



Investigating the Interpretability of Deep Learning in Medical Diagnosis

Maryam Firouzi*

CEO, Behbood Bakhsh Behina Company

Nadia Bakhtiari Bahador

R&D Manager, Behbood Bakhsh Behina Company

Abstract

Deep learning has made significant advancements in the field of medicine, with its accuracy often matching or even surpassing human experts in many tasks. However, the "black box" nature of these models poses a fundamental challenge, as their decision-making processes remain opaque, counterintuitive, and difficult to comprehend. This characteristic undermines the trust, transparency, and clinical acceptance of deep learning models. To address this challenge, extensive research has been conducted on model interpretability. This paper provides a comprehensive review of deep learning interpretability in medical diagnostics, including an examination of common interpretability methods, various applications of this approach in disease diagnosis, performance evaluation metrics, and commonly used datasets. Furthermore, the main challenges of interpretability and future research directions in this field are discussed. This review emphasizes the importance of developing transparent and reliable models to enhance medical decision-making and patient safety, while clarifying future research paths in clinical artificial intelligence.

Keywords: clinical artificial intelligence, model interpretability, deep learning, clinical decision-making

Received: 01/June/2026

Accepted: 15/June/2026

eISSN: 3115-7610

ISSN: 3115-7572

بررسی قابلیت تفسیر یادگیری عمیق در تشخیص پزشکی

مریم فیروزی*

مدیرعامل شرکت بهبود بخش بهینا

نادیا بختیاری بهادر

مدیر تحقیق و توسعه شرکت بهبود بخش بهینا

چکیده

یادگیری عمیق در حوزه پزشکی پیشرفت‌های چشمگیری داشته و در بسیاری از موارد دقت آن با متخصصان انسانی برابری کرده یا حتی از آن‌ها فراتر رفته است. با این حال، ساختار «جعبه سیاه» این مدل‌ها چالشی اساسی ایجاد می‌کند، زیرا فرایند تصمیم‌گیری آن‌ها مبهم، غیرشهودی و دشوار برای درک است. این ویژگی موجب کاهش اعتماد، شفافیت و پذیرش بالینی مدل‌های یادگیری عمیق می‌شود. به منظور رفع این چالش، پژوهش‌های گسترده‌ای در زمینه تفسیرپذیری مدل‌ها صورت گرفته است. در این مقاله، مروری جامع بر تفسیرپذیری یادگیری عمیق در تشخیص‌های پزشکی ارائه می‌گردد؛ از جمله بررسی روش‌های رایج تفسیرپذیری، کاربردهای گوناگون این رویکرد در تشخیص بیماری، معیارهای ارزیابی عملکرد و مجموعه داده‌های معمول مورد استفاده. افزون بر این، چالش‌های اصلی تفسیرپذیری و جهت‌گیری‌های آینده پژوهش در این زمینه مورد بحث قرار گرفته‌اند. این بررسی بر اهمیت توسعه مدل‌های شفاف و قابل اعتماد برای ارتقای تصمیم‌گیری پزشکی و افزایش امنیت بیماران تأکید دارد و مسیرهای تحقیقاتی آینده در زمینه هوش مصنوعی بالینی را روشن می‌سازد.

کلیدواژه‌ها: هوش مصنوعی بالینی، تفسیرپذیری مدل‌ها، یادگیری عمیق، تصمیم‌گیری بالینی

مقدمه

پیشرفت‌های شگرف در حوزه هوش مصنوعی، به‌ویژه با بهره‌گیری از شبکه‌های عصبی ژرف (یادگیری عمیق)، انقلابی در حوزه مراقبت‌های بهداشتی و تشخیص پزشکی ایجاد کرده است. مدل‌های یادگیری ژرف اکنون می‌توانند تصاویر پزشکی (مانند رادیوگرافی، MRI و پاتولوژی) را با دقتی رقابتی و حتی گاهی فراتر از دقت متخصصان انسانی، تحلیل کنند. این توانایی نویدبخش تشخیص زودهنگام‌تر، شخصی‌سازی درمان‌ها و کاهش خطاهای تشخیصی است. با این وجود، پذیرش گسترده این فناوری‌ها در محیط بالینی با مانعی بنیادین روبه‌رو است: اعتماد. پزشکان و متخصصان سلامت، که مسئولیت نهایی جان بیمار را بر عهده دارند، نمی‌توانند صرفاً براساس یک پاسخ “بله” یا “خیر” از الگوریتم پیچیده‌ای تصمیم‌گیری کنند. مدل‌های ژرف، به دلیل ساختار سلسله‌مراتبی و تعداد بالای پارامترهایشان، اغلب مانند یک جعبه سیاه عمل می‌کنند؛ آن‌ها می‌توانند تشخیص دقیقی ارائه دهند، اما سازوکار استدلالی منجر به آن نتیجه را پنهان نگه می‌دارند. این عدم شفافیت، چالش‌های اخلاقی و حقوقی بزرگی را مطرح می‌کند: چگونه می‌توان مدلی را در صورت اشتباه بازخواست کرد؟ چگونه می‌توان از صحت و عدالت آن در برابر سوگیری‌های داده‌ای اطمینان حاصل کرد؟

