستم مصنوعی (سیستم دینامیکی که به دست بشر ساخته شده است) همیشه وسوسه انگیز و مطلوب بوده است. محققان بسیاری در طول سال‌ها در این زمینه تلاش کرده‌اند، و نتیجه این فعالیت‌ها، صرف‌نظر از دست‌آوردهای بسیار با ارزش، باور به این اصل بوده است

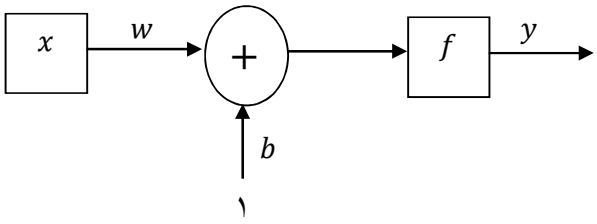
که مغز بشر دست نیافتنی است. باید اعتراف نمود که هدف عالی و کافی نبودن دانش موجود، خود دلیل انگیزش پژوهش‌های فراوانی در این زمینه بوده و خواهد بود، همان‌طور که امروز شاهد بروز این‌چنین فعالیت‌هایی در قالب شبکه‌های مصنوعی هستیم [۲۷]. بسیاری از کسانی که با این سیستم‌ها آشنایی دارند به اغراق آمیز بودن نام آن‌ها موافق هستند. باید توجه داشت که شبکه های عصبی چه در بعد آنالیز و توسعه ساختاری و چه در بعد پیاده‌سازی سخت افزاری، از نظر کمی و کیفی و توانایی، در حال رشد و پیشرفت می‌باشد و تکنیک‌های مختلف محاسبات عصبی از لحاظ تعداد، همچنان در حال افزایش است و فعالیت علمی و کاربردی اساساً در مسائل فنی مهندسی از قبیل سیستم‌های کنترلی، پردازش سیگنال‌ها و شناسایی الگو گسترش یافته است.

۲-۴-۱ کاربرد شبکه عصبی

کاربرد شبکه‌های عصبی حوزه وسیعی از مسائل در زمینه‌های مختلف فنی، مهندسی و اجتماعی از قبیل سیستم‌های هوافضا، بایولوژی، وسایل اتوماتیک خانگی، بانک‌داری، الکترونیک، صنایع دفاعی، وسایل گرم کننده، پزشکی، صدا و تصویر، روبات‌ها، ارتباط راه دور، سیستم‌های حمل و نقل و غیره خواهد گرفت. آنچه که شبکه‌های عصبی را در آینده محبوب‌تر خواهد نمود سرعت محاسباتی بالای کامپیوترها و الگوریتم‌های یادگیری سریع‌تر است که استفاده بیشتر از شبکه‌های عصبی را در مسائل صنعتی با محاسبات زیاد الزامی می‌نماید.

۲-۴-۲ ساختار شبکه‌های عصبی

یک نرون، کوچکترین واحد پردازش‌گر اطلاعات است که اساس عملکرد شبکه های عصبی را تشکیل می دهد [۲۸]. شکل ۲-۶ مدل نرون در شبکه عصبی را نشان می‌دهد.



شکل ۲-۶: مدل نرون در شبکه عصبی

میزان تاثیر x روی y بوسیله مقدار عددی w تعیین می‌شود. ورودی دیگر، که مقدار ثابت یک است در جمله بایاس b ضرب شده و سپس با حاصل‌ضرب x در w جمع می‌شود. سپس این حاصل جمع، ورودی برای تابع فعالیت f خواهد بود. بنابراین خروجی نرون با به صورت رابطه (۲-۱) تعریف می‌شود.

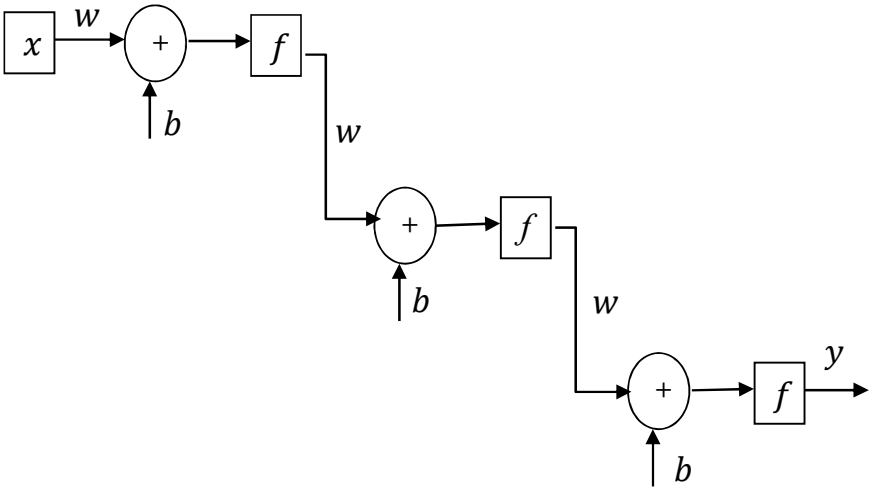
$$y = f\,(\mathrm{wx} + \mathrm{b}) \qquad \text{رابطه (۲-۱)}$$

با مقایسه این مدل تک ورودی با یک نرون عصبی، عملاً w معادل شدت سیناپس جمع کننده و تابع فعالیت معادل هسته سلول و سیگنال خروجی نرون، y معادل سیناپس گذرنده از اکسون خواهند بود. به این نکته که پارامترهای w و b قابل تنظیم هستند و تابع فعالیت f می‌تواند در انواع مختلف مانند، تابع سیگوئید[۲۹]، تانژانت هیپربولیک[۳۰]، تابع هار[۳۱] و غیره انتخاب می‌شوند. بر اساس انتخاب f و نوع الگوریتم یادگیری، پارامترهای w و b متناسب با آن تنظیم می‌شوند. به این ترتیب یادگیری در شبکه عصبی بدین معنی است که w و b طوری تغییر می‌کنند تا رابطه ورودی و خروجی نرون با هدف خاصی مطابقت نماید. در شبکه‌های تک لایه بردار ورودی x توسط نرون‌های لایه (عناصر محاسباتی) طبق رابطه (۲-۱) به بردار خروجی متصل می‌شود. این شبکه شکل بسیار ساده‌ای از شبکه‌های پیشخور[۳۲]

---

[۲۹] Sigmoid
[۳۰] tanh
[۳۱] Haar
[۳۲] Feed Forward

می‌باشد. شکل ۲-۷ نمونه‌ای از شبکه‌های پیشخور را نشان می‌دهد. شبکه‌های عصبی چند لایه دارای توانایی بیشتری نسبت به شبکه‌های عصبی تک لایه‌اند. شبکه‌های عصبی پیشخور دو لایه با توابع



شکل ۲-۷: شبکه پیش‌خور با سه لایه

سیگوئید در لایه اول قادرند هر تابعی را به دقت دلخواه تقریب بزنند. در حالی که شبکه‌های عصبی تک لایه قادر به انجام این کار نیستند. به نظر می‌رسد که باید تعداد درجات آزادی (مثلاً تعداد ورودی‌ها، خروجی‌ها و نرون‌های هر لایه) برای طراحی یک شبکه چند لایه پیشخور زیاد باشد اما توجه به این نکته مهم است که تعداد ورودی‌های شبکه و تعداد خروجی‌های شبکه بر اساس مساله‌ای که باید حل شود، بدست می‌آیند. به عبارت دیگر این پارامتر جزء پارامترهای آزاد طراح شبکه نیستند بلکه انتخاب آن‌ها به مسأله در حال بررسی بستگی دارد. علاوه بر این ویژگی‌های خروجی مطلوب شبکه در انتخاب نوع تابع فعالیت هر لایه تأثیر دارد.

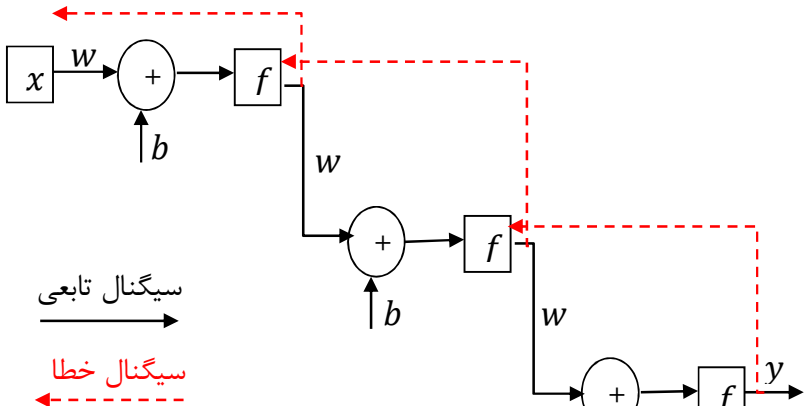
۲-۴-۳ ساختار شبکه عصبی انتشار به عقب

توسعه الگوریتم انتشار به عقب با فراهم آوردن روشی که از نظر محاسباتی بسیار کارا می‌باشد، تحولی عظیم در شبکه‌های عصبی ایجاد نمود، شبکه‌های MLP و همچنین شبکه‌های عمیق با قاعده آموزش

انتشار به عقب، امروزه بیشترین کاربرد را در حل مسائل فنی، مهندسی و شناسایی الگو دارد.

در ساختار شبکه‌های عصبی انتشار به عقب همان‌طور که در شکل ۲-۸ ملاحظه می شود، هر نرون در هر لایه به تمامی نرون‌های لایه قبل متصل می‌باشد. این شبکه عملاً از بهم پیوستن سه شبکه پرسپترون تک لایه ایجاد شده است که یکی لایه خروجی و دو لایه دیگر، لایه‌های میانی شبکه نامیده می‌شوند. خروجی‌های لایه اول بردار، به عنوان ورودی لایه دوم را تشکیل می‌دهند و به همین ترتیب بردار خروجی لایه دوم، ورودی‌های لایه سوم را می‌سازند و خروجی‌های لایه سوم پاسخ واقعی شبکه را تشکیل می‌دهند به عبارتی دقیق‌تر روند سیگنالی در شبکه در یک مسیر پیشخور، از چپ به راست از لایه‌ای به لایه دیگر صورت می‌گیرد. هر لایه در این شبکه می‌تواند از تعدادی نرون‌های مختلف با توابع فعال‌ساز[۳۳] متفاوت برخوردار باشد.

در شبکه MLP دو نوع سیگنال عموماً استفاده می‌شوند که بهتر است از هم متمایز نشان داده شوند. یک نوع سیگنال‌هایی هستند که در مسیر رفت، حرکت می‌کنند و دسته دیگر سیگنال‌هایی هستند که



شکل ۲-۸: ساختار شبکه عصبی انتشار به عقب

[۳۳] Activation Function

در مسیر برگشت حرکت می کنند. به دسته اول سیگنال‌های تابعی[۳۴] و دسته دوم را سیگنال‌های خطا[۳۵] گویند. دلیل این نام‌گذاری این است که سیگنال‌های دسته نخست بر اساس تابعی از ورودی‌های هر نرون و پارامترهای شبکه متناظر با آن محاسبه می شوند و سیگنال‌های دسته دوم به خاطر منشعب شدن از سیگنال خطا و برگشت از لایه خروجی به لایه‌های دیگر شبکه، به سیگنال‌های خطا موسوم هستند. به طور خلاصه سیگنال تابعی, در مسیر رفت در شبکه از لایه‌ای به لایه دیگر توزیع می‌شود و سیگنال‌های خطا در مسیر برگشت در شبکه منتشر می‌گردد. بنابراین هر نرون در شبکه MLP دو نوع محاسبه انجام می‌دهد. در محاسبه اول سیگنال تابعی و در محاسبه دوم تخمین لحظه‌ای از گرادیان منحنی خطا را نسبت به پارامترهایی که ورودی نرون را به خود متصل می‌کند در اختیار قرار می‌دهد. این گرادیان‌ها جهت پخش سیگنال‌های خطا در شبکه مورد نیاز می‌باشند.

۲-۴-۴ شبکه عصبی بازگشتی

شبکه‌های عصبی بازگشتی[۳۶] یا RNN در واقع برای پردازش سیگنال‌های دنباله‌دار به وجود آمدند. در یک شبکه عصبی معمولی تمام ورودی‌ها و خروجی‌ها مستقل از یکدیگر هستند، اما در بسیاری از موارد این ایده نمی‌تواند خیلی جالب باشد. به طور مثال فرض کنید در یک جمله به دنبال پیش‌بینی کلمه بعدی باشیم در صورتی که شبکه نتواند روابط بین کلمات را یاد بگیرد مسلما نمی‌تواند کلمه بعدی را به درستی پیش‌بینی کند. می‌توان گفت در نگاه اول شاید کمی عجیب به نظر برسد اما این RNNها در واقع کپی‌های متعدی از شبکه‌های عصبی معمولی هستند که کنار هم چیده شده‌اند و هر کدام پیغامی را به دیگری منتقل می‌کنند.

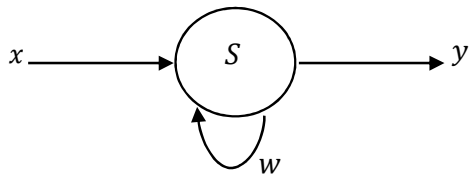
---

[۳۴] Functional Signals

[۳۵] Error Signals

[۳۶] Recurrent Neural Network

اگر از یک دیدگاه به شبکه‌های بازگشتی RNN نگاه کنیم، این شبکه‌ها دارای یک نوع حافظه هستند که اطلاعات تاکنون دیده شده را ثبت و ضبط می‌کند. در تئوری این‌طور به نظر می‌رسد که RNNها می‌توانند اطلاعات موجود در یک دنباله طولانی را ثبت و از آن‌ها استفاده کنند اما در عمل این‌گونه نیست و محدودیت‌های فراوانی دارند. به این صورت که فقط اطلاعات چند گام قبل را ضبط می‌کنند. در شکل ۲-۹ نمونه‌ای از یک شبکه عصبی بازگشتی معمولی نمایش داده شده است.



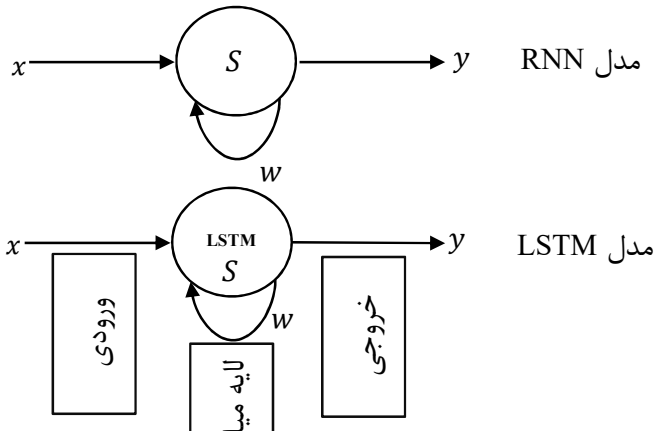
شکل ۲-۹: مدل نرون در شبکه‌های RNN

در شکل ۲-۹، x ورودی در گام زمانی t است و Sحالت پنهان در گام زمانی t است. این قسمت همان جایی است که حافظه شبکه در آن قرار گرفته است. حالت S بر اساس حالت پنهان قبلی و ورودی که در زمان جاری به آن داده می‌شود تغییر می‌کند که به شکل $S_t = f(x_t \cdot w_{s(t-1)})$ است. تابع f معمولا یک تابع غیر خطی مثل تانژانت هیپربولیک است. پارامتر y خروجی در گام زمانی t است. به عنوان مثال در محاسبه کلمه بعدی در یک جمله این خروجی می‌تواند یک بردار از احتمالات در واژگان باشد. همان‌طور که اشاره کردیم، S در واقع جایی است که حافظه شبکه در آن قرار گرفته است. حالت S اطلاعاتی درباره این‌که تاکنون در شبکه چه رخ داده است را ضبط می‌کند. خروجی S با توجه به حالت‌های قبلی محاسبه می‌شود اما همان‌طور که بیان شد S نمی‌تواند اطلاعات موجود در گام‌های زمانی (به عنوان مثال ۱۰ گام قبل) را حفظ نماید. برخلاف شبکه‌های معمولی که از پارامترهای متفاوتی در هر لایه استفاده می‌کند، یک شبکه RNN پارامترهای مشابهی را بین همه گام‌های زمانی به اشتراک می‌گذارد. این بدین معنی است که ما در هر گام زمانی عملیات مشابهی را انجام می‌دهیم فقط ورودی‌ها متفاوت هستند. با این تکنیک تعداد کلی پارامترهایی که شبکه بایستی یاد بگیرد به شدت کاهش پیدا

می‌کند. ویژگی اصلی در RNN حالت پنهان آن است که اطلاعات یک توالی را ذخیره می‌کند. همچنین نیاز نیست در هر گام زمانی یک خروجی و یا حتما یک ورودی داشته باشیم بلکه بر اساس عملیات مورد نظر، معماری RNN می‌تواند تغییر کند. شبکه‌های RNN به این علت بازگشتی نامیده می‌شوند که خروجی هر لایه به محاسبات لایه‌های ماقبل آن وابسته است. به عبارتی دیگر این شبکه‌ها دارای حافظه هستند که اطلاعات مربوط به داده‌های دیده شده را ذخیره می‌کند.

۲-۴-۵ شبکه عمیق LSTM

حافظه طولانی مدت-کوتاه مدت یا LSTM[۱] یک نوع مدل یا ساختار برای داده‌های ترتیبی است که در در سال ۱۹۹۵ برای توسعه شبکه‌های عصبی بازگشتی(RNN) معرفی شد. سپ هوچرِتر[۲] در مقاله LSTM [۲۹] بیان می‌کند که عبارت Long Term Memory به وزن‌های یادگرفته شده و Short Term Memory به حالت‌های درونی سلول‌ها گفته می‌شود. LSTM برای حل مشکل پدیده ناپدید شده گرادیان در شبکه‌های عصبی بازگشتی بوجود آمدند که تغییر عمده آن جایگزین کردن لایه میانی RNN با یک بلاک، که بلاک LSTM نام دارد است. همان‌طور که در شکل ۲-۱۰ نشان داده شده است،



شکل ۲-۱۰: مدل RNN و مدل LSTM

[۱] Long Short Term Memory

[۲] Sepp Hochreiter

تفاوت شبکه RNN با مدل LSTM این است که در مدل LSTM به جای یک واحد نرون با سلول LSTM روبرو هستیم. در بخش‌های بعدی با این واحدها آشنا می‌شویم.

بزرگ‌ترین ویژگی LSTM امکان یادگیری وابستگی بلند مدت است که توسط شبکه‌های RNN این امکان وجود نداشت. برای پیش‌بینی گام زمانی بعدی نیاز است که مقادیر وزن‌ها در شبکه بروزرسانی شوند که این کار نیازمند حفظ اطلاعات گام‌های زمانی ابتدایی است. یک شبکه RNN فقط می‌تواند تعداد محدودی از وابستگی‌های کوتاه مدت را یاد بگیرد ، اما سری‌های زمانی بلند مدت مانند ۱۰۰۰ گام زمانی قابل یادگیری توسط RNNها نیستند اما LSTMها می‌تواند این وابستگی‌های بلند مدت را به درستی یاد بگیرند. همان‌گونه که در جدول ۲-۱ (این جدول از مقاله هوچرِتر استخراج شده است [۲۹]) نمایش داده شده است، RNN نمی‌تواند حتی خروجی شبکه با ۱۰۰ تاخیر(گام زمانی) را با هر دو روش RTRL[۱]و BPTT[۲]پیش‌بینی کند اما LSTM در زمان کوتاه توانست این کار را انجام دهد.

جدول ۲-۱: مقایسه عملکرد LSTM با روش‌های RTRL و BPTT [۲۹]

Method	Delay p	Learning rate	Weights	Successful trials	Success after
RTRL	٤	١,٠	٣٦	٧٨	١٠٤٣٠٠٠
RTRL	٤	٤,٠	٣٦	٥٦	٨٩٢٠٠٠
RTRL	٤	١٠,٠	٣٦	٢٢	٢٥٤٠٠٠
RTRL	١٠	١,٠-١٠,٠	١٤٤	٠	>٥٠٠٠٠٠٠
RTRL	١٠٠	١,٠-١٠,٠	١٠٤٠٤	٠	>٥٠٠٠٠٠٠
BPTT	١٠٠	١,٠-١٠,٠	١٠٤٠٤	٠	>٥٠٠٠٠٠٠
CH	١٠٠	١,٠	١٠٥٠٦	٣٣	٣٢٤٠٠
LSTM	١٠٠	١,٠	١٠٥٠٤	١٠٠	٥٠٤٠

قبل از این‌که وارد جزییات ساختار LSTM شویم، ابتدا لازم است پدیده ناپدید شدن گرادیان که

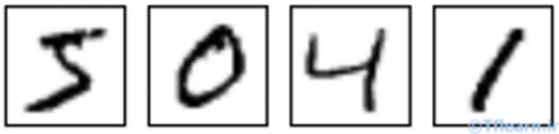
[۱] Real-Time Recurrent Learning

[۲] Back-Propagation Through Time

اولین بار در سال ۱۹۹۱ توسط هوچرِتر در پایان‌نامه ایشان مطرح شد را بررسی کنیم. دو الگوریتم برای آموزش شبکه‌های RNN با نام‌های BPTT یا RTRL وجود دارد. هر دوی این الگوریتم‌ها بر پایه مشتق‌گیری کار می‌کنند. گرادیان الگوریتم‌هایی از این نوع در انتشار رو به عقب در سری‌هایی با وابستگی بلند مدت، ممکن است ناپدید شود یا به بی‌نهایت میل کنند. این مشکل نه تنها در RNNها بلکه در شبکه‌های عمیق عصبی جایی که گرادیان از میان چندین گام عبور می‌کند نیز اتفاق می‌افتد. مکانیسمی که باعث می‌شود شبکه‌های عصبی یاد بگیرند بسیار ساده و با نام کاهش گرادیان تصادفی[۱] شناخته می‌شود. به طور خلاصه برای هر پارامتر در شبکه (وزن‌ها و بایاس‌ها)، کاری که ما بایستی انجام دهیم محاسبه مشتق تابع هزینه نسبت به پارامترها و حرکت دادن آن به مقدار کم در سمت مسیر مخالف آن است [۳۰]. این الگوریتم بسیار ساده، در برخی از شبکه ها رفتار اعجاب انگیزی را از خود نشان می دهد، بطوری که وزن هایی که نزدیک‌تر به انتهای شبکه هستند نسبت به وزن‌های موجود در لایه‌های اولیه با مقدار خیلی بیشتری تغییر می‌کنند. هر چه تعداد لایه های شبکه بیشتر باشد مقدار پارامترهای لایه های اول، کمتر دچار تغییرات می‌شود. این امر مشکل‌ساز است زیرا مقادیر وزن‌ها در ابتدای کار به صورت تصادفی مقداردهی می‌شوند و اگر این مقادیر به ندرت تغییر کنند، یا هیچ وقت تغییر نکند، نمی‌توانند مقدار صحیح وزن‌ها را پیدا کنند یا این‌که این فرآیند بسیار زمان بر خواهد بود.

