Classification of membrane proteins using a deep hybrid model

Milad Vazan

اولین همایش ملی
کاربرد زیست سامانه‌ها و بیوانفورماتیک در علوم زیستی
دانشگاه جیرفت – ۲۸ شهریور ۱۴۰۰

طبقه‌بندی پروتین‌های غشائی با استفاده از یک مدل ترکیبی عمیق: با تاکید بر شبکه‌های کانولوشنی، حافظه طولانی کوتاه-مدت و استفاده از مکانیزم توجه

میلاد وزان*
دانشگاه تبریز، گروه علوم کامپیوتر
E-mail: m.vazan97@ms.tabrizu.ac.ir

چکیده

پروتئین‌های غشائی نقش مهمی در فعالیت‌های حیاتی موجودات زنده دارند. از آن‌جایی که عملکرد پروتئین غشائی ارتباط تنگاتنگی با نوع آن‌ها دارد، پیش‌بینی انواع پروتئین غشایی می‌تواند به تحقیقات در زمینه بیوانفورماتیک کمک کرده و سرنخ‌هایی را برای درک ساختار و عملکرد پروتئین‌ها فراهم کند. علاوه بر این، به‌دلیل اکتشاف گسترده توالی‌های پروتئین، یک مدل طبقه‌بند قوی برای طبقه‌بندی انواع پروتئین غشائی با دقت بالا ضروری است. هدف اصلی این مقاله، طبقه‌بندی پروتیئن‌های غشایی با استفاده از یک مدل ترکیبی مبتنی‌بر یادگیری عمیق و استفاده از مکانیزم توجه در جهت افزایش کارآیی است. در همین راستا از دو مجموعه داده آموزشی و آزمایشی استفاده شد و پس از ارزیابی روش پیشنهادی با روش‌های پیشین مشاهده گردید که مدل ترکیبی یادگیری عمیق به دلیل استفاده از مزایای هر کدام از شبکه‌های عصبی در پردازش توالی‌ها، کارآیی بهتری را در طبقه‌بندی انواع پروتئین غشائی از خود نشان می‌دهد.

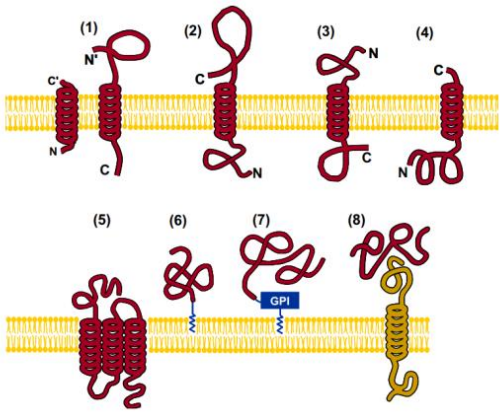
کلمات کلیدی: بیوانفورماتیک، یادگیری عمیق، پروتیئن‌های غشائی، شبکه کانولوشنی، مکانیزم توجه

۱. مقدمه

پروتئین غشائی[1] نشان‌دهنده یک نوع مهم از پروتئین‌ها است که از منظر عملکرد بسیار پرکابرد هستند. آن‌ها در بسیاری از واکنش‌های مهم سلول، از جمله حمل مواد به داخل و خارج از سلول به عنوان حامل، انتقال سیگنال و برهمکنش سلول-سلول شرکت می‌کنند. علاوه بر این، پروتئین‌های غشایی اهمیت ویژه‌ای در درمان‌های دارویی دارند (Guo et al., 2019). پروتئین‌های غشائی را می‌توان بر اساس ماهیت برهم‌کنش غشاء و پروتئین به سه دسته اصلی، پروتئین‌های غشائی سراسری[2] (همچنین پروتئین‌های غشایی[3] نامیده می‌شوند)، متصل به چربی[4] و محیطی[5] طبقه‌بندی کرد (Evaggelos Tamvakis, 2020). بر اساس رابطه مستقیم بین پروتئین‌های غشائی و لایه‌های چربی، این سه دسته را می‌توان به هشت نوع اصلی تقسیم کرد (Guo, L et al., 2019): پروتئین‌های غشائی نوع I، نوع II، نوع III، نوع IV، پروتئین غشائی چندپایه[6]، پروتئین غشائی متصل به زنجیره چربی، پروتئین غشائی متصل به GPI و پروتئین غشائی محیط (شکل ۱).