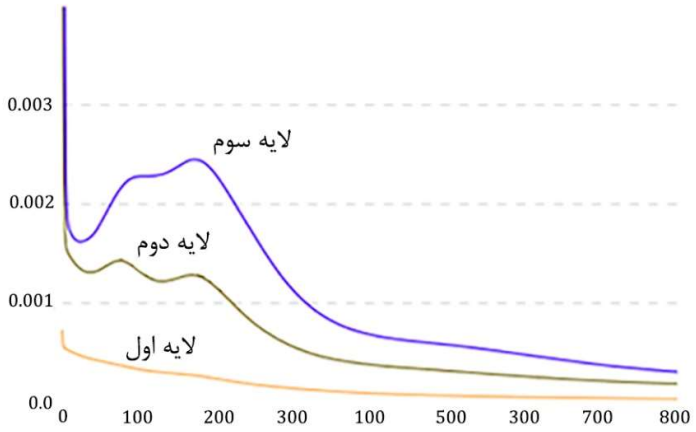
برای شفاف‌تر شدن بحث گرادیان، یک شبکه ساده برای طبقه‌بندی عکس‌های موجود در دادگان MNIST آموزش داده شده است. MNIST یک مجموعه داده ساده در زمینه بینایی ماشین(کامپیوتر) است. این مجموعه داده شامل تصاویری از ارقام دست‌نویس انگلیسی مانند شکل ۲-۱۱ است [۳۱].

[۱] Stochastic Gradient Descent



شکل ۲-۱۱: ارقام دست‌نویس در دادگان MNIST [۳۱]

همچنین در این مجموعه، برچسب‌هایی برای هر تصویر وجود دارد که بیان‌گر این است که هر تصویر نمایان‌گر چه رقمی می‌باشد. در شکل ۲-۱۲ تغییرات گرادیان در طول زمان در لایه‌های مختلف شبکه برای دادگان MNIST نمایش داده شده است:

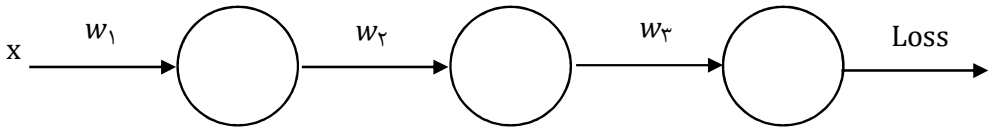


شکل ۲-۱۲: تغییرات گرادیان در طول زمان در لایه‌های مختلف شبکه [۳۱]

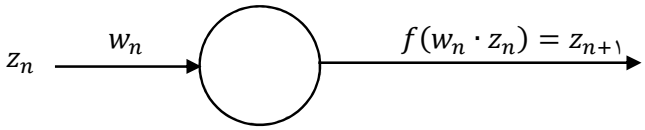
همان‌طور که در شکل ۲-۱۲ مشاهده می‌کنیم مقدار گرادیان در لایه اول نسبت به لایه سوم کمتر است. این فقط برای یک شبکه خیلی کوچک است موضوع بسیار جالب‌تر خواهد شد اگر تعداد این لایه‌ها را افزایش دهیم. از آنجایی که بقیه شبکه با مقادیر پارامترهای موجود در لایه‌های اولیه تحت تاثیر قرار می‌گیرند، اگر این پارامترها کاملا غلط باشند شبکه به صورت صحیح عمل نخواهد کرد.

۱-۵-۴-۲ دلیل ناپدید شدن گرادیان

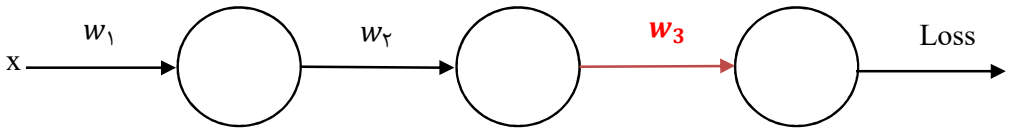
ناپدید شدن گرادیان[۱] را با یک مثال ساده و به صورت گام به گام توضیح می‌دهیم. فرض کنیم که یک شبکه با سه لایه داریم. برای سادگی فرض می‌کنیم هر کدام از لایه‌ها فقط یک نرون دارند:



هر نرون با یک w (وزن) در ارتباط است و خروجی آن تابعی از w در مقدار ورودی است. به عنوان مثال نرون اول خروجی $w_1 \cdot x$ را تولید می‌کند . در عمل این تابع می‌تواند سیگموئید یا تانژانت هیپربولیک یا تابع ReLU باشد، که در این‌جا آن را تابع $f(x)$ می‌نامیم. از این‌جا به بعد خروجی این تابع برای نرون n را z_n می‌نامیم:



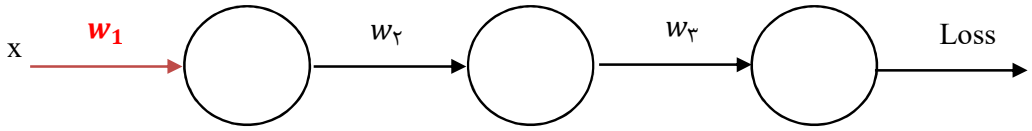
انتهای شبکه با تابع هزینه به پایان می‌رسد که این تابع تفاوت مقدار حقیقی و مقدار تولید شده را محاسبه می‌کند. این تابع در واقع خروجی لایه انتهایی خواهد شد. تابع هزینه میزان خطای هر بار اجرای شبکه‌ی عصبی را برای داده‌های آموزشی را محاسبه می‌کند. اکنون می‌خواهیم الگوریتم انتشار رو به عقب را بررسی کنیم. در گام نخست می‌بایست مقدار w_3 را بروزرسانی کنیم.



[۱] Gradiant Vanishig

برای این منظور، نیاز است که مشتق تابع هزینه نسبت به w_3 محاسبه شود. اگر ما قاعده زنجیری را طبق رابطه (۲-۲) اعمال کنیم:

$$\frac{\partial Loss}{\partial w_3} = \frac{\partial Loss}{\partial f(z_3)} \cdot \frac{\partial f(z_3)}{\partial w_3} = \frac{\partial Loss}{\partial f(z_3)} \cdot f'(z_3) \cdot w_3 \qquad \text{رابطه (۲-۲)}$$



مقدار w_3 را اندکی در مسیر مخالف حرکت می‌دهیم. برای لایه‌های قبلی این عملیات مجدد تکرار می‌شود. بنابراین اگر به ابتدای شبکه پرش کنیم می‌توانیم ببینیم که چگونه وزن اولین لایه تغییر می‌کند:

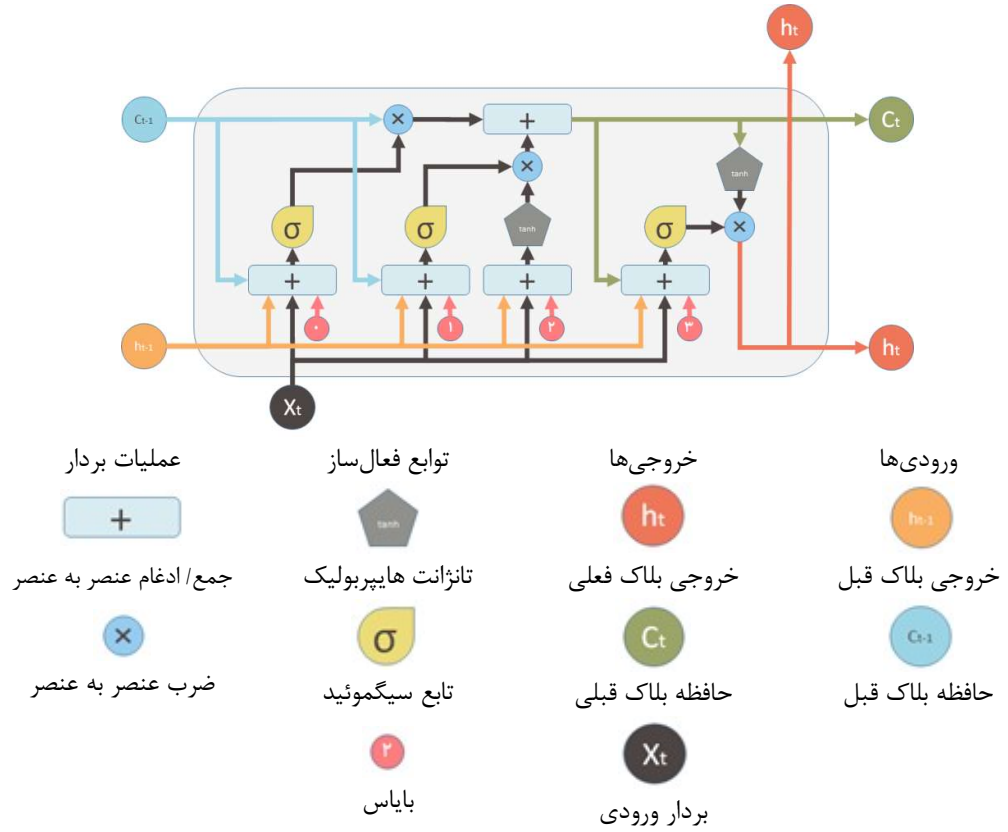
$$\frac{\partial Loss}{\partial w_1} = \frac{\partial Loss}{\partial f(z_3)} \cdot \frac{\partial f(z_3)}{\partial f(z_2)} \cdot \frac{\partial f(z_2)}{\partial f(z_1)} \cdot \frac{\partial f(z_1)}{\partial w_1}$$
$$= \frac{\partial Loss}{\partial f(z_3)} \cdot f'(z_3) \cdot w_3 \cdot f'(z_2) \cdot w_2 \cdot f'(z_1) \cdot w_1 \qquad \text{رابطه (۲-۳)}$$

همان‌طور که مشاهده می‌کنیم، در محاسبه گرادیان برای w_1 تعداد عبارت‌ها افزایش پیدا می‌کند. روش معمول برای مقداردهی اولیه پارامترها مانند وزن‌ها در یک شبکه عصبی به فرم یک تابع گوسی با میانگین صفر و انحراف معیار یک است، که گویای این موضوع است که اندازه پارامترهای شبکه (وزن‌ها) کمتر از یک می‌باشد. اگر تابع $f(x)$ ما یک تابع مانند سیگموئید باشد آن‌گاه مشتق همیشه کمتر از ۰,۲۵ خواهد شد. حالا اگر درنظر بگیریم که اگر تعداد بسیار زیادی از این اعداد کوچک در یکدیگر ضرب شوند باعث ایجاد یک عدد بسیار کوچک خواهند شد. سناریوی مخالف این شرایط نیز می‌تواند اتفاق بیفتد به طوری که مقدار وزن‌ها بزرگ شوند آنگاه ضرب این اعداد در یک دیگر باعث می‌شود گرادیان

به بی‌نهایت میل کند که به آن انفجار گرادیان[۲] می‌گویند [۳۰].

۲-۴-۵-۲ شبکه عمیق LSTM

برای این‌که مفهوم یک بلاک LSTM را بهتر متوجه شویم و از محاسبات پیچیده استفاده نکنیم و بتوانیم شبکه را از لحاظ مفهومی بیان کنیم، شکل ۲-۱۳ را از [۳۲] و [۳۳] معرفی می‌کنیم[۳].



شکل ۲-۱۳: شمایی از یک واحد LSTM [۳۱ و ۳۲]

شبکه LSTM دارای ۳ پارامتر ورودی می‌باشد:

- پارامتر x_t : ورودی گام فعلی

---

[۲] Gradiant Exploiding

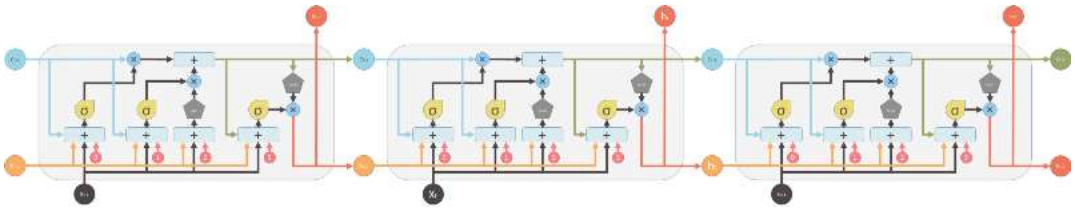
[۳] دیاگرام‌هایی که در این بخش آمده برگفته از فایل پاورپوینت آقای Shi Yan بوده که در مخزن گیت‌هاب ایشان قرار گرفته است.

- پارامتر h_{t-1}: خروجی بلاک LSTM قبلی
- پارامتر C_{t-1}: خروجی سلول حافظه بلاک LSTM قبلی

همچنین این شبکه دارای ۲ پارامتر خروجی است:

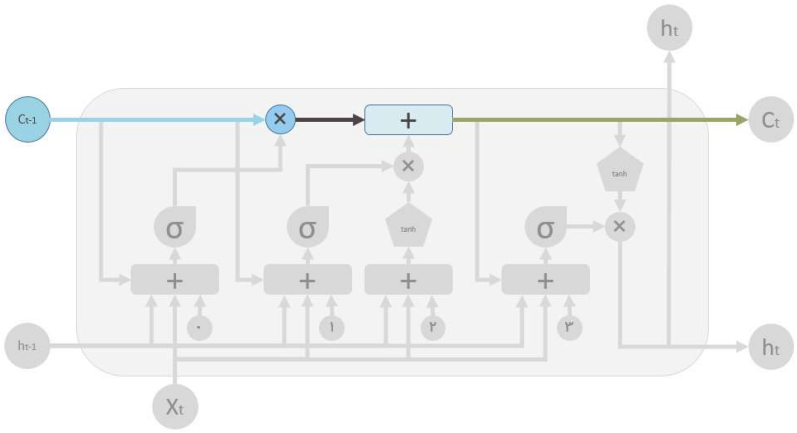
- پارامتر h_t: خروجی گام زمانی t
- پارامتر C_t: حافظه سلول کنونی

هر واحد بر اساس ورودی کنونی، خروجی واحد پیشین و حافظه واحد قبل تصمیم‌گیری می‌کند سپس یک خروجی جدید را ایجاد کرده و مقدار حافظه خود را اصلاح می‌کند.



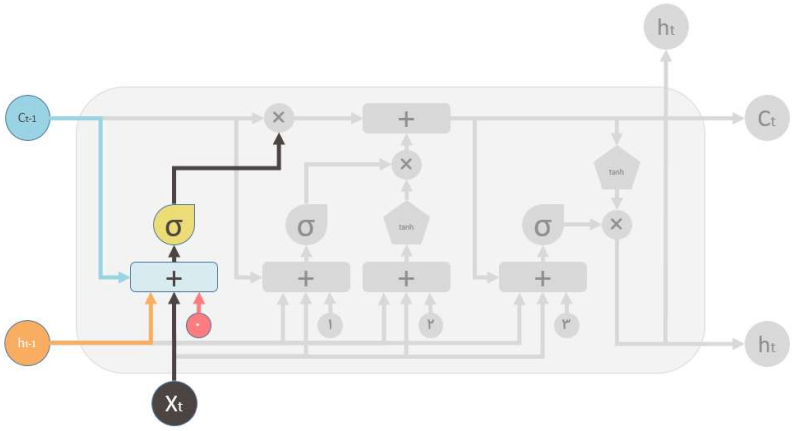
شکل ۲-۱۴: شمایی از ۳ واحد LSTM

حافظه سلول توسط دو مقدار تغییر می‌نماید. مقدار اول دروازه فراموشی است. اگر این دروازه کاملا بسته شود، حافظه قبلی کاملا پاک می‌شود اما اگر دروازه باز باشد تمام مقدار حافظه قبلی از آن عبور می‌کند. دوم مقدار حافظه جدید است. مقدار حافظه جدید با حافظه قبلی ادغام می‌شود. این‌که چه مقدار حافظه جدید بایستی وارد شود توسط مقدار دوم کنترل می‌شود.



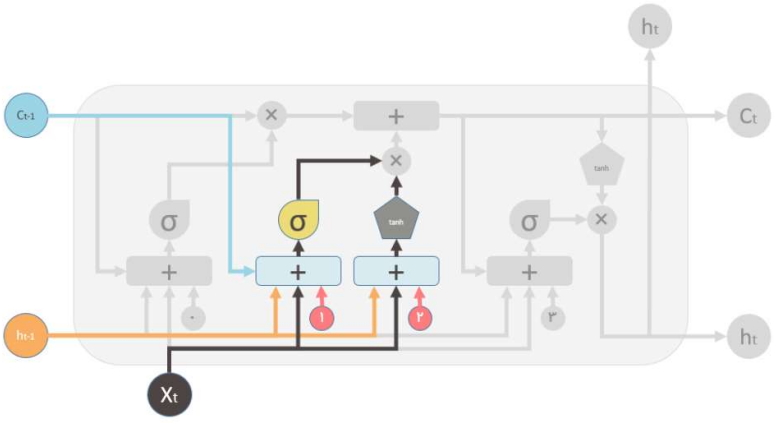
شکل ۲-۱۵ :مسیر حافظه در شبکه LSTM

شکل ۲-۱۵ مسیر حافظه در شبکه LSTM را نشان می‌دهد که در آن ورودی، مقدار حافظه قبلی است. علامت × همان‌طور که در بالا نشان داده شده است یک عملیات ضرب عنصر به عنصر است. مقدار حافظه قبلی در مسیر ورودی خود از دروازه فراموشی عبور می‌کند . نتیجه این عملیات مشخص کننده این است که آیا می‌خواهیم مقدار حافظه قبلی را نگه داریم یا این‌که آن را به فراموشی بسپاریم. بطور مثال اگر نتیجه ضرب C_t با این خروجی این دروازه نزدیک صفر باشد به این معنی است که می‌خواهیم بیشتر حافظه قبلی را فراموش کنیم اما در صورتی که مقدار یک باشد بدین معنی است که می‌خواهیم مقدار حافظه قبلی به صورت کامل از مسیر عبور کند. در مرحله دوم نتیجه عملیات از عملگر + (جمع عنصر به عنصر) عبور می‌کند. این عملگر وظیفه ادغام حافظه قبلی و حافظه کنونی را بر عهده دارد. این‌که چقدر از حافظه جدید بایستی به حافظه قبلی اضافه شود توسط عملگر × که در زیر عملگر + قرار گرفته است مشخص می‌شود. بعد از اجرای این عملیات حافظه قبلی C_{t-1} به حافظه جدید C_t تغییر می‌کند.



شکل ۲-۱۶: دروازه فراموشی

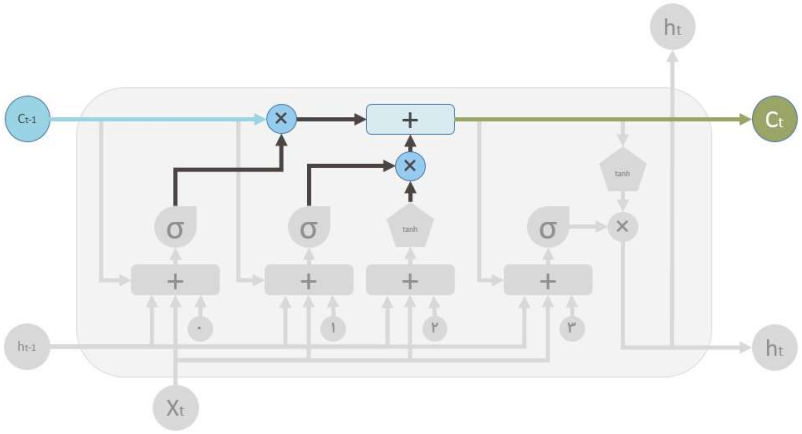
در شکل ۲-۱۶ دروازه فراموشی مشخص شده است. دروازه فراموشی در واقع یک شبکه عصبی کوچک است که سه ورودی و یک بایاس را دریافت می‌کند و از یک تابع سیگموئید به عنوان تابع فعال‌ساز استفاده می‌کند. خروجی این شبکه مقدار فراموشی را مشخص می‌کند. به طور کلی دروازه فراموشی پاسخگوی این سوال است که آیا سلول حافظه پاک شود یا می‌خواهیم حافظه را نگه داریم.



شکل ۲-۱۷: دروازه ورودی

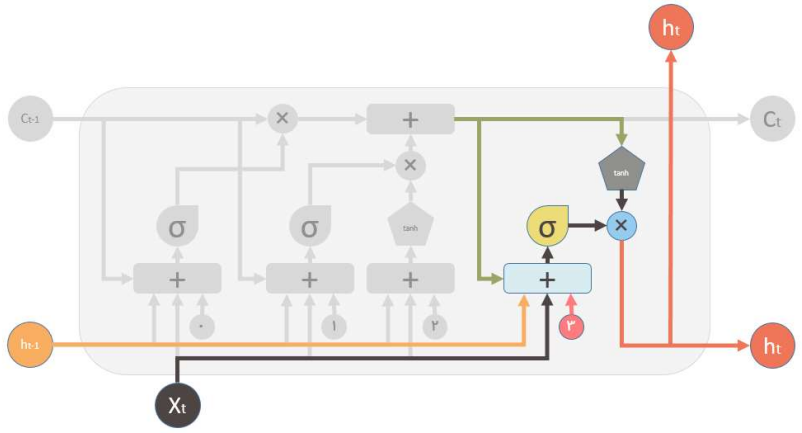
در شکل ۲-۱۷ یک شبکه عصبی با یک لایه کوچک مشخص شده است که ورودی‌هایی مشابه با دروازه فراموشی دریافت می‌کند. خروجی این شبکه میزان تاثیر حافظه جدید را بر روی حافظه قبلی تعیین می‌نماید. اگر چه حافظه جدید توسط یک شبکه دیگر تولید می‌شود که یک شبکه یک لایه است،

اما از تابع تانژانت هیپربولیک به عنوان تابع فعال‌ساز استفاده می‌کند. می توان این‌گونه بیان نمود که دروازه ورودی به سوال " آیا در سلول حافظه بنویسم یا خیر" پاسخ می‌دهد.



شکل ۲-۱۸: محل اعمال مقادیر دروازه های ورودی و فراموشی

دو علامت × مقادیر دروازه فراموشی (سمت چپ) و مقدار حافظه جدید(سمت راست) هستند.



شکل ۲-۱۹: دروازه خروجی

در پایان دروازه خروجی را خواهیم داشت. این دروازه، خروجی بلاک LSTM را تولید می‌کند که توسط حافظه جدید و سایر اجزا مشخص شده در شکل کنترل می‌شود.