---

[1] Membrane proteins
[2] Integral membrane proteins
[3] transmembrane proteins
[4] lipid-anchored
[5] peripheral
[6] multipass



شکل ۱. انواع پروتئین غشایی: (۱) غشائی نوع I، (۲) نوع II، (۳) نوع III، (۴) نوع IV، (۵) غشای چندپایه، (۶) غشای متصل به زنجیره چربی، (۷) غشای متصل به GPI، (۸) و غشای محیطی (Chou and Shen, 2007)

از آن‌جایی که عملکرد پروتئین غشائی ارتباط تنگاتنگی با نوع آن‌ها دارد، پیش‌بینی انواع پروتئین غشایی می‌تواند به تحقیقات در زمینه بیوانفورماتیک کمک کند (Guo et al., 2019). با توجه به اهمیت موضوع، پژوهش‌هایی در این خصوص صورت گرفته است. در پژوهش (Guo et al., 2019) برای پردازش داده‌های توالی از شبکه‌های عصبی کانولوشنی[۷] (CNN) یک‌بعدی و شبکه‌های حافظه طولانی کوتاه-مدت دوطرفه[۸] (Bi-LSTM) استفاده شده است که مدل آن‌ها در مقایسه با مدل‌های سنتی یادگیری ماشین نتایج بهتری را بدست آورد. علاوه بر این، آن‌ها یک روش جدید برای بازنمایی بردار[۹] به عنوان جایگزین روش کدگذاری one-hot ارائه کردند که این روش بازنمایی بردار جدید، نه تنها رابطه بین اسیدهای آمینه را به خوبی نشان می‌دهد، بلکه می‌تواند عملکرد پیش‌بینی را به‌طور موثری بهبود بخشید. در پژوهش (Evaggelos Tamvakis, 2020) مدل‌ها مختلف یادگیری ماشین و یادگیری عمیق توسعه داده شد تا عملکرد آن‌ها در طبقه‌بندی پروتئین‌های غشائی مورد ارزیابی قرار گیرد. در پایان این نتیجه حاصل شد که مدل‌های یادگیری عمیق به دلیل توانایی پردازش توالی‌های طولانی‌تر و تشخیص الگوهای محلی بهترین عملکرد را در مقایسه با مدل‌ها یادگیری ماشین سنتی دارند. در پژوهش (Alphonse et al., 2020) ساختار پروتئین‌های غشائی را با استفاده از تکنیک جدید استخراج ویژگی مبتنی‌بر الگوی تترا پپتید[۱۰] تجزیه و تحلیل کردند. در راستای این کار یک ماتریس تکرار رخداد ایجاد می‌شود که از طریق آن بردار ویژگی بدست می‌آید. این بردار ویژگی، الگوی اسیدهای آمینه را توالی پروتئین غشایی ثبت می‌کند. سپس، بردار ویژگی کاهش ابعاد پیدا کرده و از طریق شبکه باور عمیق طبقه‌بندی صورت می‌گیرد. در انتها روش پیشنهادی آن‌ها با دو مجموعه داده مورد ارزیابی قرار گرفت که در مقایسه با سایر روش‌ها پیشرفته نتایج خوبی را بدست آورد.

در این پژوهش، ما مدلی مبتنی‌بر ترکیب شبکه‌های عمیق که شامل شبکه کانولوشنی، شبکه حافظه طولانی کوتاه-مدت (LSTM) و با استفاده از مکانیزم توجه برای طبقه‌بندی و پیش‌بینی نوع پروتئین غشائی طراحی کردیم. در انتها روش پیشنهادی با پژوهش (Evaggelos Tamvakis, 2020) مورد ارزیابی قرار گرفت. در ادامه، پس از مروری کوتاه بر یادگیری عمیق و مدل‌های مناسب آن برای کار با داده‌های توالی (دنباله‌ای)، به معرفی مجموعه داده مورد استفاده در این پژوهش پرداخته و سپس جزئیات روش پیشنهادی و نتایج ارزیابی آن را شرح می‌دهیم.

[7] convolutional neural network
[8] Bidirectional Long Short-Term Memory
[9] vector representation
[10] Tetra Peptide

یادگیری عمیق روشی برای یادگیری محاسباتی مفاهیم سطح بالا در داده‌ها و بازنمایی آن‌ها با استفاده از یک ساختار سلسله مراتبی (لایه‌ها) عمیق است و در زیر مجموعه روش‌های یادگیری ماشین قرار می‌گیرد. وجود لایه‌های مختلف به یادگیری عمیق این امکان را می‌دهد تا بتواند در هر لایه ویژگی‌های خاصی از مساله را کشف کرده و از آن‌ها در جهت تصمیم‌گیری بهتر در حل مساله استفاده کند (وزان، ۱۳۹۹). شبکه‌های عصبی کانولوشنی و بازگشتی مناسب برای پردازش داده‌هایی هستند که حالت توالی دارند. شبکه‌های کانولوشنی یا به اختصار CNN عصبی متشکل از چندین لایه کانولوشن و ادغام هستند که به دنبال آن‌ها یک یا چند لایه متصل کامل اضافه می‌شود (Sorin et al., 2020; Harsha Kadam, 2020) و در پردازش داده هایی با یک ساختار مکانی معلوم و مشبکی مناسب هستند. این شبکه‌ها، ورودی‌هایی که از لحاظ ساختار مکانی نزدیک به یکدیگر باشند را به‌صورت معناداری به یکدیکر ارتباط می‌دهند. این شبکه‌ها نقش مهمی در تاریخچه یادگیری عمیق داشته‌اند و نمونه‌ای موفق و مهم در فهم ما از مطالعه مغز در کاربردهای یادگیری ماشین هستند (وزان م، ۱۳۹۹). شبکه‌های حافظه طولانی کوتاه-مدت که به اختصار LSTM نامیده می‌شوند توانایی یادگیری وابستگی‌های بلندت مدت را دارند. وجود دروازه‌ها در ساختار این شبکه‌ها با آن‌ها این امکان را می‌دهد تا داده‌ها را در مراحل مختلف پیگیری کنند. به عبارت دیگر، وجود دروازه‌ها یک معماری مبتنی‌بر حافظه را در شبکه‌های عصبی ایجاد می‌کند. شبکه‌های حافظه طولانی کوتاه-مدت دوطرفه یا به اختصار Bi-LSTM از دو شبکه LSTM به‌صورت هم‌زمان در زمان آموزش استفاده می‌کند. یکی از شبکه‌ها از طریق تغذیه توالی ورودی به همان صورتی که وارد شده‌اند و دیگری از طریق معکوس ورودی یعنی از انتها به ابتدا آموزش پیدا می‌کند. این تغذیه از دو جهت، افزایش اطلاعات را به‌همراه دارد که نتیجه آن افزایش کارایی شبکه است (Munawar et al., 2021). مکانیزم توجه (Bahdanau et al., 2016) به‌طور موفقیت‌آمیزی توانسته است در راستای افزایش عملکرد برای طیف وسیعی از مسائل در داده‌های حالت توالی مورد استفاده قرار گیرد. به‌طور کلی، مکانیزم توجه در شبکه‌های عصبی روشی برای هدایت فرآیند آموزش است؛ با اطلاع دادن به مدل در مورد این‌که بر چه قسمت‌هایی از ورودی یا ویژگی‌ها باید متمرکز شود تا بتواند پیش‌بینی بهتری را ارائه دهد (Bahuleyan, 2018).