۲-۵ اطلاعات متقابل

در بخش های قبل با CNV و شبکه عصبی LSTM آشنا شدیم. اما برای این‌که بتوانیم تحلیل بایولوژیکی از طبقه‌بندی خود داشته باشیم لازم است با مفهوم اطلاعات مقابل[۴] آشنا شویم. اکنون با مفهوم معیار اطلاعات متقابل با ذکر مثال آشنا می‌شویم.

در نظریه احتمالات[۵] و نظریه اطلاعات[۶]، اطلاعات متقابل بین دو متغیر تصادفی معیاری برای نشان دادن میزان وابستگی متقابل آن دو متغیر است. به بیان دیگر در واقع این معیار میزان اطلاعات بدست آمده در مورد یک متغیر تصادفی از طریق متغیر تصادفی دیگری را نشان می‌دهد [۳۴]. مفهوم اطلاعات متقابل بطور ذاتی مرتبط با آنتروپی یک متغیر تصادفی که میزان اطلاعات موجود در یک متغیر تصادفی را نشان می‌دهد، است.

اطلاعات متقابل از رابطه (۲-۴) بدست می‌آید [۳۵]:

$$I(x.y) = \int\int p(x.y) log \frac{p(x.y)}{p(x)p(y)} dxdy \qquad \text{رابطه (۲-٤)}$$

در رابطه (۲-۴) x و y دو بردار هستند و $p(x.y)$ توزیع احتمال توام بین دو متغیر x و y است. همچنین احتمال‌های $p(x)$ و $p(y)$ به ترتیب توزیع احتمال حاشیه بین دو متغیر x و y می‌باشند. در ادامه با ذکر مثالی ساده، به فهم معیار اطلاعات متقابل پی می‌بریم.

فرض کنیم چهار نمونه بیمار داشته باشیم که سه تا از ژن‌های آن‌ها که در واقع ویژگی‌های ما هستند، به نام‌های x_1 و x_2 و x_3دچار اختلال CNV شده است. بعد از بررسی آزمایش‌های بیماران مشخص شد که به یکی از دو کلاس سرطان y_1 و y_2مبتلا شده‌اند.

---

[۴] Mutual Information

[۵] Probability theory

[۶] Information theory

جدول ۲-۲: کلاس نمونه بیماران. سطر: کلاس سرطان، ستون: ژن‌های بیمار(ویژگی‌ها)

	$x_۱$	$x_۲$	$x_۳$
$y_۲$	-۱	۰	۱
$y_۱$	۰	۱	۱
$y_۱$	۱	-۱	۰
$y_۲$	۰	۱	۱

حال این سوال مطرح می‌شود که چگونه می‌توان اطلاعات متقابل $I(x.y)$ را در این مسئله حل کرد. روشن است که در اولین گام باید $p(x.y)$ توزیع احتمال توام بین دو متغیر x و y را بدست آورد.

برای حل مسئله اولین ویژگی $\boldsymbol{x_1}$ را انتخاب می‌کنیم و جدول مستطیلی (ژن (ویژگی) ، کلاس (نوع سرطان)) بر اساس جدول ۲-۲ طراحی می‌کنیم. همان‌طور که در جدول ۲-۲ نشان داده شده است ویژگی $\boldsymbol{x_1}$ دارای یکی از سه مقدار ۰ , ۱- , ۱+ بوده و تعداد حالت‌های کلاس‌ها در این مثال دو تا ($\boldsymbol{y_1}$ و $\boldsymbol{y_2}$) است. برای این‌که بتوان توزیع احتمال توام بین دو متغیر $\boldsymbol{x}$ و $\boldsymbol{y}$ را بدست آورد، کافی است تا تعداد حالت‌هایی که این دو متغیر با یکدیگر اتفاق می‌افتند را بشماریم و در جدول ۲-۳ قرار دهیم.

جدول ۲-۳: جدول مستطیلی بین ویژگی $\boldsymbol{x_1}$ و کلاس‌های $\boldsymbol{y_1}$ و $\boldsymbol{y_2}$

	$y_۱$	$y_۲$
$x_۱ = -۱$	۰	۱
$x_۱ = ۰$	۱	۱
$x_۱ = +۱$	۱	۰

در جدول ۲-۳ مشاهده می‌کنیم $x_۱ = ۰$ هیچ اطلاعاتی را به ما نمی‌دهد زیرا در هر دو کلاس $y_۱$ و $y_۲$ دارای مقدار یک است. محاسبه توزیع‌های توام و توزیع‌های حاشیه‌ای برای این ویژگی به سادگی بدست می‌آید.

$$p(x_1.y) = \begin{pmatrix} 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 0 \end{pmatrix}. \quad p(x_1) = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix}. \quad p(y) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

این عملیات را برای دو ویژگی دیگر x_2 و x_3 نیز تکرار می‌کنیم و با کمک رابطه (۲-۵) می‌توانیم مقدار $I(x.y)$ را به سادگی بدست آوریم.

$$I(x.y) = \sum_{i=1,2,3} \sum_{j=1,2} p(x_i.y_i)\, log \frac{p(x_i.y_i)}{p(x_i)p(y_i)}$$ رابطه (۲-۵)

۲-۶ معیار بیشترین وابستگی-کمترین افزونگی (mRMR)

همان‌طور که در بخش قبل با معیار اطلاعات متقابل آشنا شدیم، از این معیار برای نشان دادن میزان وابستگی بین دو متغیر استفاده می‌شود. همچنین در فصل پنجم از این تحقیق خواهیم دید، از این معیار جهت کاهش ابعاد دادگان استفاده می‌کنیم. برای این منظور بیشترین وابستگی که ویژگی‌ها به کلاس‌ها دارند و کمترین افزونگی که در دادگان دارند را پیدا کرده و ویژگی‌هایی که در این معیار بیشترین امتیاز را گرفته باشند، به عنوان ویژگی‌های برتر انتخاب می‌شوند.

میزان وابستگی (D) یک متغیر به یک کلاس را می‌توان به صورت رابطه (۲-۶) تعریف کرد [۳۶].

$$max\ D(S.y). \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i.y))$$ رابطه (۲-۶)

در رابطه (۲-۷)، S مجموعه ویژگی‌ها می‌باشد.

هنگامی که دو ویژگی در طبقه‌بندی به یکدیگر وابستگی شدید داشته باشند، اگر یکی از آن‌ها حذف شود، در طبقه‌بندی تاثیر چندانی ندارد. بنابراین این نتیجه حاصل می‌شود که یکی از این ویژگی‌ها اضافی بوده و می‌توان آن را حذف کرد. برای محاسبه افزونگی (R) از رابطه (۲-۸) استفاده می‌شود [۳۶].

$$min\ R(S). \quad R = \frac{1}{|S|^2} \sum_{x_i.x_j \in S} I(x_i.x_j))$$ رابطه (۲-۸)

با ترکیب رابطه (۲-۶) و رابطه (۲-۸) معیار بیشترین وابستکی-کمترین افزونگی[۷] یا به اختصار mRMR به صورت **Error! Reference source not found.**) محاسبه می‌شود [۳۶].

$$\max \varphi(D.R). \quad \varphi = D - R$$ رابطه (۲-۹)

در مقاله [۳۶] اثبات شده است که بیشینه کردن تدریجی ویژگی بر طبق این معیار، معادل بیشینه کردن رابطه (۲-۴) است. به این ترتیب به جای بررسی تک‌تک حالت ها یعنی 2^n حالت می‌توان در n مرحله ویژگی‌های مناسب را اضافه کرد و به نتایج دلخواه رسید.

۲-۷ نتیجه‌گیری

در این فصل از تحقیق ابتدا به معرفی DNA پرداختیم و با ساختمان آن آشنا شدیم. همچنین عملکرد DNA در بدن انسان را مورد بررسی قرار دادیم. در ادامه در مورد جهش‌ها درون DNA صحبت کردیم و انواع جهش را معرفی نموده و نشان دادیم که بعضی از جهش‌های مضر می‌توانند سرطان‌زا باشند که آن‌ها را عنوان تنوع تعداد کپی یا CNV معرفی کردیم. در ادامه به بررسی شبکه عصبی پرداختیم و با کاربردهای آن آشنا شدیم. سپس مدل نرون در شبکه عصبی را معرفی کردیم و با شبکه پرسپترون چند لایه MLP آشنا شدیم. همچنین شبکه‌های انتشار به عقب را مورد بررسی قرار دادیم و با شبکه‌های عصبی بازگشتی یا آشنا شدیم. دریافتیم که حافظه شبکه‌های عصبی بازگشتی بسیار کوتاه است و به دلیل مشکل ناپدید شدن گرادیان، نمی‌تواند گذشته خود را به یاد آورد. برای حل این مشکل با شبکه‌های عمیق LSTM آشنا شدیم و ساختمان آن را به زبان ساده و فارغ از ریاضیات پیچیده مورد بررسی قرار دادیم. همچنین به معرفی معیار mRMR پرداخته و با ذکر مثالی با مفهوم آن آشنا شدیم.

---

[۷] Maximum Relevance-Minumum Redundancy

فصل ۳: پیشینه تحقیق

۳-۱ مقدمه

بحث یادگیری عمیق در ژنومیک بسیار جذاب است و مطالعات متعددی در زمینه تشخیص انواع مختلف سرطان‌ها با کمک یادگیری عمیق صورت گرفته است. در این فصل با تعدادی از این مقاله‌ها آشنا می‌شویم و با کاربردهای شبکه‌های عمیق در تشخیص انواع سرطان مانند سرطان ریه، پوست و سینه آشنا می‌شویم. بسیاری از این روش‌ها از شبکه‌های کانولوشن برای تشخیص سرطان استفاده کرده‌اند.

۳-۲ یادگیری عمیق و تشخیص سرطان سینه

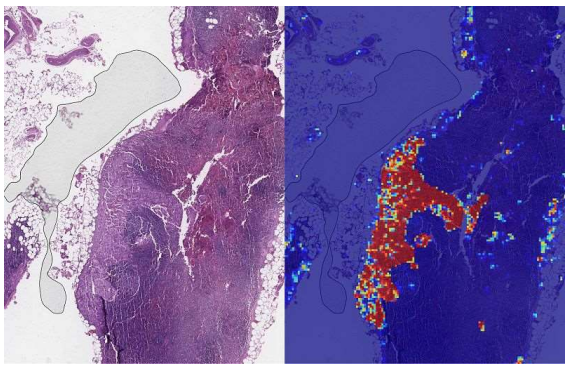
طبق آمار سازمان بهداشت جهانی، سرطان سینه یکی از سرطان‌های شایع در بین زنان و دومین عامل مرگ و میر در بین آن‌ها است [۳۷]. یکی از روش‌های مرسوم در تشخیص این سرطان نمونه‌برداری با کمک جراحی است، که بالاترین دقت تشخیص را داراست. اما این روش زمان‌بر و پرهزینه است. یکی دیگر از روش‌های غربالگری متداول در بررسی این بیماری استفاده از تکنیک‌های تصویر برداری از قبیل ماموگرافی[۸] است که معایب روش قبل را ندارد. همچنین ماموگرافی می‌تواند کوچک‌ترین تغییرات ایجاد شده در سینه را که ممکن است نشان‌هایی از وجود سرطان باشد، قبل از پزشک یا خود بیمار تشخیص دهد [۳۸].

شبکه‌های عصبی عمیق حوزه جدیدی در شبکه‌های عصبی هستند که اخیراً پیشرفت‌های زیادی را در موضوعات مختلف مرتبط با یادگیری ماشین با خود به همراه آورده‌اند. با توجه به ظهور این حوزه در یادگیری ماشین که با خود تعریف جدیدی را برای ساختارهای شبکه‌های عصبی عمیق ارائه کرده است، مشکلات مذکور رفع شده و توانایی استفاده از شبکه‌های عمیق با قابلیت‌های مختلف برای محققان

[۸] Mammography

فراهم شده است.

الگوریتم هوش مصنوعی LYNA[۹] که توسط گوگل ارائه شد برای تشخیص سرطان سینه در حوزه یادگیری عمیق به کار گرفته شده است [۳۹]. این الگوریتم سرطان سینه را با استفاده از تصاویر آسیب‌شناسی[۱۰] و مقایسه آن‌ها با یکدیگر تشخیص می‌دهد. شکل ۳-۱ نمونه‌ای از این تصاویر را نشان می‌دهد. بر اساس آزمایش‌های انجام شده، هوش مصنوعی گوگل[۱۱] با دقت ۹۹ درصد می‌تواند تصاویر مربوط به سلول‌های سرطانی و غیرسرطانی را از یکدیگر جدا نماید [۴۰]. هوش مصنوعی گوگل متاستاز[۱۲] سرطان سینه را در ابعاد کوچک نیز شناسایی می‌کند. متاستاز همان‌طور که قبلا اشاره شد به مهاجرت سلول‌های سرطانی از یک بافت به بافت‌های دیگر گفته می‌شود. این تکنولوژی علاوه بر تشخیص نرخ متاستازهای کوچک، زمان مورد نیاز برای تشخیص را به حدود یک دقیقه رسانده است [۴۱].



شکل ۳-۱: نمونه ای از تشخیص سرطان سینه توسط الگوریتم LYNA. نقاط قرمز رنگ در تصویر سمت راست نشان از تومور است [۴۱].

---

[۹] Lymph Node Assistant

[۱۰] Pathology

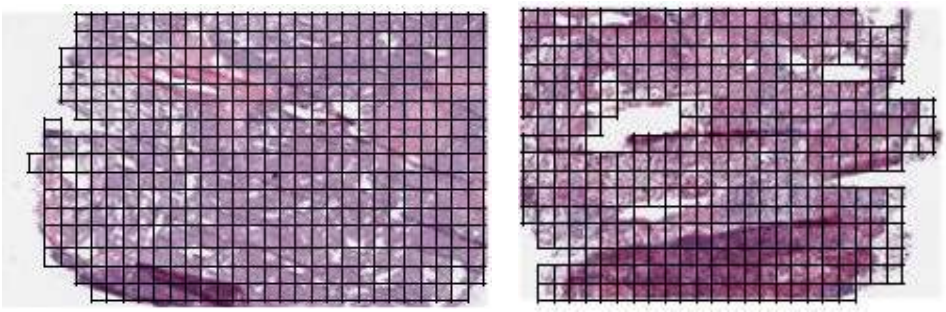
[۱۱] Google AI

[۱۲] متاستاز یا Metastasis به گسترش و مهاجرت سلول‌های سرطانی از یک بافت به بافت‌های دیگر، گفته می‌شود.

۳-۳ یادگیری عمیق و تشخیص سرطان ریه

بررسی‌های بصری از تصاویر آسیب‌شناسی یکی از روش‌های اصلی مورد استفاده آسیب‌شناسان[۱۳] برای فهمیدن این‌که تومور ریه در چه مرحله‌ای است، می‌باشد. شایع‌ترین نوع سرطان‌های ریه LUAD[۱۴] و LUSC[۱۵] می‌باشند و تمایز بین آن‌ها نیاز به بررسی بصری توسط آسیب‌شناس باتجربه دارد.

در مقاله [۴۲]، یک شبکه عصبی کانولوشن عمیق برای تصاویر تهیه شده از Atlas of Genome Cancer آموزش داده شده است تا به طور دقیق و به صورت خودکار LUAD، LUSC بافت نرمال ریه را طبقه‌بندی کرد. در این مقاله تعداد ۶۰۹ تصویر از LUSC و تعداد ۵۶۷ تصویر از LUAC و ۴۵۹ تصویر نرمال برای استفاده در شبکه عمیق استفاده شده است. روش کار آن‌ها بدین صورت بوده است که قبل از ورود تصویر به شبکه پیش‌پردازش‌هایی را روی آن صورت دهند. این پیش پردازش‌ها شامل حذف تصویر زمینه و موزائیک‌بندی باقیمانده تصاویر بوده است.در نهایت تعداد ۷۰ درصد از داده‌ها را برای آموزش و ۱۵ درصد را برای اعتبارسنجی و ۱۵ درصد باقیمانده را نیز برای آزمون در نظر گرفتند.



شکل ۳-۲: حذف نویز و موزائیک‌بندی تصویر [۴۲]

[۱۳] Pathologist
[۱۴] Lung tumor adenocarcinoma
[۱۵] Squamous cell carcinoma lung tumor

عملکرد این روش که با نام DeepPATH نام‌گذاری شده است با عمکرد آسیب‌شناسان قابل مقایسه است، به طوری که سطح متوسط زیر منحنی AUC[۱۶] ۰,۹۷ است. همان‌طور که از یادگیری ماشین می‌دانیم هر چه سطح متوسط زیر منحنی نزدیک به یک باشد کارایی طبقه‌بند مطلوب‌تر است. معیار AUC برای تعیین میزان کارایی یک دسته بند استفاده می‌شود. نمودار AUC نشان دهنده سطح زیر نمودارROC[۱۷]می‌باشد که هر چه مقدار این عدد مربوط به یک دسته بند بزرگ‌تر باشد کارایی نهایی دسته‌بند مطلوب‌تر ارزیابی می‌شود. نمودار ROC روشی برای بررسی کارایی دسته‌بندها می‌باشد. در واقع منحنی‌های ROC منحنی‌های دو بعدی هستند که در آن‌ها نرخ تشخیص صحیح دسته مثبت روی محور y و بطور مشابه نرخ تشخیص غلط دسته منفی روی محور x رسم می‌شوند. به بیان دیگر یک منحنی ROC مصالحه نسبی میان سودها و هزینه‌ها را نشان می‌دهد.

مدل DeepPATH بر روی مجموعه‌های مستقل از بافت‌های یخ زده، بافت‌های پارافین و بیوپسی‌های[۱۸] ثابت شده با فرمالین مورد تایید قرار گرفت. مدل DeepPATH بر روی ده ژن متداول در LUAG آموزش داده شد. سپس DeepPATH دریافت که شش ژن STK۱۱ ، EGFR، FAT۱، SETBP۱، KRAS و TP۵۳ را می‌توان از تصاویر آسیب‌شناسی پیش‌بینی کرد، که AUC از ۰,۷۳۳ به ۰,۸۵۶ برآورد شده است. این یافته‌ها حاکی از آن است که مدل‌های یادگیری عمیق می‌توانند متخصصان آسیب‌شناس را در تشخیص جهش‌های ژنی کمک کنند.

۴-۳ یادگیری عمیق و تشخیص سرطان پوست

سرطان پوست، شایع‌ترین نوع بدخیم، عمدتا به صورت بصری و توسط متخصص تشخیص داده می‌شود.

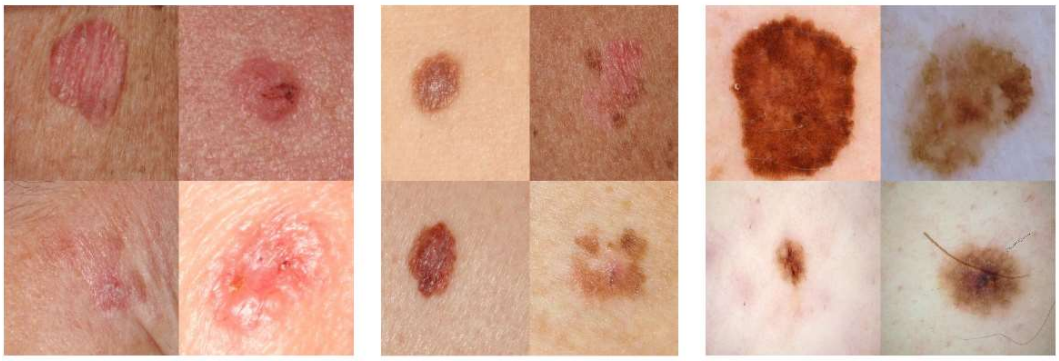
---

[۱۶] Area Under Curve

[۱۷] Receiver Operating Characteristic

[۱۸] Biopsie

طبقه‌بندی خودکار ضایعات پوستی با استفاده از تصاویر یک عمل چالش برانگیز به دلیل تنوع ریزدانه در ضایعات پوستی است. شبکه‌های عصبی کانولوشن عمیق نشان دادند که پتانسیل تشخیص سرطان پوست را با کمک تصاویر گرفته شده دارند. در مقاله [۴۳] طبقه‌بندی ضایعات پوستی با استفاده از شبکه عصبی کانولوشن بیان شده است.



شکل ۳-۳: نمونه‌ای از تصاویر بالینی از بیماران مبتلا به سرطان پوست [۴۳]

شکل ۳-۳ نمونه ای از تصاویر استفاده شده که دچار این نوع سرطان شده‌اند را نشان می‌دهد. این شبکه تنها با کمک پیکسل‌ها و برچسب‌های بیماری به تشخیص سرطان می‌پردازد. این مقاله یک شبکه کانولوشن را با استفاده از مجموعه‌ای از ۱۲۹،۴۵۰ تصویر بالینی آموزش می‌دهد، که شامل ۲۰۳۲ بیماری مختلف است. با کمک شبکه کانولوشن الگوریتم تفاوت بین تصاویر سرطانی و غیرسرطانی را تشخیص می‌دهد. در نهایت بر روی داده‌های آزمون مشخص می‌کند که آیا تصویر ورودی دچار سرطان پوست شده است یا خیر. پس از انجام آزمایش‌های گسترده عملکرد خود را در برابر عملکرد ۲۱ متخصص پوست مورد مقایسه قرار می‌دهد. در انتها نشان داده شده است که عملکرد شبکه کانولوشن با عملکرد تمامی متخصصان، قابل مقایسه است.

۳-۵ نتیجه‌گیری

همان‌طور که مشاهده کردیم، هوش مصنوعی با کمک یادگیری عمیق می‌تواند با دقت بسیار بالایی به

تشخیص پزشکی کمک کند. این تشخیص‌ها گاهی از پیش‌بینی‌های متخصصان نیز پیشی می‌گیرد. شاید بتوان گفت با استفاده از شبکه‌های عصبی عمیق و با کمک تکنولوژی‌های می‌توان به طور بالقوه دسترسی پزشکان به خارج از کلینیک‌ها را گسترش دهند. در نتیجه این امر کمک قابل ملاحظه‌ای به کاهش هزینه‌ها و تشخیص‌های موثر می‌کند.