۲. مواد و روش‌ها

مجموعه داده‌های آموزشی و آزمایشی مورد استفاده در این پژوهش برگرفته از پژوهش (Evaggelos Tamvakis, 2020) می‌باشد. مجموعه داده آموزشی در این پژوهش شامل ۵۵۰۲ پروتئین غشائی و مجموعه داده آزمایشی، شامل ۳۲۴۹ پروتئین غشائی است. جدول ۱ نمایش توزیع پروتئین‌ها را برای هر ۸ نوع پروتئین نشان می‌دهد.

جدول ۱. نمایش توزیع پروتئین‌ها در مجموعه داده آموزشی و آزمایشی

نوع پروتئین غشائی	مجموعه داده آموزشی	مجموعه داده آزمایشی
نوع I	۶۲۶	۶۱۰
نوع II	۲۹۹	۳۱۲
نوع III	۴۲	۲۴
نوع IV	۷۳	۴۴
غشای چندپایه	۲۴۳۷	۱۳۱۶
غشای متصل به زنجیره چربی	۴۰۳	۱۵۱
غشای متصل به GPI	۱۷۲	۱۸۲
غشای محیطی	۱۴۵۰	۶۱۰
مجموع	**۵۵۰۲**	**۳۲۴۹**

هر نمونه ورودی در این دو مجموعه داده شامل یک شماره دسترسی، توالی، نوع پروتئین و طول دنباله است. در جدول ۲ نمایشی از ۵ نمونه اول مجموعه داده آموزشی را نشان می‌دهد.

جدول ۲. نمایشی از ۵ نمونه اول در مجموعه داده آموزشی

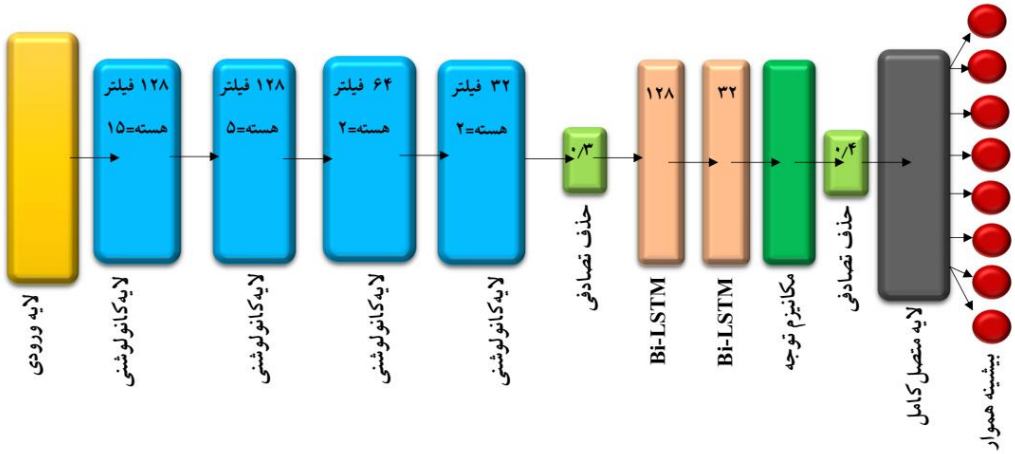
	accNo	Sequence	Type	Length
0	A6NFA1	MHAALAGPLLAALLATARARPQPPDGGQCRPPGSQ...	0	517
1	A8MVS5	MPWTILLFAAGSLAIPAPSIRLVPPYPSSQEDPIHIAC...	0	230
2	A8MVW5	MGQDAFMEPFGDTLGVFQCKIYLLLFGACSGLKV...	0	462
3	B0F2B4	MPAPVPALLCLALALASAQPSPPPPPPFPVVATNY...	0	945
4	B3LS11	MRFSMLIGFNLLTALSSFCAAISANNSDNVEHEQE...	0	225

اکنون به تشریح روش پیشنهادی در حل مساله طبقه‌بندی پروتیئن‌های غشایی با استفاده از شبکه‌های عمیق می‌پردازیم. شبکه‌های عصبی و با عبارت کلی‌تر، الگوریتم‌های طبقه‌بندی توانایی پردازش ورودی متون به‌صورت خام را ندارند. از همین‌رو باید آن‌ها را به‌گونه‌ای نگاشت کرد تا توانایی پردازش آن‌ها را داشته باشند. برای این کار داده باید به بردارهای عددی تبدیل شوند. در جهت بردارسازی[11] داده‌ها روش‌های مختلفی وجود دارد. در این پژوهش ما از روش one-hot استفاده می‌کنیم. در این روش به هر یک از ۲۰ اسیدآمینه یک مقدار صحیح نگاشت می‌شود. شکل ۱ مثالی از این تبدیل را برروی مجموعه داده نشان می‌دهد. از آن‌جایی که برای تغذیه ورودی‌ها به شبکه باید طول همه بردارها یکسان باشد، یک نقطه برش را انتخاب کردیم. بنابراین، پروتین‌هایی را که طول آن‌ها بیش از ۱۵۰۰ باشد را از آن نقطه برش دادیم. انتخاب عدد ۱۵۰۰ به این دلیل است که در مجموعه داده آموزشی ۹۶.۴۷ درصد پروتئین‌ها دارای طولی کم‌تر از ۱۵۰۰ و در مجموعه داده آزمایشی ۹۷.۵۶ درصد این پروتئین‌ها کم‌تر از این طول را دارند. هم‌چنین، برای نمونه ورودی‌هایی که طول آن‌ها کم‌تر ۱۵۰۰ است از تکنیک لایه‌گذاری[12] استفاده کردیم. به این صورت که برای هر توالی کم‌تر از طول ۱۵۰۰ توالی از عدد صفر را قرار داده تا طول بردار آن برابر با ۱۵۰۰ شود.

شکل ۱. مثالی از تبدیل‌گذاری One-Hot برروی مجموعه داده

[11] Vectorizing
[12] padding

بعد از آن که داده‌ها بردارسازی شده و برای پردازش توسط شبکه مناسب شدند، نوبت به تغذیه آن‌ها به شبکه است. تغذیه ورودی‌ها از طریق لایه ورودی صورت می‌گیرد. بعد از لایه ورودی در معماری پیشنهادی، ابتدا از چهار لایه کانولوشنی به‌همراه ادغام حداکثری[۱۳] استفاده شد. اندازه فیلتر این چهار لایه کانولوشنی به ترتیب ۱۲۸، ۱۲۸، ۶۴ و ۳۲ با اندازه هسته‌های ۱۵، ۵، ۲ و ۲ است. پس از آن برای جلوگیری از بیش‌برازش[۱۴] از یک لایه حذف تصادفی[۱۵] با احتمال ۰.۳ استفاده شد. پس از آن دو لایه LSTM به ترتیب با ۱۲۸ و ۳۲ نرون هوشمند در معماری قرار داده شد. بعد از آن‌ها مکانیزم توجه استفاده شد. بار دیگر یک لایه حذف تصادفی با احتمال ۰.۴ بعد از مکانیزم توجه قرار داده شد. در انتها از یک لایه متصل کامل در راستای طبقه‌بندی با ۸ خروجی به تعداد کلاس‌ها قرار داده شد. از آن‌جایی که مساله یک طبقه‌بندی چندگانه[۱۶] است، در لایه خروجی از تابع فعال‌سازی بیشینه‌هموار[۱۷] استفاده شده است. معماری روش پیشنهادی در شکل ۲ قابل مشاهده است.



شکل ۲. معماری روش پیشنهادی

پیاده‌سازی روش پیشنهادی این پژوهش با استفاده از زبان پایتون و استفاده از کتاب‌خانه کراس[۱۸] بر بستر تنسورفلو[۱۹] صورت گرفته است. کراس یک رابط کابری سطح بالا است که به کاربران اجازه می‌دهد با استفاده از مجموعه‌ای ابزار از پیش‌ساخته، به طراحی مدل‌های یادگیری عمیق بپردازند. هم‌چنین به دلیل محدودیت‌های سخت‌افزاری (GPU)، برای آموزش مدل‌ها از گوگل کولب[۲۰] استفاده شده است. آموزش مدل در ۲۰ دوره[۲۱] با اندازه‌های ۲۵۶ و با بهینه‌ساز RMSprop، صورت گرفته است.