فصل ٤ : معرفی روش پیشنهادی

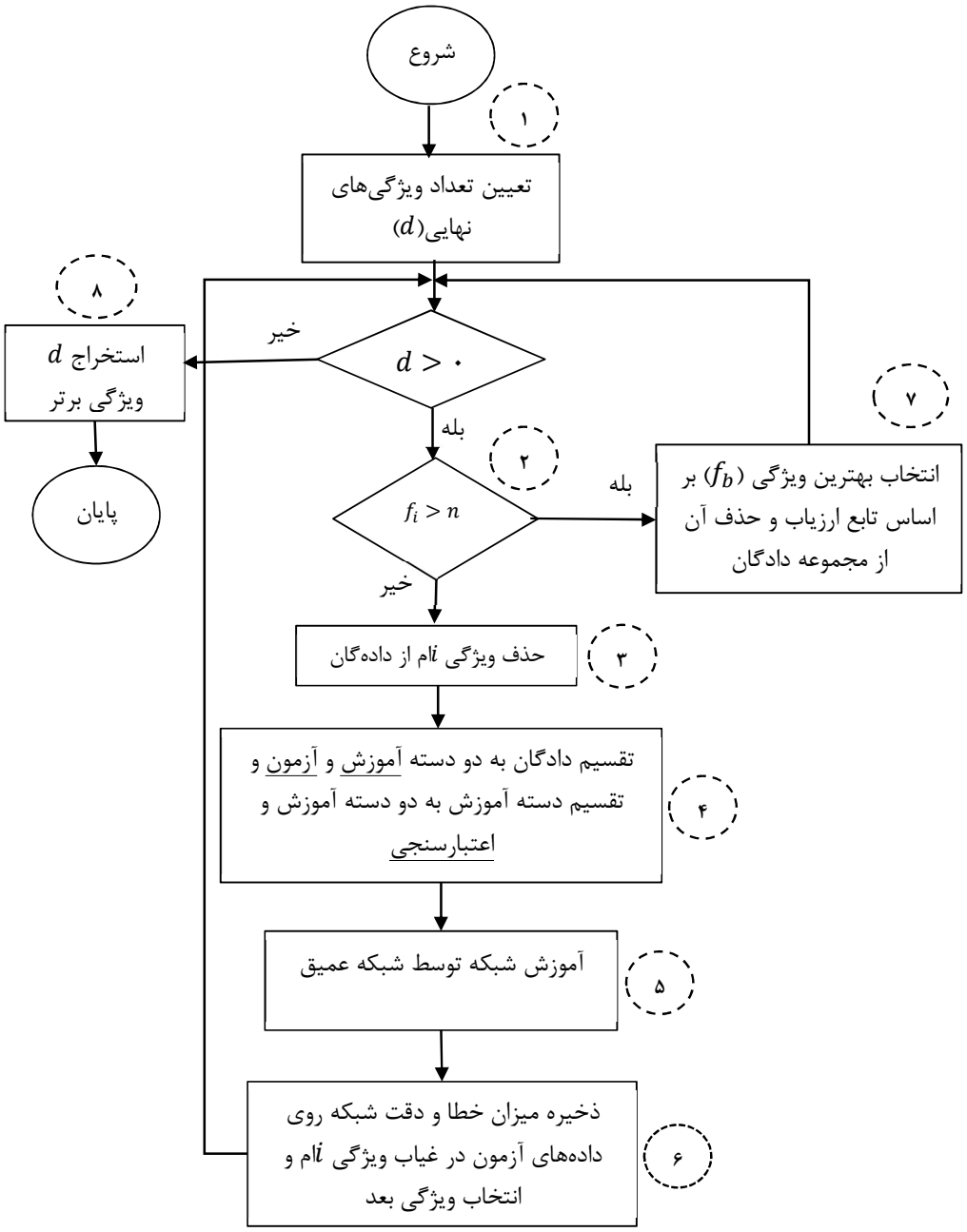
۴-۱ مقدمه

در این فصل ابتدا الگوریتم پیشنهادی با نام DeepCNV معرفی می‌شود. در ابتدا فلوچارتی از الگوریتم پیشنهادی ارائه شده و سپس به شرح گام به گام مراحل آن خواهیم پرداخت. در انتهای فصل مشکلات پیش آمده در آموزش مانند مشکل بیش‌برازش معرفی می‌شود و راهکار مناسب برای حل آن ارائه می‌شود.

۴-۲ الگوریتم DeepCNV

شکل ۴-۱ فلوچارت الگوریتم DeepCNV را نشان می‌دهد. این الگوریتم تعداد مشخصی ویژگی (d) را در خروجی از مجموعه دادگان برمی‌گرداند. در ابتدا الگوریتم SBS عملیات مورد نظر خود را با مجموعه تمام ویژگی‌ها شروع کرده، سپس در هر تکرار یک ویژگی با استفاده از تابع ارزیاب مورد استفاده، به مجموعه جواب اضافه می‌کند. الگوریتم این عمل را تکرار کرده تا زمانی که تعداد ویژگی مشخص شده (d) انتخاب شود. تابع ارزیابی مورد استفاده معیار بیشترین میزان خطا و کمترین میزان درصد دقت در طبقه‌بند است. همچنین در مجموعه دادگان با مجموعه‌ای متشکل از n ویژگی روبرو بوده که در آن میزان خطا و درصد دقت شبکه در غیاب هر کدام از آن ویژگی‌ها با کمک شبکه عمیق محاسبه می‌شود. بعد از بررسی تمام ویژگی‌ها، بهترین ویژگی که دارای بیشترین میزان خطا و کمترین میزان دقت بوده، انتخاب شده و در نهایت آن را از مجموعه دادگان حذف می‌نماییم. الگوریتم این ویژگی را به عنوان ویژگی برتر ذخیره می‌کند.

الگوریتم DeepCNV از ۸ گام اصلی تشکیل می‌شود. ادامه این بخش شرح جزئیات هر کدام از گام‌ها را شرح می‌دهد. هر کدام از گام‌های توضیح داده شده با دایره نقطه چین در شکل ۴-۱ نشان داده شده‌اند.



شکل ٤-۱: فلوچارت الگوریتم DeepCNV

گام ۱: تعیین تعداد ویژگی اولیه

مساله انتخاب ویژگی، یکی از مسائلی است که در بحث یادگیری ماشین و همچنین شناسایی آماری

الگو مطرح است. این مساله در بسیاری از کاربردها مانند طبقه‌بندی اهمیت فراوانی دارد، زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد، که بسیاری از آن‌ها یا غیرقابل استفاده بوده یا بار اطلاعاتی مفیدی ندارند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد. علاوه بر این باعث می‌شود تا اطلاعات غیرمفید زیادی را به همراه داده‌های مفید ذخیره کنیم.

برای شناسایی ویژگی‌هایی از دادگان که بیشترین تاثیر را در طبقه‌بندی داشتند تعداد $\boldsymbol{d}$ ویژگی انتخاب شد. مقدار $\boldsymbol{d}$ مشخص کننده تعداد ویژگی‌هایی از دادگان است که با کمک الگوریتم DeepCNV استخراج می‌شود. تعداد $\boldsymbol{d}$ ویژگی استخراج شده، خروجی‌های الگوریتم DeepCNV بوده که بیشترین تاثیر را در طبقه‌بندی داشتند. این بخش از الگوریتم بخشی از SBS است که $\boldsymbol{d}$ در آن، مشخص کننده تعداد ویژگی‌هایی است که SBS برمی‌گرداند.

گام ۲: بررسی تمام $\boldsymbol{n}$ ویژگی

به ازای تمام ویژگی‌ها با اندازه n میزان خطا و میزان دقت شبکه را برای یک ویژگی محاسبه می‌کنیم. همان‌طور که گفته شد الگوریتم SBS کارش را با مجموعه‌ای شامل تمام ویژگی‌ها شروع کرده و در هر بار تکرار الگوریتم، ویژگی که بوسیله تابع ارزیابی انتخاب می‌شود، از مجموعه ویژگی خود حذف می‌کند. الگوریتم این کار را تا زمانی ادامه می‌دهد که تعداد d ویژگی به عنوان خروجی برنامه استخراج شود.

گام ۳: حذف ویژگی $\boldsymbol{i}$ام از دادگان

ویژگی iام در عمل همان ویژگی است که باید در الگوریتم SBS مورد بررسی قرار گیرد. برای این‌که بتوان بهترین ویژگی را پیدا کرد، و به اهمیت ویژگی iام پی برد، اقدام به حذف این ویژگی می‌کنیم. با حذف این ویژگی میزان خطا و درصد دقت شبکه را محاسبه می‌کنیم. هر چقدر خطای شبکه با حذف این ویژگی بیشتر باشد و میزان دقت شبکه پایین‌تر، می‌توان به با اهمیت بودن این ویژگی در شبکه پی

برد به طوری که حذف این ویژگی خطای شبکه را افزایش می‌دهد. در واقع این عمل همان تابع ارزیابی است که در الگوریتم SBS استفاده می‌شود. با حذف این ویژگی دادگان خود را به دو دسته داده‌های آموزشی و داده‌های آزمون تقیسم می‌کنیم و برای بررسی میزان خطا و دقت شبکه، آن را به شبکه عمیق وارد می‌کنیم.

گام ۴: تشکیل مجموعه آموزش و آزمون

جهت آموزش و آزمون شبکه، برای هر نمونه بیمار، تعداد ۲۴۱۷۴ ویژگی از دادگان خود را در اختیار داشتیم. تعداد ویژگی‌ها، که در واقع همان ژن‌های دارای اختلال CNV بوده را به عنوان لایه ورودی به شبکه عمیق اعمال می‌کنیم. الگوریتم DeepCNV از روی داده‌های آموزش یادگیری را انجام می‌دهد. همچنین از روی داده‌های آزمون می‌توان فهمید که الگوریتم و مدل ساخته شده توسط آن، با چه دقتی طبقه‌بندی را انجام می‌دهد. وقتی الگوریتم عملیات یادگیری را انجام داد و یک مدل را از روی این داده‌ها ساخت، حال می‌توان از روی این مدل، عملیات داده‌کاوی را بر روی داده‌های جدید انجام دهیم.

در الگوریتم DeepCNV برای آموزش شبکه از ۹۰ درصد نمونه‌ها، جهت آموزش شبکه استفاده شد و ۱۰ درصد باقیمانده برای آزمون شبکه استفاده کردیم. در ادامه با مشخصات داده‌های آموزش و آزمون الگوریتم پیشنهادی آشنا می‌شویم.

- **داده آموزشی**

تعداد ۳۱۳۲ نمونه از بیماران، به عنوان داده آموزشی در نظر گرفته شده است. از این تعداد، فقط ۸۰ درصد در تغییر وزن‌های شبکه تاثیر داده شد و ۲۰ درصد باقیمانده برای مرحله اعتبارسنجی استفاده گردید. دلیل آن به خاطر کاهش بیش‌برازش[۱۹] و جلوگیری از حفظ کردن داده‌های آموزش توسط شبکه

[۱۹] Overfitting

عمیق است. با توجه به این‌که تعداد مراحل اجرای الگوریتم DeepCNV برابر ۱۰۰۰ درنظر گرفته شد، بنابراین به تعداد ۱۰۰۰ بار، داده‌های ورودی را به شبکه اعمال کرده تا بهترین مقادیر وزن‌ها را داشته باشیم.

- **داده آزمون**

برای این‌که ارزیابی درستی از شبکه خود داشته باشیم، تعداد ۱۰ درصد ار نمونه‌ها را به صورت دیده نشده[۱] در نظر گرفتیم. با کمک این داده‌ها در شبکه عمیق، می‌توان ارزیابی مناسبی از عملکرد شبکه بدست آورد.

گام ۵: آموزش شبکه

در یادگیری ماشین با الگوریتم‌های متعدد یادگیری روبرو هستیم. از جمله این الگوریتم‌ها می‌توان به الگوریتم SVM، KNN یا شبکه‌های عصبی اشاره نمود. معمولا در یادگیری ماشین دو روش زیر در زمینه یادگیری مورد استفاده قرار می‌گیرد:

یادگیری تحت نظارت[۲]: نوعی از یادگیری مربوط به یادگیری ماشین است که در آن ورودی و خروجی مشخص بوده و به اصطلاح ناظری وجود دارد که اطلاعاتی را در اختیار یادگیرنده قرار می‌دهد، و به این ترتیب سیستم سعی می‌کند تا تابعی را از ورودی به خروجی بیابد.

یادگیری بدون نظارت[۳]: بر خلاف یادگیری نظارت شده، داده‌های مشخصی از قبل وجود ندارد و هدف هدف ارتباط ورودی و خروجی نیست، بلکه تنها طبقه‌بندی آن‌ها مهم است و این یادگیرنده است که

---

[۱] unseen

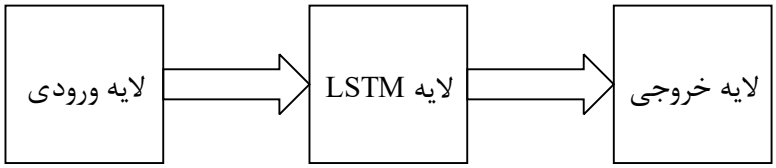
[۲] Supervised Learning

[۳] Unsupervised Learning

باید در مجموعه داده‌ها به دنبال ساختاری خاص بگردد. به بیان دیگر در این شیوه، یادگیری بر روی داده‌های بدون برچسب انجام می‌شود و سیستم خودش باید الگوهای پنهان در داده‌ها را کشف کند.

همان‌طور که در فصل ۵ از تحقیق خواهیم دید، در الگوریتم پیشنهادی از هر دو شبکه عصبی MLP و شبکه عمیق LSTM استفاده شد. اما به دلیل برتری LSTM نسبت به MLP (دلیل برتری آن بیان خواهد شد) از این الگوریتم جهت آموزش شبکه استفاده گردید. در ادامه توضیحاتی در مورد شبکه عمیق LSTM استفاده شده در الگوریتم DeepCNV می‌دهیم. ضمنا یادگیری در این شبکه به صورت نظارت شده است.

همان‌طور که در شکل ۴-۲ نشان داده شده است بخش آموزش الگوریتم از ۳ لایه تشکیل شده است. در ادامه این بخش به شرح لایه های شبکه می‌پردازیم.



شکل ٤-۲: شمایی از لایه‌های الگوریتم DeepCNV

در ادامه به شرح هر کدام از لایه‌های شبکه عمیق مورد استفاده می‌پردازیم.

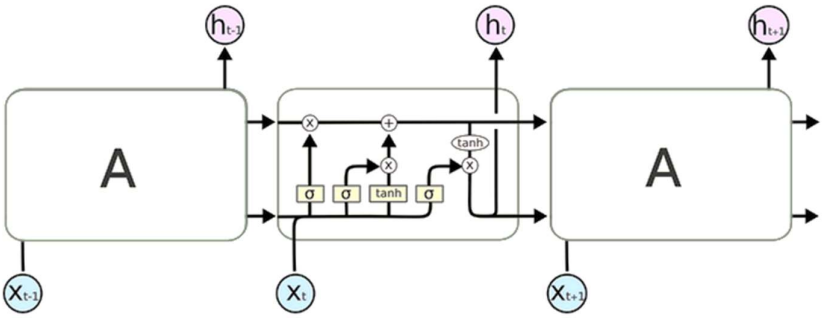
- **لایه ورودی**

تعداد نرون‌های این لایه برابر تعداد ویژگی‌های هر بیمار است. به طور مثال برای یک نمونه، تعداد ۲۴۱۷۴ ویژگی داریم. مقادیر سطح CNV هر کدام از نرون‌های لایه ورودی به تک تک نرون‌های لایه LSTM متصل است.

- **لایه LSTM**

این لایه دارای ۱۰۰ نرون است. هر کدام از نرون‌ها نشان دهنده یک واحد LSTM است. شکل ۴-۳

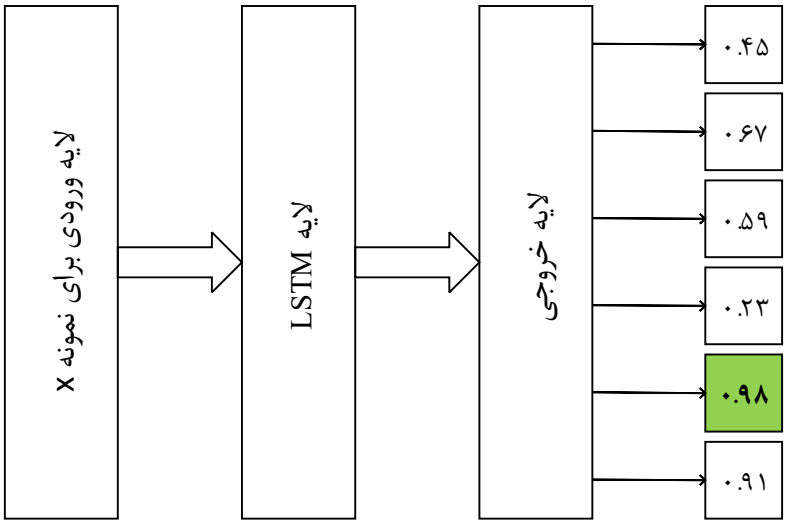
نمونه‌ای از این واحدها با ۳ واحد LSTM را نشان می‌دهد [۳۲]. ورودی این لایه همان نرون‌های لایه ورودی بوده و خروجی حاصل از آن به لایه خروجی متصل می‌گردد.



شکل ۴-۳: نمونه‌ای از ۳ واحد LSTM [۳۱]

- **لایه خروجی**

این لایه دارای ۶ نرون است. هر کدام از نرون‌ها در این لایه معرف یک نمونه از کلاس سرطان می‌باشند. هر کدام از واحدهای لایه LSTM به هر کدام از لایه‌های خروجی، متصل می‌گردد. انتخاب کلاس مناسب برای هر نمونه بدین صورت است که، برای یک ورودی مشخص، مقدار خروجی هر کدام از نرون‌ها بیشتر باشد، نمونه ورودی به آن کلاس سرطان تعلق می‌گیرد. برای تک تک نمونه‌های آموزش، خروجی به همین ترتیب محاسبه می‌شود و وزن‌های شبکه بروزرسانی می‌گردد. شکل ۴-۴ نمونه‌ای از این عمل را نشان می‌دهد.



شکل ٤-٤: نمونه‌ای از نحوه انتساب یک ورودی به یک کلاس

گام ۶: ذخیره میزان خطا و دقت شبکه (تابع ارزیاب)

برای این‌که بتوان میزان تاثیر ویژگی iام را در شبکه عمیق داشته باشیم باید در پایان آموزش شبکه و در غیاب ویژگی iام، میزان خطا و درصد دقت شبکه را ذخیره کنیم. بعد از ذخیره میزان خطا و دقت، ویژگی بعدی را انتخاب نموده و تا زمانی که تمام n ویژگی بررسی نشده است، این عمل را برای تمام ویژگی‌ها تکرار می‌کنیم.

گام ۷: انتخاب بهترین ویژگی با کمک تابع ارزیاب

همان‌طور که در گام ۶ ملاحظه کردیم، به ازای تمام n ویژگی میزان خطا و درصد دقت شبکه را محاسبه نمودیم. این عملیات را برای تمام ویژگی‌ها انجام دادیم. در این گام بهترین ویژگی (f_b) را با کمک معیار بیشترین میزان خطا و کمترین میزان درصد دقت محاسبه می‌کنیم. در نهایت این ویژگی را به مجموعه ویژگی‌های برتر خود اضافه کرده و از مجموعه دادگان حذف می‌نماییم.

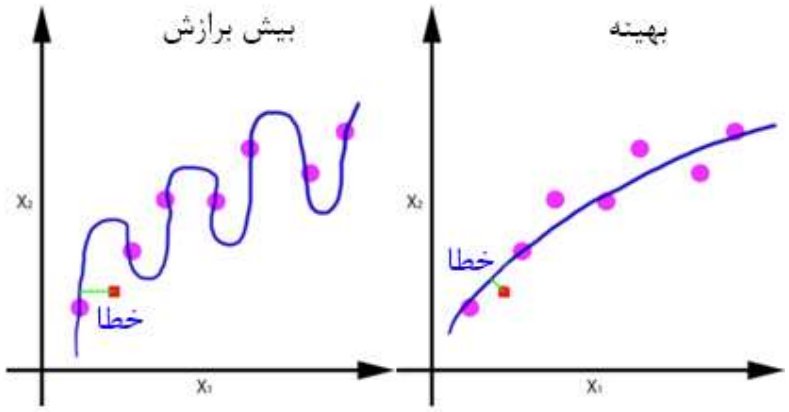
گام ۸: انتخاب d ویژگی برتر

در مراحل قبل تعداد d ویژگی را از دادگان خود حذف نمودیم. حال می‌توان از این d ویژگی به عنوان

خروجی برنامه استفاده کرد زیرا در غیاب هر کدام از آن‌ها، شبکه دارای بیشترین میزان خطا و کم‌ترین میزان دقت را دارا بوده است.

۴-۳ مشکل بیش‌برازش

اگر فرایند آموزش خیلی طولانی باشد، بیش‌برازش پیش می‌آید. بیش‌برازش یعنی شبکه خیلی به داده‌های آموزش حساس می‌شود و اگر داده‌های جدید کمی متفاوت باشند، نتیجه‌ی دقیقی حاصل نمی‌شود [۴۴]. به طورکلی بیش‌برازش همان‌طور که در شکل ۴-۵ نشان داده شده است، این است که شبکه داده‌ها را حفظ نکند [۴۵].

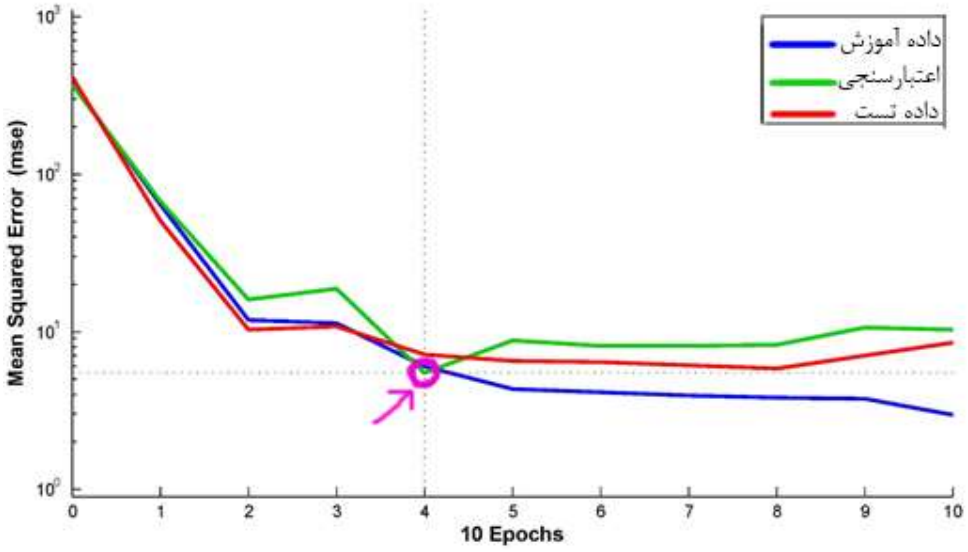


شکل ۴-۵: نمایش بیش‌برازش [۴۴]

همان‌طور که در شکل ۴-۵ مشخص است، شبکه در سمت چپ داده‌ها را به خاطر می‌سپارد. اما در شکل سمت راست که بیش برازش کم است، و شبکه بهترین خط را تخمین زده است، شبکه داده‌های آموزش را حفظ نمی‌کند. بنابراین اگه نمونه جدیدی وارد شود، خطای شبکه سمت چپ بیشتر از شبکه سمت راست است. در این تحقیق برای حل مشکل بیش‌برازش در شبکه عمیق از دو راه‌کار زیر استفاده گردید:

۱- استفاده از داده‌های اعتبارسنجی

اهمیت داده‌های اعتبارسنجی این است که از وقوع بیش‌برازش جلوگیری می‌کند. زمانی که فرایند آموزش توسط داده‌های بخش آموزش انجام می‌گیرد، توسط داده‌های اعتبارسنجی بررسی می‌کنیم که سیستم خیلی وابسته به داده‌های آموزش نباشد. در شکل ۴-۶ نمونه فرآیند آموزش و میزان‌خطاهای آموزش، اعتبارسنجی و آزمایش در هر مرحله تکرار از الگوریتم را نشان داده شده است. همان‌طور که مشاهده می‌شود هر چه تعداد مراحل آموزش شبکه بیشتر شود، میزان خطای آموزش کاهش می‌یابد اما به نقطه‌ای می‌رسیم که کم کم خطای اعتبارسنجی افزایش می‌یابد. این نقطه همان جایی است که ممکن است از آن به بعد بیش‌برازش اتفاق افتد. به همین دلیل فرایند آموزش در آن متوقف می‌شود. در شکل ۴-۶ محلی که بهتر است آموزش متوقف شود نشان داده شده است [۴۶].