۳. نتایج و بحث

ارزیابی در یادگیری ماشین توانایی مدل در برخورد با نمونه‌هایی است که در فرآیند آموزشی آن‌ها را ندیده است. مجموعه داده‌های آزمایشی همان نمونه‌هایی هستند که در فرآیند آموزش شرکت داده نشده‌اند. علاوه براین، برای بهینه‌سازی ابرپارمترهای مدل در

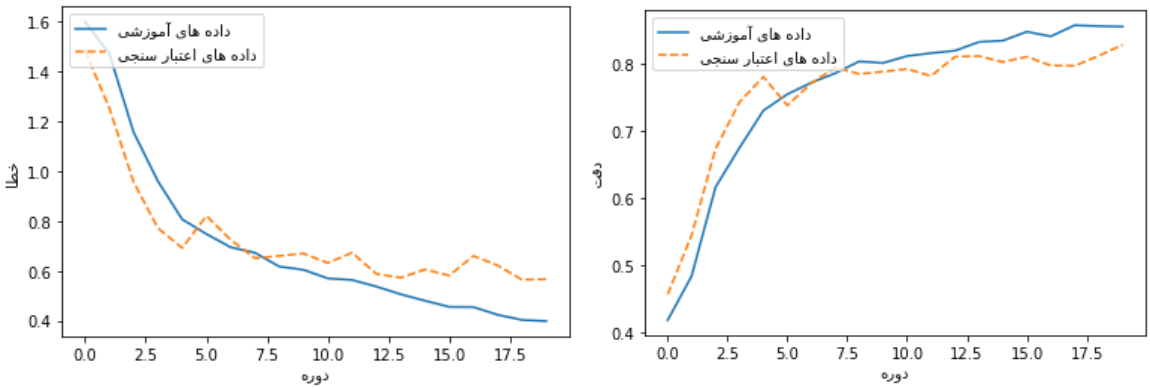
[13] max pooling
[14] overfitting
[15] dropout
[16] multi class classification
[17] softmax
[18] https://keras.io
[19] https://www.tensorflow.org
[20] https://colab.research.google.com/notebooks/intro.ipynb
[21] epochs

حین آموزش به‌طور معمول داده‌های آموزشی به دو قسمت داده‌های آموزشی و اعتبارسنجی تقسیم می‌شوند. از همین‌رو، ما ۳۰۰۰ نمونه اول را به عنوان مجموعه داده آموزشی و بقیه آن‌ها را به عنوان داده‌های اعتبارسنجی تقسیم‌بندی کردیم. داده‌های اعتبارسنجی نیز همانند داده‌های آزمایشی در فرآیند آموزش شرکت نمی‌کنند و تنها برای ارزیابی مدل استفاده می‌شوند. به‌طور معمول در مسائل طبقه‌بندی در راستای ارزیابی مدل‌ها از معیار دقت استفاده می‌شود. دقت، نسبت نمونه‌هایی است که مدل در خروجی به‌درستی طبقه‌بندی کرده است. در جدول ۳ نتایج روش پیشنهادی با سایر روش‌ها در مجموعه داده اعتبارسنجی و آزمایشی قابل مشاهده است.

جدول ۳. مقایسه کارآیی روش پیشنهادی با سایر روش‌ها

مدل	مجموعه داده اعتبارسنجی	مجموعه داده آزمایشی
Wavenet+one-hot (Evaggelos Tamvakis, 2020)	۰.۸۰	۰.۸۴
Wavenet+ Physicochemical (Evaggelos Tamvakis, 2020)	۰.۸۱	۰۸۲
روش پیشنهادی	**۰.۸۳**	**۰.۸۴۵**

نمودار دقت و تابع زیان روش پیشنهادی برروی مجموعه داده آموزشی و اعتبارسنجی در شکل ۳ قابل مشاهده است.