شکل ۴-۶: بهترین زمان برای توقف آموزش برای جلوگیری از بیش‌برازش [۴۳]

۲- استفاده از لایه Dropout

بعد از پیاده‌سازی و مشاهده نتایج الگوریتم، به این نتیجه رسیدیم، برای کاهش بیش‌برازش در شبکه LSTM، علاوه بر انتخاب داده‌های اعتبارسنجی از لایه میانی به نام Dropout استفاده کنیم

[۴۷]. لایه Dropout تضمین می‌کند که بیش‌برازش به حداقل برسد. لایه Dropout این کار را با انتخاب نرون‌ها به صورت تصادفی و نادیده گرفتن آن‌ها در طول آموزش انجام می‌دهد. در واقع Dropout، یک روش برای جلوگیری از بیش‌برازش بر روی داده‌های آموزشی با حذف تصادفی نرون‌ها در یک شبکه‌ی عصبی است. در عمل، نرون‌ها با احتمال p حذف یا با احتمال $۱ - p$ حفظ می‌شوند. در آزمایش‌های ما، Dropout برابر ۲۰ درصد در نظر گرفته شده است. بدین معنی که ۲۰ درصد از نرون‌ها به طور تصادفی در هر مرحله از الگوریتم در بروزرسانی وزن‌ها دخالت ندارند.

۴-۴ نتیجه‌گیری

در این فصل از تحقیق الگوریتم پیشنهادی DeepCNV معرفی شد. بعد از معرفی فلوچارت الگوریتم، تمامی گام‌های آن مورد بررسی قرار گرفت. برخی از مشکلات پیش آمده در پیاده‌سازی الگوریتم مانند مشکل بیش‌برازش بیان شده و راه‌حل‌های مناسب ارائه شد.

فصل ۵ : ارزیابی روش پیشنهادی

۵-۱ مقدمه

در فصل قبل با الگوریتم DeepCNV آشنا شدیم و گام‌های آن را به طور کامل شرح دادیم. همچنین با مشکلات پیش آمده مانند مشکل بیش‌برازش آشنا شده و راه‌حل‌های مناسب ارائه شد. در این فصل ابتدا دادگان مورد استفاده در این تحقیق معرفی می‌شود. برای این منظور هر کدام از سرطان‌های موجود در این دادگان را به اختصار شرح خواهیم داد. سپس به پیاده‌سازی الگوریتم پیشنهادی می‌پردازیم. پیاده‌سازی طبقه‌بند را با دو شبکه MLP و شبکه عمیق LSTM انجام داده و تفاوت‌های آنها را مورد بررسی قرار می‌دهیم. همچنین دلیل برتری LSTM نسبت به MLP را بیان می‌کنیم. در نهایت خروجی الگوریتم DeepCNV و نتایج حاصل از آن را مورد ارزیابی قرار خواهیم داد.

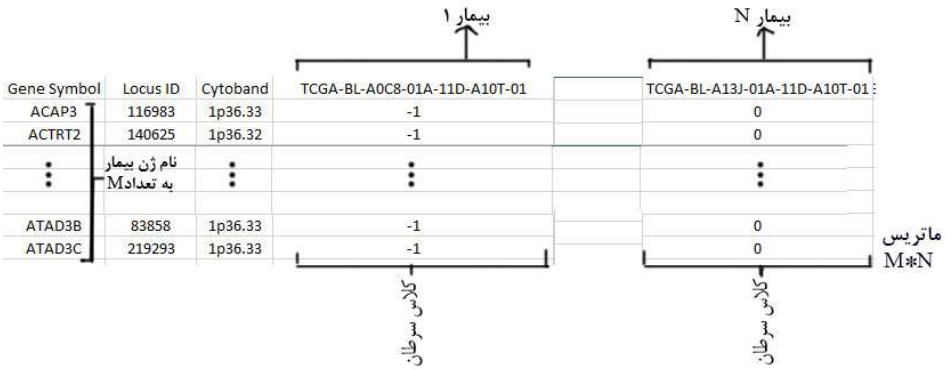
۵-۲ معرفی دادگان

در این تحقیق از دادگان cBioPortal [۱۵] استفاده شد. این دادگان اطلاعات CNV از تعداد ۱۱ نوع مختلف سرطان را در خود ذخیره کرده است. از این تعداد، ۶ سرطان که اطلاعات بیشتر از ۴۰۰ نمونه بیمار را در اختیار داشت، انتخاب شد. اطلاعات این ۶ نوع سرطان در جدول ۵-۲ نشان داده شده است که از یک ماتریس M در N بُعدی تشکیل می‌شود. در این ماتریس، M تعداد ژن‌ها با اختلال CNV است و N نشان دهنده تعداد نمونه‌ها (بیماران) است. بنابراین در هر ستون این ماتریس یک بیمار و در هر ردیف آن یک ژن قرار گرفته است. در این دادگان تشخیص اختلالات CNV با کمک روش ریزآرایه[۱] صورت گرفته است. ریزآرایه، شامل فناوری بررسی فعالیت ده‌ها، صدها و هزاران ژن یا پروتئین در یک سطح کوچک (یک ریزآرایه در حد چند سانتی‌متر مربع) است. از ریزآرایه جهت مقایسه، مشابهت و بررسی تغییر، کاهش، افزایش و عدم تغییر در ساختار و فعالیت ژن‌ها یا پروتئین‌های نمونه

[۱] Microarrays

بیمار با نمونه‌های سالم استفاده می‌شود. شکل ۵-۱ نمونه‌ای از این دادگان را نشان می‌دهد. لازم به ذکر است که در این پایان‌نامه از یادگیری با نظارت استفاده می‌کنیم، زیرا کلاس‌ها در دادگان مورد استفاده مشخص هستند. مقدار هر عنصر در این ماتریس یکی از ۵ عدد فهرست شده در

جدول ۵-۱ است.



شکل ۵-۱: بخشی از داده‌های دادگان مورد استفاده

جدول ۵-۱: مقادیر هر عنصر در دادگان

۲-	Deep Loss
۱-	Shallow Loss
۰	Diploid
۱+	Low-Level gain
۲+	High-Level amplification

پس از انتخاب دادگان مربوطه، به تعریف و تعیین ویژگی‌ها می‌پردازیم. با مشخص شدن ویژگی‌ها، ورودی جهت انجام پیش‌پردازش داده‌ها امکان پذیر می‌شود. بنابراین می‌توان از آن به عنوان ماتریس ورودی شبکه عمیق استفاده نمود. در اینجا، مقادیر سطح CNV هر کدام از ژن‌ها را به عنوان ورودی به شبکه در نظر می‌گیریم که مقادیر آن در

جدول **۵-۱** مشخص شده است. اکنون تعداد ویژگی‌ها بیش از ۲۴۰۰۰ است. جدول ۵-۲ جزئیات این دادگان را نشان می‌دهد. همچنین ذکر این نکته لازم است که کلاس‌ها در این دادگان به صورت رشته‌ای می‌باشند که با کمک بردار one-hot به مقادیر باینری جهت طبقه‌بندی تبدیل شده‌اند.

جدول ۵-۲: اطلاعات کامل از دادگان مورد استفاده

نوع سرطان	تعداد ویژگی‌ها (ژن‌ها)	تعداد نمونه (بیماران)
سرطان سینه	۲۴۱۷۴	۸۴۷
سرطان روده بزرگ	۲۴۱۷۴	۵۷۵
گلیوبلاستوما	۲۴۱۷۴	۵۶۳
سرطان کلیه	۲۴۱۷۴	۴۹۰
سرطان تخمدان	۲۴۱۷۴	۵۶۲
سرطان رحم	۲۴۱۷۴	۴۴۳
	تعداد ویژگی: ۲۴۱۷۴	مجموع: ۳۴۸۰

- **سرطان سینه (BRCA)**

در این سرطان، سلول‌های سرطانی در نرمه‌های کوچک یا غدد شیری به وجود می‌آیند و سپس به تمام بدن منتقل می‌شوند. در سرطان سینه تهاجمی، ابتدا تومور در قفسه سینه یا غدد سینه ایجاد می شود و بعد از آن به داخل بافت‌های سینه سرایت می‌کند. در نهایت تومور می‌تواند به غدد لنفاوی[1] نزدیک‌تر و بعد به بقیه بخش‌های بدن سرایت کند. در صورت ابتلا به سرطان سینه مهاجم، بانوان حس می‌کنند که پوست سینه‌شان ضخیم شده است [۴۸].

- **سرطان روده بزرگ (COAD)**

[1] این غدد در واقع محل تجمع سلول‌های دفاعی بدن هستند.

این سرطان مربوط به روده بزرگ، قسمت پایین دستگاه گوارش می‌شود. این سرطان چند سانتی‌متر آخر روده بزرگ را مبتلا می‌کند. در اغلب موارد سرطان روده بزرگ به شکل انبوهی از سلول‌های غیرسرطانی به نام پولیپ[۲]، آغاز شده و به مرور زمان برخی از این پولیپ‌ها تبدیل به سرطان می‌شوند. این پولیپ‌ها کوچک هستند، اما در عین حال علائمی را در بدن ایجاد می‌کنند. از این جهت پزشک آزمایش‌های غربال‌گری منظم را توصیه می‌کند تا از بروز سرطان روده بزرگ پیشگیری شود [۴۹].

- **گلیوبلاستوما (GBM)**

تومور GBM یا گلیوبلاستوما شایع‌ترین نوع تومور بدخیم اولیه سیستم عصبی است که در نخاع یا مغز بروز می‌کند. این سرطان بدخیم است و به گونه‌ای تهاجمی گسترش می‌یابد. این سرطان اغلب در افراد بزرگسال بین ۴۵ تا ۷۰ سال دیده می‌شود. برای مطالعه بیشتر به [۵۰-۵۲] مراجعه شود.

- **سرطان کلیه (KIRC)**

سرطان کلیه، سرطانی است که منشاء آن در کلیه‌هاست. کلیه دو اندام لوبیا شکل است، که هر کدام تقریبا به اندازه مشخص است. کلیه پشت اندام‌های شکمی واقع شده و هر کلیه در یک طرف ستون فقرات قرار گرفته است. در بزرگسالان، شایع‌ترین نوع سرطان کلیه کارسینوم سلول کلیوی است. انواع دیگر سرطان کلیه به ندرت رخ می‌دهد. سرطان سلول کلیوی (RCC)[۳] همیشه، به خصوص در مراحل ابتدایی علائم مشخصی ندارد. در نتیجه، سرطان تا رسیدن به مرحله پیشرفته قابل شناسایی نیست. درمان سرطان سلول کلیوی ممکن است شامل برداشتن قسمتی یا تمام کلیه شود. در بعضی افراد، از دارو جهت کند کردن پیشرفت سرطان استفاده می‌شود. برای افراد مبتلا به کارسینوم سلول کلیوی، در

---

[۲] پولیپ یا Polyp، به صورت رشد غیرطبیعی توده‌های خوش‌خیم در جدار اعضای بدن تعریف می‌شود مانند پولیپ روده، معده، بینی و پولیپ رحم. پولیپ، سرطان نیست ولی سرطان می‌تواند منشا آن باشد.

[۳] Renal Cell Carcinoma

صورت عدم رفع کامل بیماری پس از جراحی، از درمان‌های دارویی می‌توان استفاده کرد. جراحی به منظور برداشتن کلیه یا نواحی خارج از کلیه (جایی که سرطان گسترش و متاستاز داده است) را می‌توان قبل از درمان دارویی انجام داد. در افراد دارای RCC پیشرفته، جراحی، سبب معالجه سرطان نشده بلکه علائم آن را کاهش می‌دهد. برای مطالعه بیشتر در مورد این بیماری به مقالات [۵۳-۵۵] رجوع شود.

- **سرطان تخمدان (OV)**

سرطان تخمدان نوعی از سرطان است که از تخمدان شروع می‌شود [۵۶]. این بیماری منجر به ایجاد سلول‌های غیرعادی شده که توانایی حمله یا گسترش به دیگر بخش‌های بدن را دارند. ممکن است در آغاز این فرایند علائم مبهم یا غیرآشکار باشند، اما با پیشرفت سرطان احتمال بروز علائم بیشتر شود. طبق آمار در سال ۲۰۱۲، این بیماری به عنوان هفتمین نوع رایج سرطان و هشتمین عامل مرگ ناشی از سرطان در زنان شناخته شد [۵۷].

- **سرطان رحم (UCEC)**

نوعی سرطان است که در رحم ایجاد می‌شود. این سرطان هنگامی رخ داده که سلول‌های آندومتر[۴] دست‌خوش تغییر بدخیمی شده و بدون مکانیسم‌های کنترلی محدود کننده رشد، و شروع به تکثیر و افزایش می‌کنند. رشد و تکثیر این نوع سلول موجب شکل‌گیری تومور سرطانی و بدخیم می‌شود. سرطان به این دلیل خطرناک است که سلول‌های سالم را با اشغال فضا و مصرف اکسیژن و مواد مغذی که برای بقا نیاز دارند، از بین می‌برد. سرطان همچنین ممکن است به دیگر اندام‌ها سرایت کرده و متاستاز داده و آن‌ها را نیز درگیر سرطان کند [۵۸].

[۴] Endometrium

۵-۳ پیاده‌سازی و ارزیابی الگوریتم DeepCNV

برای پیاده‌سازی و آموزش شبکه از دو مدل پرسپترون چند لایه و شبکه عمیق LSTM استفاده شد. پیاده‌سازی الگوریتم با سیستمی با مشخصات جدول ۵-۳ استفاده شد.

جدول ۵-۳ : مشخصات سیستم استفاده شده

RAM	CPU	SYSTEM TYPE	GRAPHIC
۳۲ GB	Intel Core i۷-۶۷۰۰K ۴,۰۰GHz	Win۱۰ x۶۴-based	NVIDIA GEFORCE GTX ۱۰۶۰ ۶GB

همچنین پیاده‌سازی خود را به زبان پایتون و در محیط برنامه‌نویسی PyCharm انجام دادیم. برای پیاده‌سازی شبکه MLP و LSTM از کتابخانه‌های Keras و TensorFlow و Scikit-learn استفاده شد. ضمن این‌که مدل DeepCNV در مخزن گیت‌هاب[۵] در دسترس است. همچنین جهت پیاده‌سازی، داده‌های آموزش و آزمون همان‌طور که قبلا به آن اشاره شد، به صورت زیر در نظر گرفته شدند:

- تعداد کل نمونه بیماران: ۳۴۸۰
- تعداد داده‌های آموزش: ۳۱۳۲
- تعداد داده‌های اعتبارسنجی (۱۰٪ از داده‌های آموزش): ۳۱۳
- تعداد داده آزمون (۱۰٪): ۳۴۸

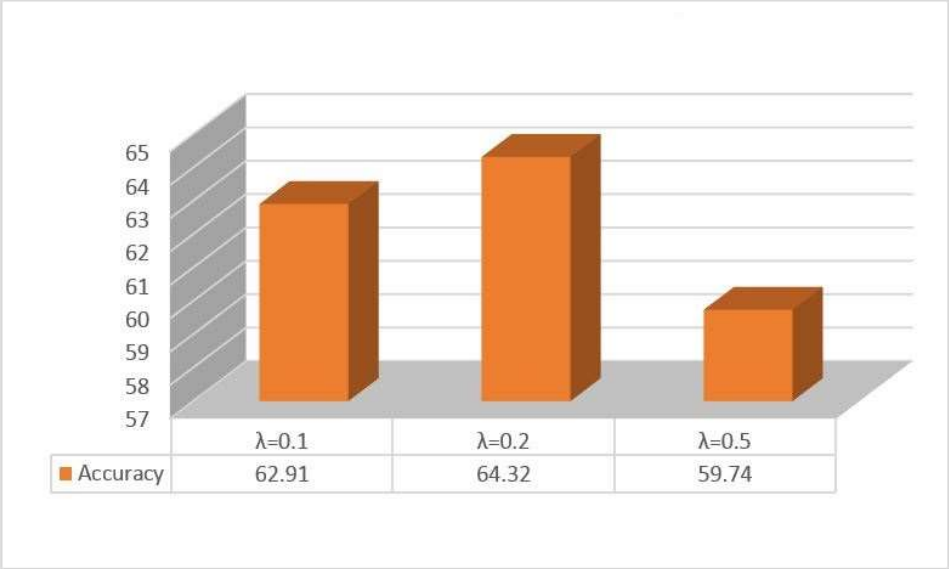
• ۵-۳-۱ پیاده‌سازی الگوریتم با کمک شبکه MLP

در این بخش ابتدا مشخصات پیاده‌سازی‌های انجام شده را بیان می‌کنیم و در انتها خروجی‌های حاصل از الگوریتم ارائه می‌شود. ضمن این‌که برای پیاده‌سازی از کتابخانه TensorFlow در زبان پایتون استفاده شد [۵۹].

[۵] https://github.com/borzadaran/DeepCNV

- **مشخصات شبکه MLP پیاده‌سازی شده**

برای پیاده‌سازی شبکه از ۴ لایه استفاه شد. همچنین تعداد مراحل الگوریتم MLP (Epoch) برابر ۱۰۰۰ تنظیم شد. ضمنا برای این‌که از حافظه سیستم بهترین بهره‌برداری را داشته باشیم، در هر مرحله، داده‌ها را به صورت دسته‌ای[۶] و تعداد ۵۰ نمونه در هر دسته را وارد شبکه کردیم و وزن‌ها را با کمک آن بروزرسانی نمودیم. در پیاده‌سازی نرخ یادگیری شبکه برابر ۰,۲ در نظر گرفته شد. مقدار ۰,۲ را با سعی و خطا به دست آوردیم. همان‌طور که در شکل ۵-۲ نشان داده شده است مقدار نرخ یادگیری یا λ به ازای ۰,۲ دارای بیشترین درصد دقت است. بنابراین این مقدار برای نرخ یادگیری انتخاب شد.



شکل ۵-۲: میزان دقت شبکه MLP برای مقادیر مختلف نرخ یادگیری

جدول ۵-۴ مشخصات شبکه پرسپترون چند لایه MLP برای طبقه‌بندی را نشان می‌دهد. همان‌طور که مشاهده می‌کنیم لایه خروجی برابر ۶ نرون در نظر گرفته شد که برابر تعداد کلاس‌ها است.

جدول ۵-۴: مشخصات شبکه MLP

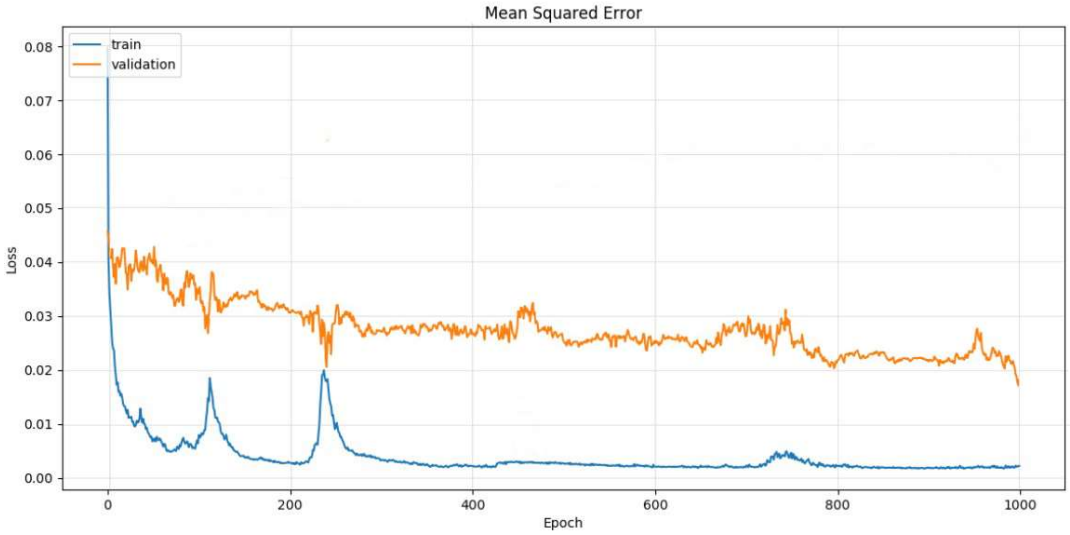
نرخ یادگیری	۰,۲

[۶] Batch mode

لایه‌های شبکه	لایه اول: ۵۰ نرون لایه دوم: ۵۰ نرون لایه سوم: ۱۰۰ نرون لایه خروجی: ۶ نرون

- **نتایج و ارزیابی شبکه MLP**

شکل ۵-۳ نمودار خطا در شبکه MLP را نشان می‌دهد. همان‌طور که در شکل ۵-۳ ملاحظه می‌کنیم آموزش شبکه از مرحله تقریبا ۲۴۰ به بعد خطای شبکه تقریبا به صفر نزدیک شده است. همچنین خطا روی داده های اعتبارسنجی به سمت صفر میل می‌کند. با کمک این الگوریتم به درصد دقت **۶۴,۳۲** درصد دست یافتیم.



شکل ۵-۳: نمودار میزان خطا در شبکه MLP

- ۵-۳-۲ پیاده‌سازی الگوریتم به کمک شبکه عمیق LSTM

در این قسمت ابتدا مشخصات شبکه مورد استفاده بیان می‌شود. سپس نتایج حاصل از آن را توضیح می‌دهیم. همچنین برای پیاده‌سازی شبکه LSTM از کتابخانه Keras در زبان پایتون استفاده گردید

.[۶۰]

- **مشخصات شبکه LSTM پیاده‌سازی شده**

جدول ۵-۵ مشخصات شبکه عمیق LSTM برای طبقه‌بندی را نشان می‌دهد. همان‌طور که ملاحظه می‌کنیم از تعداد ۱۰۰ واحد LSTM در لایه اول شبکه استفاده شد. در لایه دوم از Dropout استفاده کردیم. در این لایه ما Dropout را برابر ۲۰ درصد انتخاب کردیم. به این معنی که ۲۰ درصد از نرون‌های لایه قبل در تغییر وزن‌ها دخالت داده نمی‌شود.

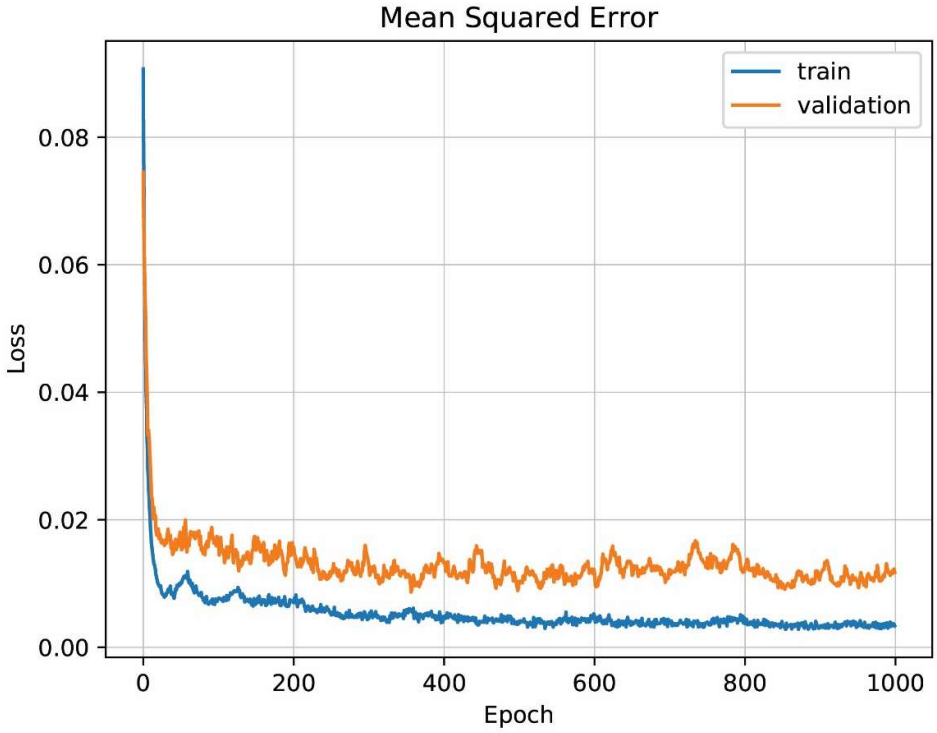
جدول ۵-۵: مشخصات شبکه LSTM

لایه اول	تعداد ۱۰۰ واحد LSTM
لایه دوم	Dropout برای جلوگیری از بیش‌برازش
لایه سوم	لایه خروجی با ۶ نرون (تعدادکلاس‌ها)

همچنین لایه خروجی برابر ۶ در نظر گرفته شده که بیان‌گر تعداد کلاس‌های سرطان است. همان‌طور که قبلا گفته شد تعداد مراحل الگوریتم در شبکه برابر ۱۰۰۰ در نظر گرفته شد.

- **نتایج و ارزیابی شبکه LSTM**

شکل ۵-۴ نمودار خطا شبکه LSTM را نشان می‌دهد.



شکل ٥-٤: نمودار میزان خطا در شبکه LSTM

همان‌طور که در شکل ۵-۴ مشاهده می‌کنیم میزان خطا از حدود مرحله ۲۰ به بعد کاهش یافته و میزان خطای داده‌های اعتبارسنجی نیز تا مرحله ۱۰۰۰ تقریبا ثابت مانده است.

با آموزش شبکه توسط LSTM برای آزمایش داده‌های دیده نشده با دقت **۹۲,۴۳** درصدی روبرو بودیم. بدین معنی که طبقه‌بند توانست ۳۴۸ داده که در آموزش از آن‌ها استفاده نشد، با دقت حدود ۹۲ درصد به درستی طبقه‌بندی کند.

۵-۴ دلیل برتری LSTM نسبت به MLP

همان‌طورکه مشاهده کردیم شبکه عمیق LSTM با دقت بسیار بالای ۹۲ درصدی توانست طبقه‌بندی را به درستی انجام دهد. در حالی که طبقه‌بند MLP با حدود ۲۵ درصد کمتر از LSTM این طبقه‌بندی را انجام داد. شبکه LSTM برای حل مشکل پدیده ناپدید شده گرادیان در شبکه های عصبی بازگشتی

بوجود آمد که تغییر عمده آن جایگزین کردن لایه میانی RNN با بلاک LSTM بود. با توجه به اینکه در شبکه MLP تعداد لایه‌ها افزایش یافته محاسبه گرادیان برای وزن‌ها نیز افزایش می‌یابد. همان‌طور که گفته شد مقدار دهی اولیه پارامترها در یک شبکه عصبی به فرم یک تابع گوسی با میانگین صفر و انحراف معیار یک، بدین معنی است که اندازه پارامترهای شبکه مانند وزن‌ها کمتر از یک است. همچنین با توجه به این‌که در لایه های اول تا سوم شبکه MLP از تابع فعال‌ساز سیگموئید استفاده شد آنگاه مشتق همیشه کمتر از ۰,۲۵ می‌شود. حال اگر تعداد زیادی از این اعداد کوچک در یکدیگر ضرب شوند باعث ایجاد یک عدد بسیار کوچک شده که تقریبا صفر است. در طرف مخالف ممکن است مقدار وزن‌ها بزرگ شوند آنگاه ضرب این اعداد در یک دیگر باعث شده تا گرادیان به بی نهایت میل کند. شبکه MLP راه‌حلی را برای ناپدید شدن گرادیان طراحی نکرده است. اما شبکه LSTM با کمک دروازه فراموشی می‌تواند تا چندین گام قبل مقدار حافظه خروجی را بازگرداند. در واقع اگر این دروازه کاملا بسته شود، حافظه قبلی کاملا پاک شده اما اگر دروازه باز باشد تمام مقدار حافظه قبلی از آن عبور می‌کند.

۵-۵ الگوریتم DeepCNV و معیار mRMR

در این بخش به دنبال بهترین ویژگی‌ها هستیم که در طبقه‌بند بیشترین تاثیر را داشته‌اند. به دلیل دقت بسیار بالای LSTM نسبت به MLP ما از این شبکه در یافتن برترین ویژگی‌ها بهره می‌گیریم.

با آموزش شبکه و طبقه‌بندی شبکه با کمک الگوریتم یادگیری عمیق LSTM به دقت حدود ۹۲ درصدی می‌رسیم. اما سوالی که پیش می‌آید این است که چگونه می‌توان این طبقه‌بندی را توصیف کرد یا از منظر دیگر چه تحلیل بایولوژیکی می‌توان از این طبقه‌بندی داشته باشیم. برای پاسخ به این دو مسئله نیاز است تا بهترین ویژگی‌هایی که در طبقه‌بندی بیشترین تاثیر را داشته‌اند شناسایی کنیم. جهت انتخاب بهترین ویژگی‌ها از معیار mRMR استفاده شد. ابتدا با کمک این معیار تعداد ۲۰۰ ویژگی را از دادگان خود استخراج کردیم. نکته جالب توجه در این آزمایش این بود که با انتخاب ۲۰۰ ویژگی

از ۲۴۱۷۴ ویژگی اولیه، باز هم با دقت ۸۱,۲۱ درصدی در آموزش شبکه روبرو بودیم. بنابراین می‌توان به اهمیت این ویژگی‌ها در طبقه‌بندی سرطان پی برد. در نهایت با کمک الگوریتم یادگیری عمیق LSTM و با ترکیب آن با الگوریتم SBS تعداد ۱۰ ویژگی برتر را استخراج کردیم.

۱-۵-۵ انتخاب ویژگی به کمک mRMR

با کمک این الگوریتم تعداد دویست ویژگی که بیشترین وابستگی به کلاس‌های سرطان را داشته و کمترین افزونگی رو به یکدیگر داشتند استخراج گردید جدول ۵-۶ این ویژگی‌ها را نشان می‌دهد.

جدول ۵-۶: ژن‌های استخراج شده توسط معیار mRMR

ترتیب	ژن	امتیاز	ترتیب	ژن	امتیاز
۱	APC۲	۰,۳۳۸	۱۰۱	SLC۳۹A۳	۰,۳۲۹
۲	MED۱۶	۰,۳۳۷	۱۰۲	C۱۹orf۷۱	۰,۳۲۹
۳	WDR۱۸	۰,۳۳۷	۱۰۳	ZNF۵۵۴	۰,۳۲۹
۴	PRTN۳	۰,۳۳۷	۱۰۴	STK۱۱	۰,۳۲۹
۵	R۳HDM۴	۰,۳۳۷	۱۰۵	CTBP۲	۰,۳۲۹
۶	KISS۱R	۰,۳۳۷	۱۰۶	MFSD۱۲	۰,۳۲۹
۷	ELANE	۰,۳۳۷	۱۰۷	GIPC۳	۰,۳۲۹
۸	CFD	۰,۳۳۷	۱۰۸	HCN۲	۰,۳۲۹
۹	RPS۱۵	۰,۳۳۷	۱۰۹	THOP۱	۰,۳۲۹
۱۰	C۱۹orf۲۵	۰,۳۳۶	۱۱۰	ZNF۵۵۶	۰,۳۲۹
۱۱	PLK۵	۰,۳۳۶	۱۱۱	SGTA	۰,۳۲۸
۱۲	GADD۴۵B	۰,۳۳۶	۱۱۲	ZNF۵۷	۰,۳۲۸
۱۳	ADAMTSL۵	۰,۳۳۶	۱۱۳	ZNF۵۵۵	۰,۳۲۸
۱۴	CDKN۲A	۰,۳۳۶	۱۱۴	TJP۳	۰,۳۲۸
۱۵	PCSK۴	۰,۳۳۶	۱۱۵	ODF۳L۲	۰,۳۲۷
۱۶	LMNB۲	۰,۳۳۶	۱۱۶	C۱۹orf۲۹	۰,۳۲۷

ترتیب	ژن	امتیاز	ترتیب	ژن	امتیاز
۱۷	ARID۳A	۰,۳۳۵	۱۱۷	HMG۲۰B	۰,۳۲۶
۱۸	MBD۳	۰,۳۳۵	۱۱۸	TBXA۲R	۰,۳۲۶
۱۹	HMHA۱	۰,۳۳۵	۱۱۹	MRPL۵۴	۰,۳۲۶
۲۰	TIMM۱۳	۰,۳۳۵	۱۲۰	MATK	۰,۳۲۶
۲۱	C۱۹orf۶	۰,۳۳۵	۱۲۱	BSG	۰,۳۲۶
۲۲	GRIN۳B	۰,۳۳۵	۱۲۲	GZMM	۰,۳۲۶
۲۳	IL۱۷RC	۰,۳۳۵	۱۲۳	RAX۲	۰,۳۲۶
۲۴	MIR۴۷۴۵	۰,۳۳۵	۱۲۴	CDC۳۴	۰,۳۲۶
۲۵	TCF۳	۰,۳۳۵	۱۲۵	TPGS۱	۰,۳۲۶
۲۶	UQCR۱۱	۰,۳۳۵	۱۲۶	APBA۳	۰,۳۲۶
۲۷	MIR۳۱۸۷	۰,۳۳۵	۱۲۷	C۱۹orf۲۹-AS۱	۰,۳۲۶
۲۸	hsa-mir-۳۱۸۷	۰,۳۳۵	۱۲۸	PIP۵K۱C	۰,۳۲۵
۲۹	AZU۱	۰,۳۳۵	۱۲۹	SHC۲	۰,۳۲۵
۳۰	LPPR۳	۰,۳۳۵	۱۳۰	ZFR۲	۰,۳۲۵
۳۱	ATP۸B۳	۰,۳۳۵	۱۳۱	C۲CD۴C	۰,۳۲۵
۳۲	MEX۳D	۰,۳۳۵	۱۳۲	MADCAM۱	۰,۳۲۵
۳۳	DAZAP۱	۰,۳۳۵	۱۳۳	MIER۲	۰,۳۲۴
۳۴	SCAMP۴	۰,۳۳۵	۱۳۴	hsa-mir-۱۳۰۲-۱۱	۰,۳۲۴
۳۵	GAMT	۰,۳۳۵	۱۳۵	PPAP۲C	۰,۳۲۴
۳۶	MIR۱۲۲۷	۰,۳۳۴	۱۳۶	OR۴F۱۷	۰,۳۲۴
۳۷	GNG۷	۰,۳۳۴	۱۳۷	WASH۵P	۰,۳۲۴
۳۸	NFIC	۰,۳۳۴	۱۳۸	ATCAY	۰,۳۲۴
۳۹	ONECUT۳	۰,۳۳۴	۱۳۹	THEG	۰,۳۲۴
۴۰	MIR۱۹۰۹	۰,۳۳۴	۱۴۰	ITGB۱BP۳	۰,۳۲۲

ترتیب	ژن	امتیاز	ترتیب	ژن	امتیاز
۴۱	FAM۱۰۸A۱	۰,۳۳۴	۱۴۱	ZFP۹۰	۰,۳۲۲
۴۲	KLF۱۶	۰,۳۳۴	۱۴۲	SMPD۳	۰,۳۲۱
۴۳	LOC۱۰۰۲۸۸۱۲۳	۰,۳۳۴	۱۴۳	PRMT۷	۰,۳۲۱
۴۴	REXO۱	۰,۳۳۴	۱۴۴	SLC۷A۶OS	۰,۳۲۱
۴۵	MIR۴۷۰۳	۰,۳۳۴	۱۴۵	SLC۷A۶	۰,۳۲۱
۴۶	CSNK۱G۲	۰,۳۳۴	۱۴۶	HAS۳	۰,۳۲
۴۷	IZUMO۴	۰,۳۳۴	۱۴۷	RLTPR	۰,۳۲
۴۸	CSNK۱G۲-AS۱	۰,۳۳۴	۱۴۸	DAPK۳	۰,۳۲
۴۹	ADAT۳	۰,۳۳۴	۱۴۹	CREM	۰,۳۲
۵۰	POLR۲E	۰,۳۳۴	۱۵۰	ZBTB۷A	۰,۳۲
۵۱	LSM۷	۰,۳۳۴	۱۵۱	PIAS۴	۰,۳۲
۵۲	AMH	۰,۳۳۴	۱۵۲	SNORD۳۷	۰,۳۲
۵۳	SEMA۶A	۰,۳۳۴	۱۵۳	hsa-mir-۶۳۷	۰,۳۱۹
۵۴	PLEKHJ۱	۰,۳۳۴	۱۵۴	EEF۲	۰,۳۱۹
۵۵	TMPRSS۹	۰,۳۳۴	۱۵۵	PLA۲G۱۵	۰,۳۱۹
۵۶	BTBD۲	۰,۳۳۳	۱۵۶	ESRP۲	۰,۳۱۹
۵۷	PTBP۱	۰,۳۳۳	۱۵۷	MIR۶۳۷	۰,۳۱۹
۵۸	NDUFS۷	۰,۳۳۳	۱۵۸	CDH۳	۰,۳۱۹
۵۹	DIRAS۱	۰,۳۳۳	۱۵۹	ACD	۰,۳۱۸
۶۰	GNA۱۵	۰,۳۳۳	۱۶۰	PARD۶A	۰,۳۱۸
۶۱	MMD۲	۰,۳۳۳	۱۶۱	GZMM	۰,۳۱۸
۶۲	MUM۱	۰,۳۳۳	۱۶۲	RAX۲	۰,۳۱۸
۶۳	EFNA۲	۰,۳۳۲	۱۶۳	GFOD۲	۰,۳۱۸
۶۴	SPPL۲B	۰,۳۳۲	۱۶۴	CDC۳۴	۰,۳۱۸
۶۵	hsa-mir-۴۳۲۱	۰,۳۳۲	۱۶۵	C۱۶orf۴۸	۰,۳۱۸

ترتیب	ژن	امتیاز	ترتیب	ژن	امتیاز
۶۶	C۱۹orf۳۵	۰,۳۳۲	۱۶۶	C۱۶orf۸۶	۰,۳۱۸
۶۷	JSRP۱	۰,۳۳۲	۱۶۷	NFATC۳	۰,۳۱۸
۶۸	MIR۴۳۲۱	۰,۳۳۲	۱۶۸	CUL۲	۰,۳۱۸
۶۹	RFFL	۰,۳۳۲	۱۶۹	CDH۱	۰,۳۱۸
۷۰	AP۳D۱	۰,۳۳۲	۱۷۰	SLC۱۲A۴	۰,۳۱۸
۷۱	OAZ۱	۰,۳۳۲	۱۷۱	TMCO۷	۰,۳۱۸
۷۲	LINGO۳	۰,۳۳۲	۱۷۲	DPEP۳	۰,۳۱۸
۷۳	DOT۱L	۰,۳۳۲	۱۷۳	CENPT	۰,۳۱۸
۷۴	C۱۹orf۲۱	۰,۳۳۲	۱۷۴	EDC۴	۰,۳۱۷
۷۵	SBNO۲	۰,۳۳۲	۱۷۵	PSKH۱	۰,۳۱۷
۷۶	GNA۱۱	۰,۳۳۲	۱۷۶	NRN۱L	۰,۳۱۷
۷۷	CELF۵	۰,۳۳۲	۱۷۷	LOC۲۸۳۰۷۰	۰,۳۱۷
۷۸	CIRBP-AS۱	۰,۳۳۲	۱۷۸	DUS۲L	۰,۳۱۷
۷۹	MKNK۲	۰,۳۳۲	۱۷۹	SNTB۲	۰,۳۱۷
۸۰	PALM	۰,۳۳۲	۱۸۰	LCAT	۰,۳۱۷
۸۱	ABCA۷	۰,۳۳۲	۱۸۱	PSMB۱۰	۰,۳۱۷
۸۲	S۱PR۴	۰,۳۳۲	۱۸۲	TSNAXIP۱	۰,۳۱۷
۸۳	RBFOX۱	۰,۳۳۲	۱۸۳	DDX۲۸	۰,۳۱۷
۸۴	C۱۹orf۲۴	۰,۳۳۲	۱۸۴	CHTF۸	۰,۳۱۷
۸۵	CIRBP	۰,۳۳۲	۱۸۵	CTRL	۰,۳۱۷
۸۶	C۱۹orf۲۶	۰,۳۳۲	۱۸۶	RANBP۱۰	۰,۳۱۷
۸۷	GPX۴	۰,۳۳۱	۱۸۷	CIRH۱A	۰,۳۱۷
۸۸	AES	۰,۳۳۱	۱۸۸	CCNY	۰,۳۱۶
۸۹	ATP۵D	۰,۳۳۱	۱۸۹	MAP۲K۲	۰,۳۱۶
۹۰	NCLN	۰,۳۳۱	۱۹۰	COG۸	۰,۳۱۶

ترتیب	ژن	امتیاز	ترتیب	ژن	امتیاز
۹۱	DOHH	۰,۳۳۱	۱۹۱	CTCF	۰,۳۱۶
۹۲	PRSS۵۷	۰,۳۳	۱۹۲	DPEP۲	۰,۳۱۶
۹۳	FZR۱	۰,۳۳	۱۹۳	UCN۳	۰,۳۱۶
۹۴	TLE۶	۰,۳۳	۱۹۴	CALB۲	۰,۳۱۶
۹۵	FGF۲۲	۰,۳۳	۱۹۵	COG۸	۰,۳۱۶
۹۶	FSTL۳	۰,۳۳	۱۹۶	CTCF	۰,۳۱۶
۹۷	RNF۱۲۶	۰,۳۳	۱۹۷	IL۲RA	۰,۳۱۶
۹۸	MOB۳A	۰,۳۳	۱۹۸	AKR۱C۴	۰,۳۱۶
۹۹	ZNF۷۷	۰,۳۳	۱۹۹	CCDC۳	۰,۳۱۶
۱۰۰	C۱۹orf۷۷	۰,۳۳	۲۰۰	TUBAL۳	۰,۳۱۶

۲-۵-۵ انتخاب ویژگی‌های برتر به کمک DeepCNV و معیار mRMR

برای انتخاب ویژگی‌های برتر از معیار mRMR استفاده شد. این معیار در ابتدا تعداد ۲۰۰ ویژگی که بیشترین وابستگی و کمترین افزونگی را دارا بوده به عنوان خروجی برمی‌گرداند (مقدار ۲۰۰ را خودمان به دلخواه به معیار mRMR می‌دهیم). در مرحله بعدی تمام ۲۰۰ ویژگی ($n = ۲۰۰$) را به الگوریتم DeepCNV می‌دهیم. در ادامه همان‌طور که از الگوریتم DeepCNV می‌دانیم جهت پی بردن به اهمیت یک ویژگی، اقدام به حذف آن ویژگی می‌کنیم. با حذف این ویژگی میزان خطا و درصد دقت شبکه که همان تابع ارزیاب بوده، را محاسبه می‌کنیم. بنابراین هر چه میزان خطای شبکه با حذف این ویژگی بیشتر باشد و درصد دقت شبکه پایین‌تر، می‌توان به با اهمیت بودن این ویژگی پی برد. بنابراین حذف این ویژگی خطای شبکه را افزایش می‌دهد. این عملیات را برای تمام ۲۰۰ ویژگی انجام می‌دهیم. در نهایت ۱۰ ویژگی که با حذف آن، شبکه دارای خطای بالاتر و درصد دقت پایین‌تر شده است استخراج می‌کنیم. در واقع مقدار ۱۰ همان مقدار d در الگوریتم DeepCNV است که به دلخواه در این‌جا برابر

۱۰ درنظر گرفته شده است. با انجام عملیات بالا ده ژن استخراج شده به صورت جدول ۵-۷ می‌باشند.