(آ) (ب)

شکل ۴. نمودارهای (آ) دقت و (ب) تابع زیان برروی مجموعه داده آموزشی و اعتبارسنجی

همان‌طور که در جدول ۱ مشاهده می‌شود، روش پیشنهادی در مقایسه با سایر روش‌ها کارآیی بهتری دارد. در مجموعه داد اعتبارسنجی روش پیشنهادی سبب افزایش ۲ درصدی دقت گردیده است. همچنین در مجموعه داده آزمایشی افزایش کارآیی ۰.۵ درصدی را به‌همراه داشته است. با این حال باید ذکر شود که در پژوهش (Evaggelos Tamvakis, 2020) مدل‌های مختلف یادگیری ماشین و یادگیری عمیق با یکدیگر مقایسه شده‌اند و ما در این پژوهش تنها دو مدل با بهترین کارآیی (از دو رویکرد متفاوت بردارسازی) را برای مقایسه انتخاب کرده‌ایم (در مقایسه با روش‌های سنتی یادگیری ماشین و روش‌های بدون ترکیب، روش پیشنهادی بسیار بهتر عمل کرده است). علاوه براین، همچنان که در نمودارهای شکل ۴ مشاهده می‌شود، در روش پیشنهادی روند تغییرات در داده‌های آموزشی و اعتبارسنجی به‌صورت پیوسته نزدیک به یکدیگر می‌باشد. از همین‌رو می‌توان گفت مدل پیشنهای توانسته است از بیش‌برازش جلوگیری کند.

۴. نتیجه‌گیری

در این مقاله ما یک مدل ترکیبی از شبکه‌های عمیق را برای طبقه‌بندی پروتئین‌های غشائی طراحی کردیم. شبکه‌های کانولوشنی توانایی بسیار قدرتمندی در استخراج ویژگی‌های محلی دارند. از طرف دیگر، شبکه‌های حافظه طولانی کوتاه-مدت دوطرفه می‌تواند داده‌ها از هر دو طرف پردازش کرده و اطلاعات مفیدی را بدست آورند. همچنین، مکانیزم توجه می‌تواند به قسمت‌ها مهم در حین آموزش بیشتر متمرکز شود. با ترکیب کردن این شبکه‌ها در راستای بدست آوردن مزایای همه آن‌ها یک معماری ترکیبی از شبکه‌های عمیق را ارائه دادیم. پس از مقایسه با نتایج تجربی با سایر روش‌های پیشین، مشاهده گردید که روش پیشنهادی سبب بهبود عملکرد در طبقه‌بندی پروتئین‌های غشائی شده است.

References

1- Vazan, Milad. (2021). Deep learning: principles, concepts and approaches.
2- Guo, L., Wang, S., Li, M. and Cao z. (2019). Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning. BMC Bioinformatics 20, 700.
3- Evaggelos Tamvakis, P. (2020). Data mining techniques and their applications in biological databases, MSc Thesis, University of Athens.
4- Alphonse, A.S., Mary, N.A.B. and Starvin, M.S. (2020). Classification of membrane protein using Tetra Peptide Pattern. Analytical Biochemistry, 606, p.113845.
5- Sorin, V., Barash, Y., Konen, et al. (2020). Deep Learning for Natural Language Processing in Radiology Fundamentals and a Systematic Review. Journal of the American College of Radiology.
6- Harsha Kadam, S., (2020). Text analysis for email multi label classification, MSc Thesis, University of Gothenburg.
7- Munawar, S., Asif, M., Kabir, B., Pamir, Ullah, A. and Javaid, N. (2021). Electricity Theft Detection in Smart Meters Using a Hybrid Bi-directional GRU Bi-directional LSTM Model. Complex, Intelligent and Software Intensive Systems, pp.297–308.
8- Bahuleyan, H. P. (2018). Natural Language Generation with Neural Variational Models, MSc Thesis, University of Waterloo.
9- Bahdanau, D., Cho, K., Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.
10- Chou, K.-C. and Shen, H.-B. (2007). MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and Biophysical Research Communications, 360(2), pp.339–345.