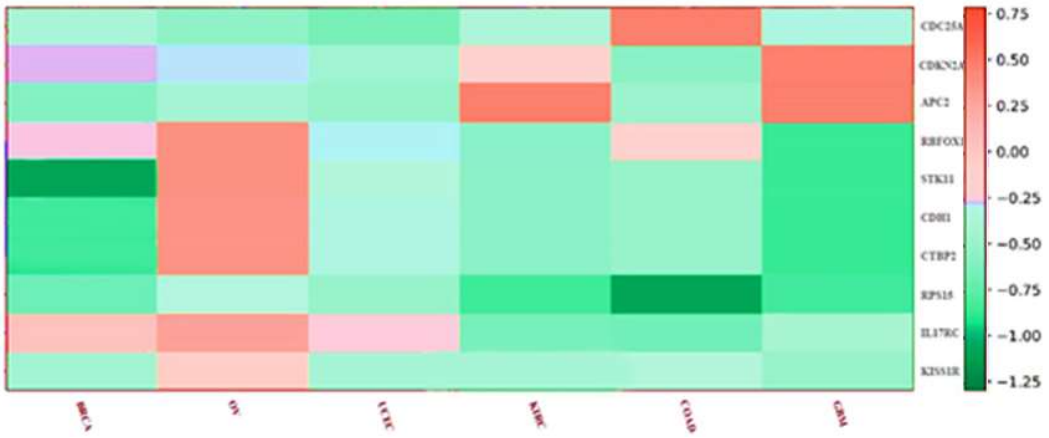
جدول ۵-۷: استخراج ده ژن برتر با کمک LSTM و SBS

CTBP۲	CDH۱
RPS۱۵	CDKN۲A
RBFOX۱	APC۲
CDC۲۵A	STK۱۱
KISS۱R	IL۱۷RC

در ادامه این بخش از تحقیق، نمودار Heat-Map[۷] را رسم می‌کنیم. سپس به شرح هر کدام از ژن‌ها پرداخته و میزان تاثیر هر کدام از آن‌ها را در ایجاد سرطان‌های مختلف با کمک منابع معتبر در این زمینه توضیح می‌دهیم.

۵-۶ نمودار Heat-Map

برای فهم تاثیر هر ژن در کلاس سرطان نمودار Heat-Map آن را طراحی کردیم. این نمودار میزان تاثیر هر کدام از ژن‌ها در کلاس‌های سرطان را نشان می‌دهد. شکل ۵-۵ این نمودار را نشان می‌دهد.



شکل ۵-۵: میانگین وقوع حذف و درج ژن (CNV) در شش نوع سرطان در مجموعه دادگان

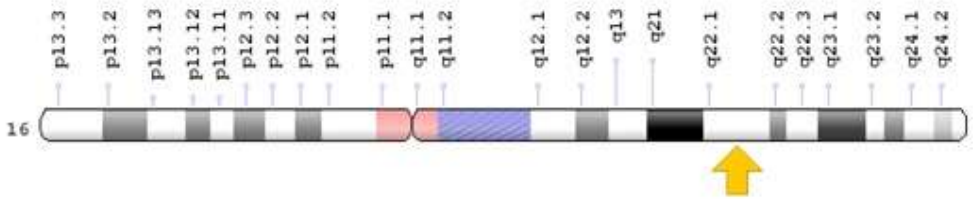
[۷] نقشه گرما

برای بدست آوردن این نمودار برای هر ژن میانگین سطح CNV در تمام بیماران برای هر سرطان مختلف را محاسبه کردیم. به طور مثال برای ژن APC۲ که در نمودار نشان داده شده است، در سرطان GBM، میزان سطح CNV این ژن را در تمام بیماران با یکدیگر جمع کرده و در نهایت بر تعداد بیماران تقسم کردیم. در این نمودار رنگ قرمز نشان از CNV با مضاعف شدن یا درج[۱] و رنگ سبز نشان دهنده حذف[۲] است. سطرهای این نمودار نشان دهنده هر کدام از ژن‌های استخراج شده بوده و ستون‌ها نشان دهنده ۶ کلاس سرطان مختلف است. هر چه میزان غلظت رنگ سبز یا قرمز بیشتر باشد نشان دهنده این است که به ترتیب تاثیر درج یا حذف در آن ژن، در ایجاد سرطان زیادتر است. در ادامه این بخش به شرح جزئیات ویژگی‌های استخراج شده با کمک الگوریتم DeepCNV می‌پردازیم. ضمن این‌که تصاویر محل قرار گرفتن هر ژن در کروموزم و همچنین توضیحات مربوط به آن ژن، از [۶۱] استخراج شد.

نام ژن: **CDH۱**

نام کامل: Cadherin ۱

نام‌های دیگر: Arc-۱, CADH۱_HUMAN, CAM ۱۲۰/۸۰, CD۳۲٤, CDHE, E-cadherin, ECAD, LCAM, liver cell adhesion molecule, UVO, uvomorulin



شکل ۵-۶: مکان قرارگرفتن CDH۱ در کروموزم شماره ۱۶ [۶۱].

ژن CDH۱ دستورالعمل‌هایی برای ساخت پروتئینی به نام کادرفین اپیدمی یا E-Cadherin فراهم

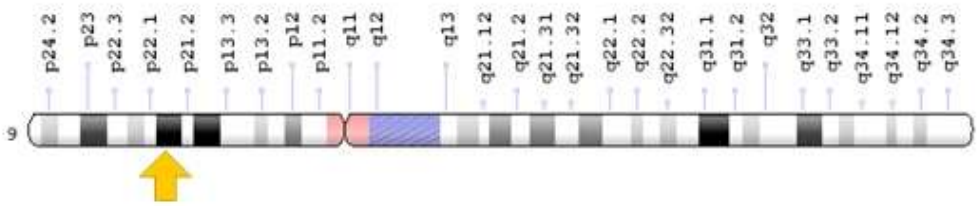
[۱] Amplification
[۲] Deletion

می‌کند. این پروتئین در داخل غشا قرار دارد که سلول‌های اپیتلیالی (سلول‌هایی هستند که سطوح و حفره‌های بدن را مانند داخل درون پلک‌ها و دهان قرار می‌دهند)، در بر می‌گیرد. پروتئین E-Cadherin متعلق به یک خانواده از پروتئین‌ها به نام Cadherins است که با کمک آن سلول‌های همسایه به یکدیگر چسبیده و بافت‌ها را تشکیل دهند. طبق منابع بسیاری مانند [۶۲-۶۴] جهش در این ژن عامل بسیاری از سرطان‌ها مانند سرطان سینه، OV و سرطان پروستات است.

نام ژن: **CDKN۲A**

نام کامل: Cyclin dependent kinase inhibitor ۲A

نام‌های دیگر: ARF, CDK۴ inhibitor, p۱۶-INK۴, INK۴A, MLM ,MTS۱, P۱۴, P۱۴ARF, P۱۶, P۱۹ARF



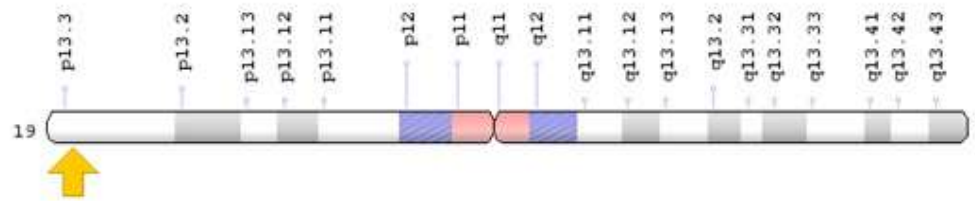
شکل ۵-۷: مکان قرارگرفتن CDKN۲A در کروموزم شماره ۹ [۶۱].

ژن CDKN۲A دستورالعمل‌هایی برای ساخت چندین پروتئین را فراهم می‌کند. دو مورد مهم از پروئین‌های ساخته شده، پروتئین p۱۶ (INK۴A) و p۱۴ (ARF) است. هر دو این پروتئین‌ها به عنوان سرکوب کننده‌های تومور عمل می‌کنند. این بدین معنی است که سلول‌ها را از رشد و تقسیم بیش از حد نگه می‌دارند. هر دو این پروتئین‌ها در متوقف کردن تقسیم سلولی در سلول‌های بزرگ‌تر نقش دارند. حذف در این ژن می‌تواند منجر به ایجاد سرطان GBM شود [۶۵]. همچنین جهش در این ژن در بسیاری دیگر از انواع سرطان مانند سرطان ریه تاثیرگذار است [۶۶].

نام ژن: **APC۲**

نام کامل: APC۲, WNT signaling pathway regulator

نام‌های دیگر: APCL



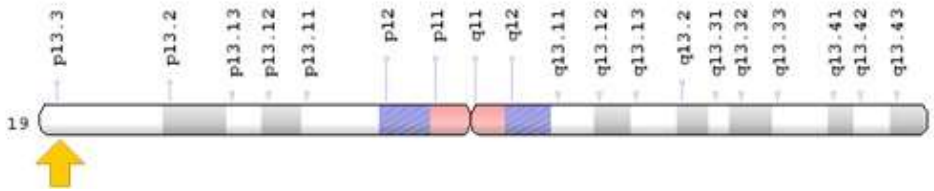
شکل ۵-۸: مکان قرارگرفتن APC۲ در کروموزم شماره ۱۹ [۶۱].

این ژن اثرات مشابه سرکوب کننده تومور را دارد. این ژن دارای بالاترین بیان[۱] در سیستم عصبی مرکزی است و در توسعه مغز و در نرون‌ها نقش مهمی ایفا می‌کند. طبق [۶۷] جهش در این ژن عامل مهم ایجاد سرطان‌هایی مانند GBM است.

نام ژن: **STK۱۱**

نام کامل: Serine/threonine kinase ۱۱

نام‌های دیگر: LKB۱, PJS



شکل ۵-۹: مکان قرارگرفتن STK۱۱ در کروموزم شماره ۱۹ [۶۱].

ژن STK۱۱ دستورالعملی را برای ساخت یک آنزیم به نام serine/threonine kinase ۱۱ ارایه می‌کند. این آنزیم یک متوقف کننده تومور است. بدین معنی که به نگهداری سلول‌ها از رشد و تقسیم سریع و یا

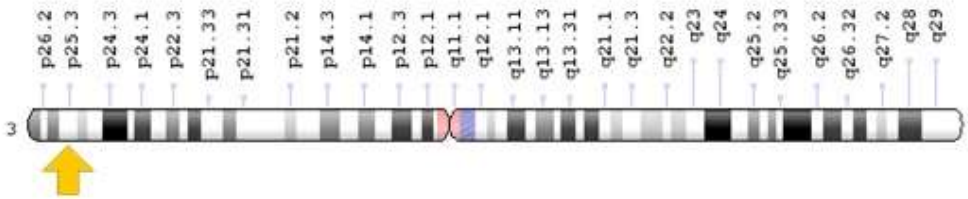
[۱] بیان ژن فرایندی است که در آن اطلاعات درون ژن استفاده شده تا یک محصول کاربردی از آن بدست آید.

به روش کنترل نشده کمک می‌کند. جهش در این ژن یکی از عوامل اصلی در ایجاد سرطان‌های سینه و OV و ریه است [۶۸-۷۰].

نام ژن: **IL۱۷RC**

نام کامل: Interleukin ۱۷ receptor C

نام‌های دیگر: IL-۱۷ receptor C, IL-۱۷RL, IL۱۷F receptor, interleukin-۱۷ receptor homolog, interleukin-۱۷ receptor-like protein



شکل ۵-۱۰: مکان قرارگرفتن IL۱۷RC در کروموزوم شماره ۳ [۶۱].

ژن IL۱۷RC دستورالعملی برای ساخت پروتئینی است که در عملکرد سیستم ایمنی خصوصا در دفاع از بدن در برابر قارچی به نام کاندیدا[۲] دخیل است. هنگامی که سیستم ایمنی بدن کاندیدا را تشخیص می‌دهد، سلول‌هایی به نام سلول‌های Th۱۷ تولید می‌کند. این سلول‌ها نقش اساسی در فرایند التهاب[۳] دارند [۷۱]. علاوه بر این تولید برخی از بخش‌های ضد میکروبی پروتئینی را که از رشد کاندیدا بر روی سطح غشای مخاطی جلوگیری می‌کند، ترویج می‌دهد. همچنین جهش در IL۱۷RC نقش مهمی در ایجاد سرطان‌هایی مانند پروستات را داراست [۷۲].

---

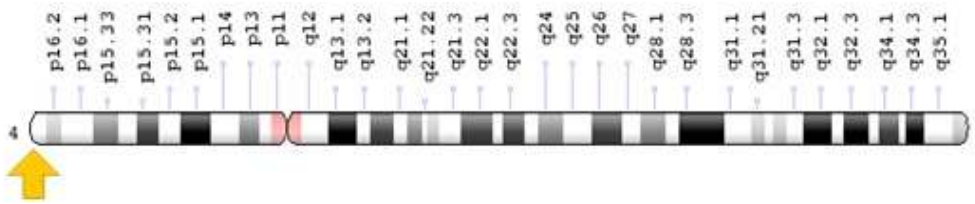
[۲] Candida

[۳] التهاب نوعی پاسخ موضعی بدن است که به دنبال خراش، بریدگی، سوختگی یا هر نوع آسیب بافتی دیگر بروز می‌کند و باعث تورم، قرمزی، گرمی، و خارش محل آسیب دیده می‌شود.

نام ژن: **CTBP۲**

نام کامل: C-terminal binding protein ۲

نام‌های دیگر: BARS, HADDTS



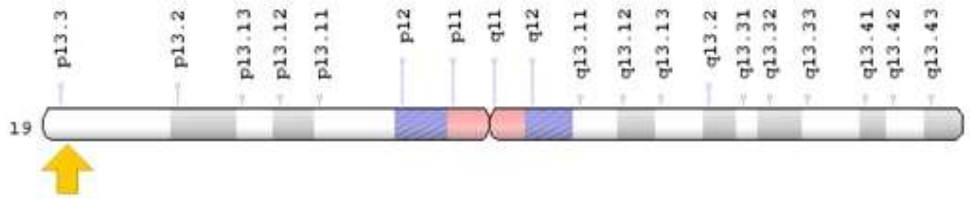
شکل ۵-۱۱: مکان قرارگرفتن CTBP۲ در کروموزم شماره ۴ [۶۱].

طبق [۷۳] این ژن عامل مهمی در ایجاد سرطان OV است. همچنین جهش در این ژن در ایجاد سرطان‌هایی مانند سرطان سینه [۷۴-۷۵] و همچنین سرطان روده بزرگ [۷۶] نقش دارد.

نام ژن: **RPS۱۵**

نام کامل: Ribosomal protein S۱۵

نام‌های دیگر: RIG, S۱۵



شکل ۵-۱۲: مکان قرار گرفتن RPS۱۵ در کروموزم ۱۹ [۶۱].

این ژن در محل سیتوپلاسم قرار داشته و ریبوزوم پروتئین را کدگذاری می‌کند. طبق تحقیقات بسیاری در زمینه سرطان‌های سینه و OV، درج در ژن RPS۱۵ می‌تواند یکی از دلایل ایجاد این نوع سرطان‌ها باشد [۷۷]. همچنین طبق مقاله [۷۸] درج در این ژن باعث ایجاد سرطان‌هایی مانند سرطان روده بزرگ شود.

نام ژن: **RBFOX۱**

نام کامل: RNA binding fox-۱ homolog ۱

نام‌های دیگر: ۲BP۱, FOX۱, A۲BP۱, FOX-۱, HRNBP۱

این ژن در کروموزم شماره ۱۶ قرار دارد. طبق [۷۹] حذف شدن بخشی از این ژن می‌تواند عامل مهمی در ایجاد سرطان روده بزرگ باشد.

نام ژن: **CDC۲۵A**

نام کامل: Cell division cycle ۲۵A

نام‌های دیگر: PAQR۱۰

این ژن در کروموزم شماره ۳ قرار دارد. جهش در این ژن عامل بسیاری از سرطان‌ها مانند سرطان روده بزرگ [۸۰] و سر و گردن [۸۱] بوده است.

نام ژن: **KISS۱R**

نام کامل: KISS۱ receptor

نام‌های دیگر: HH۸, CPPB۱, GPR۵۴, AXOR۱۲, KISS-۱R, HOTY۷T۱۷۵

این ژن در کروموزم شماره ۱۹ قرار دارد. KISS۱R یک ژن سرکوب کننده متاستاز بوده و همچنین سرکوب کننده تومور در سرطان‌های متعدد انسان است. مطالعات اخیر نشان می‌دهد که جهش در KISS۱R نقش مهمی در سرطان سینه ایفا می‌کند [۸۲].

۷-۵ نتیجه‌گیری

در این فصل از تحقیق الگوریتم DeepCNV را با کمک شبکه پرسپترون چند لایه MLP آموزش دادیم. سپس الگوریتم را با شبکه عمیق LSTM آموزش دادیم. تفاوت‌های بین این دو آموزش را مشاهده

کردیم. شبکه LSTM با حل مشکل ناپدید شدن گرادیان درصد دقت بالایی را برای ما فراهم نمود. برای این‌که بتوانیم تحلیل بایولوژی از کار خود داشته باشیم، ابتدا با کمک معیار mRMR تعداد ۲۰۰ ویژگی را استخراج نمودیم. سپس به طبقه‌بندی آن پرداختیم. نشان دادیم که با این ۲۰۰ ویژگی نیز می‌توان با درصد دقت بالایی طبقه‌بندی را انجام داده که نشان از اهمیت بالای این ویژگی‌هاست. در نهایت با کمک الگوریتم SBS تعداد ده ویژگی برتر را استخراج نمودیم. در پایان این فصل به مطالعه هر کدام از این ژن‌ها پرداخته و به تاثیر آن‌ها در انواع مختلف سرطان‌ها آشنا شدیم.

فصل ٦ : جمع بندی و پژوهش های آینده

۶-۱ جمع‌بندی

هدف ما از این تحقیق، طبقه‌بندی سرطان و شناسایی ژن‌های مهم در ایجاد سرطان‌های مختلف با استفاده از یادگیری عمیق بود. همان‌طور که گفته شد CNV، به عنوان یکی از انواع مختلف جهش در DNA، دلیل اصلی ایجاد سرطان در انسان بود. برای پی بردن به اهمیت این موضوع، به کمک CNVها، به طبقه‌بندی ۶ کلاس مختلف سرطان در انسان با استفاده از مقادیر سطح CNV و با استفاده از یادگیری عمیق LSTM پرداختیم. برای این منظور اطلاعات CNV از ۲۴۱۷۴ ژن از تعداد ۳۴۸۰ بیمار سرطانی که دچار هر کدام از این ۶ نوع سرطان شده بودند برای ساخت طبقه‌بند استفاده گردید. نتایج طبقه‌بندی ما دقت ۹۲ درصدی را نشان داد. در گام بعدی برای تحلیل بایولوژی به شناسایی ژن‌هایی که بیشترین تاثیر را در ایجاد سرطان‌ها داشتند پرداختیم. بدین سبب از معیار mRMR جهت شناسایی ژن‌های موثر در سرطان‌ها استفاده گردید. با استفاده از این الگوریتم تعداد ۲۰۰ ژن برتر را شناسایی کردیم و سپس طبقه‌بندی را با استفاده از این ۲۰۰ ژن انجام دادیم که دقت بیش از ۸۰ درصدی نشان از اهمیت این ژن‌ها در وقوع سرطان‌های مختلف بود. در گام آخر از بین این ۲۰۰ ژن تعداد ده ژن را با کمک الگوریتم DeepCNV استخراج کردیم و نمودار Heat-Map آن‌ها را ترسیم نمودیم. در نهایت عملکرد ژنتیکی این ژن‌ها را مورد بررسی قرار دادیم. امید است نتایج حاصل از الگورتیم DeepCNV کمک فراوانی به شناخت انواع سرطان‌های مختلف انسان کرده و پیشنهاداتی را برای تجویز درمان‌های مناسب بیماران ارائه دهد.

۶-۲ پژوهش‌های آینده

در این تحقیق ما از یک طبقه‌بند قدرتمند هوش مصنوعی بهره گرفتیم و با کمک آن توانستیم با درصد دقت بسیار بالا داده‌های CNV بیماران برای شش کلاس مختلف سرطان را طبقه‌بندی کنیم. همچنین با کمک این طبقه‌بند توانستیم عامل‌های مهم ژن، در این شش نوع کلاس سرطان را شناسایی کنیم.

در انجام این مطالعه ما تنها از داده‌های CNV شش سرطان استفاده نمودیم. در پژوهش‌های آینده می‌توان از تعداد کلاس بیشتری بهره برد و با کمک الگوریتم DeepCNV می‌توان به شناسایی عامل‌های مختلف سرطان پرداخت. با شناسایی این عامل می‌توان با کمترین هزینه درمان‌های موثری را تجویز نمود.

همان‌طور که می‌دانیم یادگیری ماشین با طراحی تضمینی اطلاعات مناسب شروع می‌شود. عملکرد بهتر اغلب با استفاده از ویژگی‌های بدست آمده از ورودی اصلی حاصل می‌گردد. تکنیک‌های انتخاب ویژگی چه در عمل و چه در تئوری تأثیر خود را در افزایش کارایی یادگیری، افزایش دقت پیش‌بینی و کاهش پیچیدگی نتایج آموزش داده شده، نشان می‌دهد. هدف اصلی انتخاب ویژگی در مسئله طبقه‌بندی، انتخاب زیرمجموعه متغیرهای ورودی با حذف ویژگی‌های نامربوط یا ویژگی‌هایی که فاقد اطلاعات مفید هستند، است. انتخاب بهترین ویژگی‌ها، پیچیدگی زمانی مسئله طبقه‌بندی را کاهش می‌دهد و همچنین در یافتن زیرمجموعه بهترین ویژگی‌ها کمک می‌کند. به طور کلی می‌توان گفت که هدف اصلی از انتخاب ویژگی، بهبود عملکرد پیش‌بینی، ارائه‌ی پیش‌بینی سریع‌تر و مقرون به صرفه‌تر و ارائه‌ی درک بهتر از روند اطلاعات تولید شده است. در این تحقیق برای شناسایی مسئله انتخاب ویژگی در بانک اطلاعاتی از معیار mRMR استفاده گردید. در پژوهش‌های آینده می‌توان از معیارهای متفاوت دیگری مانند آن‌چه در[۸۳] پرداخته شده است، جهت حذف وابستگی‌ها در دادگان استفاده نمود و در نهایت با مقایسه آن‌ها با یکدیگر به طبقه‌بندی دادگان پرداخت.

فهرست واژگان

ا

آدنین..... Adenine

آلل..... Allele

اطلاعات متقابل..... Mutual Information

الگوریتم انتشار به عقب..... Back Propagation Algorithm

ب

بیشترین وابستکی-کمترین افزونگی..... Maximum Dependency-Minumum Redundancy

پ

پولیپ..... Polyp

ت

تابع تانژانت هیپربولیک..... tanh

تابع هار..... Haar

تابع سیگموئید..... Sigmoid

تنوع تعداد کپی..... Copy Number Variation

تیامین..... Thymine

ج

جهش..... Mutatin

ح

حافظه طولانی کوتاه-مدت..... Long Short Term Memory

حذف..... Deletion

دی ان ای..... Deoxyribonucleic acid

دروازه ورودی..... Input gate

درج..... Amplification

دیده نشده..... Unseen

ر

ریزآرایه..... Microarrays

س

سرطان..... Cancer

سیگنال‌های تابعی..... Functional Signals

سیگنال‌های خطا..... Error Signals

سیتوزین..... Cytosine

ش

شبکه پیشخور..... Feed Forward Network

شبکه‌های عصبی بازگشتی..... Recurrent Neural Network

ک

کاهش گرادیان تصادفی..... Stochastic Gradient Descent

گ

گوانین..... Guanine

م

متاستاز..... Metastasis

د

دروازه خروجی.....Output gate

دروازه فراموشی.....Forget gate

نظریه اطلاعات.....Information theory

ی

یادگیری با نظارت.....Supervised Learning

ن

نظریه احتمالات.....Probability theory

نظریه احتمالات.....Probability theory

نمودار نقشه گرما..... Heat Map diagram

مراجع

[١] A.s, Daar, S.w, Scherer, and R.a, Hegele, "Implications of copy-number variation in the human genome: A time for questions," *Nature Reviews Genetics*, vol. ٧, no. ٦, p. ٤١٤, ٢٠٠٦.

[٢] B. Fan, Z. Du, D. M. Gorbach, and M. F. Rothschild, "Development and Application of High-density SNP Arrays in Genomic Studies of Domestic Animals methodologies and technologies have been developed," *The Asian-Aust. J. Anim. Sci.*, vol. ٢٣, no. ٧, pp. ٨٣٣–٨٤٧, ٢٠١٠.

[٣] P. D. D. de Koning A. P. Jason, Gu Wanjun, Castoe Todd A., Batzer Mark A., "Repetitive Elements May Comprise Over Two-Thirds of the Human Genome | PLOS Genetics - For Physicians," *PLoS genetics*, vol. ٧, no. ١٢, p. e١٠٠٢٣٨٤, ٢٠١١.

[٤] S. A. McCarroll and D. M. Altshuler, "Copy-number variation and association studies of human disease," *Nature Genetics*, vol. ٣٩, no. ٧S, pp. S٣٧–S٤٢, ٢٠٠٧.

[٥] A. J. Sharp *et al.*, "Segmental Duplications and Copy-Number Variation in the Human Genome," *The American Journal of Human Genetics*, vol. ٧٧, no. ١, pp. ٧٨–٨٨, ٢٠٠٥.

[٦] H. Nuytten *et al.*, "Accurate determination of copy number variations (CNVs): Application to the α- and β-defensin CNVs," *Journal of Immunological Methods*, vol. ٣٤٤, no. ١, pp. ٣٥–٤٤, ٢٠٠٩.

[٧] Ian Goodfellow, Y. Bengio, and A. Courville, "Deep learning," *Nature Methods*, vol. ١٣, no. ١, p. ٣٥, ٢٠١٧.

[٨] Y. Park and M. Kellis, "Deep learning for regulatory genomics," *Nature Biotechnology*, vol. ٣٣, no. ٨, pp. ٨٢٥–٨٢٦, ٢٠١٥.

[٩] D. Pinkel *et al.*, "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nature Genetics*, vol. ٢٠, no. ٢, pp. ٢٠٧–٢١١, ١٩٩٨.

[١٠] B. E. Stranger *et al.*, "Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes," *Science*, vol. ٣١٥, no. ٥٨١٣, pp. ٨٤٨–٨٥٣, ٢٠٠٧.

[١١] M. E. Hurles *et al.*, "Origins and functional impact of copy number variation in the human genome," *Nature*, vol. ٤٦٤, no. ٧٢٨٩, pp. ٧٠٤–٧١٢, ٢٠١٠.

[١٢] G. JT *et al.*, "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes.," *Nature*, vol. ٤٥٩, no. ٧٢٤٦, p. ٥٦٩, ٢٠٠٩.

[١٣] J. van de Leemput *et al.*, "Genotype, haplotype and copy-number variation in worldwide human populations," *Nature*, vol. ٤٥١, no. ٧١٨١, pp. ٩٩٨–١٠٠٣, ٢٠٠٨.

[١٤] D. Editor, "GIW ٢٠١٦ Submission ٩٦ - Cover Letter DeepGene : An Advanced Cancer Type Classifier Based on Deep Learning and Somatic Point Mutations,"

vol. ١٧, no. ١٧, p. ٤٧٦, ٢٠١٦.
[١٥] cBioPortal, "cBioPortal for Cancer Genomics. Retrieved ٠٥/١٠/٢٠١٨ from Memorial Sloan Kettering Cancer Center (MSK) on the World Wide Web: http://cbio.mskcc.org/cancergenomics/pancan_tcga/ Release ٢/٤/٢٠١٣."
[١٦] T. Kobayashi, A. Hidaka, and T. Kurita, "Selection of Histograms of Oriented Gradients," in *Neural Information Processing*, ٢٠٠٨, pp. ٥٩٨–٦٠٧.
[١٧] J. H. Wilson and T. Hunt, *Molecular biology of the cell : a problems approach*. Garland Science, ٢٠٠٢.
[١٨] J. Watson, *Dna the Secret of Life*. Knopf, ١٩٠٠.
[١٩] Coburn, Alvin F. in biology Coburn and medicine, "Oswald Theodore Avery and DNA," vol. ١٢, no. ٤, pp. ٦٢٣–٦٣٠, ١٩٦٩.
[٢٠] C. Freudenrich, "How DNA Works. Retrieved ١٢/١١/٢٠١٨ from howstuffworks.com on the World Wide Web: https://science.howstuffworks.com/life/cellular-microscopic/dna.htm."
[٢١] L. Antonov, *Structural Bioinformatics ٢٠١٣*, vol. ٤٤. John Wiley & Sons, ٢٠١٣.
[٢٢] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nature Structural Biology*, vol. ١٠, no. ١٢, pp. ٩٨٠–٩٨٠, ٢٠٠٣.
[٢٣] C. R. Calladine and H. Drew, *Understanding DNA: the molecule and how it works*. Academic press, ١٩٩٧.
[٢٤] E. H. Corder *et al.*, "Gene dose of apolipoprotein E type ٤ allele and the risk of Alzheimer's disease in late onset families," *Science*, vol. ٢٦١, no. ٥١٢٣, pp. ٩٢١–٩٢٣, ١٩٩٣.
[٢٥] C. Roche-Lestienne *et al.*, "Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI٥٧١, and they can pre-exist to the onset of treatment," *Blood*, vol. ١٠٠, no. ٣, pp. ١٠١٤–١٠١٨, ٢٠٠٢.
[٢٦] B. E. Stranger *et al.*, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science*, vol. ٣١٥, no. ٥٨١٣, pp. ٨٤٨–٨٥٣, ٢٠٠٧.
[٢٧] and N. N. Haykin, Simon, "A comprehensive foundation," *Neural Networks*, vol. ٢, no. ٢٠٠٤, p. ٤١, ٢٠٠٤.
[٢٨] The Mendeley Support Team, *Neural network design*. Martin Hagan, ٢٠١١.
[٢٩] S. Hochreiter, "Long Short-term Memory," vol. ٩, no. ٨, pp. ١٧٣٥–١٧٨٠, ٢٠١٦.
[٣٠] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. ٦, no. ٢, pp. ١٠٧–١١٦, ١٩٩٨.
[٣١] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: an extension of MNIST to handwritten letters," *Internaltional Joint Conference On Neural Networks (IJCONN)*, ٢٠١٧.
[٣٢] Colah, "Understanding LSTM Networks. Retrieved ١٢/١٠/٢٠١٨ from github.io on the World Wide Web: http://colah.github.io/posts/٢٠١٥-٠٨-Understanding-LSTMs."
[٣٣] S. Yan, " Understanding LSTM and its diagrams. Retrieved ١٢/١٣/٢٠١٨ from

github.io on the World Wide Web: https://medium.com/mlreview/understanding-lstm-and-its-diagrams-٣٧e٢f٤٦f١٧١٤."

[٣٤] A. M. Fraser and H. L. Swinney, "http://dx.doi.org/١٠,١١٠٣/PhysRevA.٣٣,١١٣٤ Independent coordinates for strange attractors from mutual information}," *Phys. Rev. A*, vol. ٣٣, no. ٢, pp. ١١٣٤–١١٤٠, ١٩٨٦.

[٣٥] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. I. transactions on M. I. Suetens, "Multimodality Image Registration by Maximization of Mutual Information," vol. ١٦, no. ٢, pp. ١–١٢, ٢٠٠٩.

[٣٦] Hauchan Peng, Fuhui Long, and C. Ding, "Feature selection based on joint mutual information: Criteria of Max-Dependecy, Max-Relevance, and Min-Redundancy," *Ieee Transactions on Patterns Analysis and Machine Intelligence*, vol. ٢٧, no. ٨, pp. ١٢٢٦–١٢٣٨, ١٩٩٩.

[٣٧] Z. Jiao, X. Gao, Y. Wang, and J. Li, "Neurocomputing A deep feature based framework for breast masses classi fi cation," *Neurocomputing*, vol. ١٩٧, pp. ٢٢١–٢٣١, ٢٠١٦.

[٣٨] C. K. Kuhl *et al.*, "Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer," *Journal of clinical oncology*, vol. ٢٣, no. ٣٣, pp. ٨٤٦٩–٨٤٧٦, ٢٠٠٥.

[٣٩] Y. Liu *et al.*, "Detecting Cancer Metastases on Gigapixel Pathology Images," ٢٠١٧.

[٤٠] Y. Liu *et al.*, "Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection," *Archives of Pathology & Laboratory Medicine*, p. arpa.٢٠١٨-٠١٤٧-OA, ٢٠١٨.

[٤١] D. F. Steiner *et al.*, "Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer," *American Journal of Surgical Pathology*, vol. ٤٢, no. ١٢, pp. ١٦٣٦–١٦٤٦, ٢٠١٨.

[٤٢] N. Coudray, A. L. Moreira, T. Sakellaropoulos, D. Fenyö, N. Razavian, and A. Tsirigos, "Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning Keywords," vol. ٢٤, no. ١٠, p. ١٥٥٩, ٢٠١٧.

[٤٣] A. C. F. Barata *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. ٢٠١٥–Novem, no. ٧٦٣٩, pp. ٧١١–٧١٥, ٢٠١٧.

[٤٤] D. M. Hawkins, "The Problem of Overfitting," *Journal of Chemical Information and Computer Sciences*, vol. ٤٤, no. ١, pp. ١–١٢, ٢٠٠٤.

[٤٥] G. Cawley, "Over-fitting in Model Selection and Its Avoidance Example of Over-fitting in Training," in *IDA*, ٢٠١٢, no. October, p. ١.

[٤٦] I. Alexander, I. V Tetko, and T. David, "Neural network studies. ١. Comparison of overfitting and overtraining," vol. ٣٥, no. ٥, pp. ٨٢٦–٨٣٣, ١٩٩٥.

[٤٧] I. Sutskever, G. Hinton, A. Krizhevsky, and R. R. Salakhutdinov, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. ١٥, no. ١, pp. ١٩٢٩–١٩٥٨, ٢٠١٤.

[٤٨] E. Rheinbay *et al.*, “Recurrent and functional regulatory mutations in breast cancer,” *Nature*, vol. ٥٤٧, no. ٧٦٦١, pp. ٥٥–٦٠, ٢٠١٧.

[٤٩] Y. Yang, P. Li, X. Ding, X. Ming, and P. R. Guo, “MiR-١٤٥ inhibits proliferation of primary colon adenocarcinoma cells via induction of apoptosis, cell cycle arrest and inhibition of cell migration,” vol. ١٧, no. ٨, pp. ١٤٩١–١٤٩٥, ٢٠١٨.

[٥٠] L. Alizadeh, A. Gorizan, M. Akbari Dana, and S. K. Ghaemi, “Immunotherapy of Glioblastoma Multiforme Tumors: From Basic to Clinical Trial Studies,” vol. ٣, no. ٢, pp. ٧٧–٨٤, ٢٠١٥.

[٥١] D. R. Laks, K. Visnyei, and H. I. Kornblum, “Brain tumor stem cells as therapeutic targets in models of glioma,” *Yonsei Medical Journal*, vol. ٥١, no. ٥, pp. ٦٣٣–٦٤٠, ٢٠١٠.

[٥٢] P. B. Dirks, “Brain tumor stem cells: The cancer stem cell hypothesis writ large,” *Molecular Oncology*, vol. ٤, no. ٥, pp. ٤٢٠–٤٣٠, ٢٠١٠.

[٥٣] Y. Yamada *et al.*, “Clinical features of renal cell carcinoma less than ٢٥ millimeters in diameter,” vol. ٩, no. ١٢, pp. ٦٦٣–٦٦٧, ٢٠٠٢.

[٥٤] M. Kulkarni, S. S. Khandeparkar, A. Joshi, and V. Kothikar, “A rare case of renal metastasis from squamous cell carcinoma of the cervix,” *Journal of Mid-life Health*, vol. ٧, no. ٢, p. ٩٤, ٢٠١٦.

[٥٥] G. Scelo and T. L. Larose, “Epidemiology and Risk Factors for Kidney Cancer,” *Journal of Clinical Oncology*, vol. ٣٦, no. ٣٦, pp. ٣٥٧٤–٣٥٨١, ٢٠١٨.

[٥٦] PDQ Screening and Prevention Editorial Board, “Ovarian, Fallopian Tube, and Primary Peritoneal Cancer Prevention (PDQ): Patient Version,” in *PDQ Cancer Information Summaries [Internet]*, National Cancer Institute (US), ٢٠١٨.

[٥٧] D. Z. Chu *et al.*, “Peritoneal carcinomatosis from non-gynecologic malignancies: Results of the EVOCAPE ١ multicentric prospective study,” *Cancer*, vol. ٨٨, no. ٢, pp. ١٩١–١٩٧, ١٩٨٩.

[٥٨] R. FaraJi, N. Esmaeilpour, and B. Behfar, “Association of endometrial hyperplasia with some risk factors,” ٢٠٠٩.

[٥٩] M .Abadi, Ashish Agarwal, and Paul Barham, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.” ٢٠١٥.

[٦٠] F. Chollet and others, “Keras.” GitHub, ٢٠١٥.

[٦١] “Bethesda (MD): National Library of Medicine (US); ٢٠٠٧ - [updated ٢٠١٥ Oct ٢;٢٠١٩/٧/١]. Available from: https://ghr.nlm.nih.gov/gene.”

[٦٢] F. Carneiro, C. Oliveira, G. Suriano, and R. Seruca, “Molecular pathology of familial gastric cancer, with an emphasis on hereditary diffuse gastric cancer,” *Journal of Clinical Pathology*, vol. ٦١, no. ١, pp. ٢٥–٣٠, ٢٠٠٨.

[٦٣] G. Corso *et al.*, “E-cadherin germline mutation carriers: clinical management and genetic implications,” *Cancer and Metastasis Reviews*, vol. ٣٣, no. ٤, pp. ١٠٨١–١٠٩٤, ٢٠١٤.

[٦٤] J. Gronwald *et al.*, “Hereditary breast and ovarian cancer,” *Hereditary Cancer in Clinical Practice*, vol. ٦, no. ٢, pp. ٨٨–٩٨, ٢٠٠٨.

[٦٥] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander, "Analysis Emerging landscape of oncogenic signatures across human cancers," *Nature Publishing Group*, vol. ٤٥, no. ١٠, p. ١١٢٧, ٢٠١٣.

[٦٦] M. Potrony *et al.*, "Increased prevalence of lung, breast, and pancreatic cancers in addition to melanoma risk in families bearing the cyclin-dependent kinase inhibitor ٢A mutation: Implications for genetic counseling," *Journal of the American Academy of Dermatology*, vol. ٧١, no. ٥, pp. ٨٨٨–٨٩٥, ٢٠١٤.

[٦٧] H. Nakagawa, K. Koyama, M. Monden, and Y. Nakamura, "Analysis of APCL, a brain-specific adenomatous polyposis coli homologue, for mutations and expression in brain tumors," *Japanese Journal of Cancer Research*, vol. ٩٠, no. ٩, pp. ٩٨٢–٩٨٦, ١٩٩٩.

[٦٨] Z.-G. Zhuang, G.-H. Di, Z.-Z. Shen, J. Ding, and Z.-M. Shao, "Enhanced Expression of LKB١ in Breast Cancer Cells Attenuates Angiogenesis, Invasion, and Metastatic Potential," *Molecular Cancer Research*, vol. ٤, no. ١١, pp. ٨٤٣–٨٤٩, ٢٠٠٦.

[٦٩] S. N. Wingo *et al.*, "Somatic LKB١ mutations promote cervical cancer progression," *PLoS ONE*, vol. ٤, no. ٤, p. e٥١٣٧, ٢٠٠٩.

[٧٠] D. Grahame Hardie and A. Dario R, "LKB١ and AMPK and the cancer-metabolism link –ten years after," *BMC Biology*, vol. ١١, no. ١, p. ٣٦, ٢٠١٣.

[٧١] Y. Varghaiyan, A. Aflatoonian, A. Mirghanizadeh, and S. %J S. Najafi, "Evaluation of Th١ and Th١٧ cells cytokines in cell culture stimulated in women with recurrent spontaneous abortion," vol. ٢١, no. ٤, pp. ٥٠٥–٥١٣, ٢٠١٣.

[٧٢] Z. You, Y. Dong, X. Kong, Y. Zhang, R. L. Vessella, and J. Melamed, "Differential expression of IL-١٧RC isoforms in androgen-dependent and androgen-independent prostate cancers.," *Neoplasia (New York, N.Y.)*, vol. ٩, no. ٦, pp. ٤٦٤–٧٠, ٢٠٠٧.

[٧٣] Y. Zhang *et al.*, "Pinin interacts with C-terminal binding proteins for RNA alternative splicing and epithelial cell identity of human ovarian cancer cells," *Oncotarget*, vol. ٧, no. ١٠, p. ١١٣٩٧, ٢٠١٦.

[٧٤] C. N. Birts *et al.*, "Expression of CtBP family protein isoforms in breast cancer and their role in chemoresistance.," *Biology of the cell / under the auspices of the European Cell Biology Organization*, vol. ١٠٣, no. ١, pp. ١–١٩, ٢٠١٠.

[٧٥] K. I. Takayama *et al.*, "Androgen-responsive long noncoding RNA CTBP١-AS promotes prostate cancer," *EMBO Journal*, vol. ٣٢, no. ١٢, pp. ١٦٦٥–١٦٨٠, ٢٠١٣.

[٧٦] P. Cristina *et al.*, "The expression levels of the transcriptional regulators p٣٠٠ and CtBP modulate the correlations between SNAIL, ZEB١, E□cadherin and vitamin D receptor in human colon carcinomas," *International Journal of Cancer*, vol. ١١٩, no. ٩, pp. ٢٠٩٨–٢١٠٤, ٢٠٠٦.

[٧٧] K., Yoshihara. *et al.*, "Germline copy number variations in BRCA١-associated ovarian cancer patients," *Genes Chromosomes and Cancer*, vol. ٥٠, no. ٣, pp. ١٦٧–١٧٧, ٢٠١١.

[٧٨] L. Mao-De and X. Jing, “Ribosomal Proteins and Colorectal Cancer,” *Current Genomics*, vol. ٨, no. ١, pp. ٤٣–٤٩, ٢٠٠٧.

[٧٩] N. Sengupta *et al.*, “Analysis of colorectal cancers in British Bangladeshi identifies early onset, frequent mucinous histotype and a high prevalence of RBFOX١ deletion,” *Molecular Cancer*, vol. ١٢, no. ١, p. ١, ٢٠١٣.

[٨٠] H. Huang., W. Wang., C. Chang., K. Lin, T. T Tok., “CDC٢٥A, VAV١, TP٧٣, BRCA١ and ZAP٧٠ gene overexpression correlates with radiation response in colorectal cancer,” *Oncology Reports*, vol. ٢٥, no. ٥, pp. ١٢٩٧–١٣٠٦, ٢٠١١.

[٨١] D. Gasparotto, R. Maestro, S. Piccinin, T. Vukosavijevic, L. Barzan, and S. Sulfaro, “Advances in Brief Overexpression of CDC٢٥A and CDC٢٥B in Head and Neck Cancers ’,” *Accounts of Chemical Research*, vol. ٥٧, no. ١٢, pp. ٢٣٦٦–٢٣٦٩, ١٩٩٧.

[٨٢] D. Cvetković, A. V. Babwah, and M. Bhattacharya, “Kisspeptin/KISSIR system in breast cancer,” *Journal of Cancer*, vol. ٤, no. ٨, pp. ٦٥٣–٦٦١, ٢٠١٣.

[٨٣] Yang, Yiming and Pedersen, Jan O, “A comparative study on feature selection in text categorization,” Icml, vol. ٩٧, no. ٨, pp. ٤١٢-٤٢٠, ١٩٩٧

Abstract

Cancer is one of the most common diseases in recent decades has attracted the attention of many researchers in various fields of science. Treatment of this disease using of common therapies is often costly, or fails, or the patient experiences severe side effects. For this reason, the need to develop new therapies is well felt.

Genomic variations in DNA cause a variety of cancers in humans. The copy number variations or CNV, as one of a variety of mutations in DNA, has caused various cancers in humans. In order to understand the difference between cancers using of CNVs, in this study, we classify six different classes of cancer in humans using levels of CNV and using of deep learning. In recent years, deep learning has been used to diagnose types of cancer such as lung, skin and breast cancers. Many of these methods have used convolutional neural networks to detect cancer. We use LSTM deep neural network for classification. For this purpose, CNV data from ٢٤١٧٤ genes were used as features for building our classifier.

Our results of the experiment show an accuracy of ٩٢٪ in the classification. Then, to analyse the biologically, we identified the genes that had the greatest impact on the development of cancers. For this reason, the maximum dependency and minimum redundancy (mRMR) criterion was used to identify effective genes. Using of mRMR criterion, we identified ٢٠٠ important genes and then classify these genes with our classifier. The accuracy of more than ٨١٪ in the classification indicates the importance of these genes. Finally, we identified ١٠ important genes that had the most effect on the classification. We examined the genetic function of

these genes in order to give statistical results and biological analysis of these genes to the recognition of different types of human cancers and offer suggestions for appropriate treatment for patients.

Keywords: cancer, deep learning, copy number variation, classification

Shahrood University of Technology

Faculty of Computer Engineering

M.Sc Thesis in Artificial Intelligence Engineering

Cancer Detection Based on Deep Learning and Copy Number Variations

By: Saeed Hassani Borzadaran

Supervisors:

Dr. Mohsen Rezvani

Dr. Ali Akbar Pouyan

Advisor:

Dr. Hamid Alinejad Rokney

Dr. Mansoor Fateh

January ۲۰۱۹