اینجاست که تبیین‌پذیری هوش مصنوعی نقشی حیاتی ایفا می‌کند. تبیین‌پذیری، که شاخه‌ای کلیدی از هوش مصنوعی مسئولانه است، ابزارهایی را فراهم می‌آورد تا بتوانیم «چرا»ی یک پیش‌بینی پزشکی را کشف کنیم. هدف، نمایش مناطقی از تصویر است که مدل به آن‌ها بیشترین توجه را داشته است (مانند مکان تومور)، یا شناسایی ویژگی‌هایی در داده‌های بیمار که بیشترین وزن را در تشخیص نهایی داشته‌اند. این مقاله، به بررسی روش‌های پیشرو در حوزه تبیین‌پذیری برای مدل‌های پزشکی پرداخته است و بحث می‌کند که چگونه افزایش شفافیت، نه تنها اعتماد بالینی را تقویت می‌کند، بلکه مسیر را برای ادغام امن و مؤثر هوش مصنوعی در فرایندهای حیاتی تشخیص پزشکی هموار می‌سازد.

طبقه‌بندی روش‌های تفسیرپذیری

اگرچه تاکنون مفهومی واحد و استاندارد به منظور تفسیرپذیری تعریف نشده است، مطالعات متعددی مجموعه‌ای غنی از روش‌ها و چارچوب‌ها را برای این منظور پیشنهاد داده‌اند [۱]. با این حال، می‌توان تفسیرپذیری را به‌طور کلی به‌عنوان فرایندی تعریف کرد که جزئیات و دلایل منطقی نهفته در نتایج نهایی مدل‌های پیچیده را به گونه‌ای ارائه می‌دهد که درک آن‌ها برای کاربر نهایی آسان‌تر شود. روش‌های تفسیرپذیری معمولاً براساس زمان اعمال و دامنه اثرشان طبقه‌بندی می‌شوند. به‌طور کلی، می‌توان آن‌ها را به دو دسته اصلی تقسیم کرد: تفسیرپذیری پیش از وقوع و تفسیرپذیری پس از وقوع [۲]. علاوه بر این، طبقه‌بندی پس از وقوع خود به دو زیرمجموعه مهم تقسیم می‌شود: تفسیرپذیری سراسری که هدف آن درک رفتار کلی مدل در تمام ورودی‌هاست، و تفسیرپذیری محلی که بر توضیح یک پیش‌بینی یا تصمیم خاص تمرکز دارد. در این بخش، با تمرکز بر کاربردهای بالینی، روش‌های تفسیرپذیری که معمولاً در حوزه پزشکی استفاده می‌شوند را در سه دسته اصلی طبقه‌بندی کرده و شرح مفصلی از هریک ارائه خواهد شد. لازم به ذکر است رویکردهای ارائه‌شده لزوماً مطلق نیستند؛ با توجه به ویژگی‌های ذاتی روش‌های تفسیرپذیری، ممکن است ابزاری خاص به‌طور هم‌زمان به دسته‌های متعددی تعلق گیرد یا مرزهای بین این طبقه‌بندی‌ها هم‌پوشانی داشته باشد.

روش‌های تفسیرپذیری بر تجسم

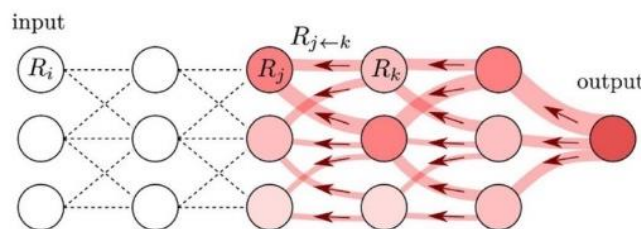
روش «YSIS» که یک تجسم کیفی امیدبخش و دیداری است، راهکاری برای تفسیر عمیق مکانیسم‌های درونی مدل‌های یادگیری ژرف ارائه می‌کند. این روش نه تنها نواحی پیش‌بینی را مستقیماً آشکار می‌سازد، بلکه به مثابه یک ابزار تأیید نیز عمل می‌کند تا بررسی شود آیا نتایج پیش‌بینی با مقادیر واقعی هم‌خوانی دارند یا خیر [۳].

در این راستا، عمده تمرکز این پژوهش بر سه شیوه برجسته‌سازی و تجسم مبتنی بر تمرکز قرار گرفته است:

۱. روش‌های پس‌انتشار (بازگشت به عقب): این شیوه‌ها مبتنی بر محاسبه و انتشار تغییرات به عقب در شبکه هستند.

۲. روش‌های نقشه‌برداری فعال‌سازی کلاسی (CAM): این دسته‌بندی بر تولید نقشه‌هایی متمرکز است که نشان‌دهنده اهمیت نواحی مختلف در ورودی هستند.

۳. روش‌های تفسیرپذیری مبتنی بر اختلال (اختلال): این رویکردها با ایجاد تغییرات هدفمند در ورودی، تأثیر آن بر خروجی مدل را می‌سنجند.



شکل ۱. فرایند LRP. هر نورون به همان اندازه که از لایه بالاتر دریافت کرده است، به لایه پایین‌تر توزیع می‌شود

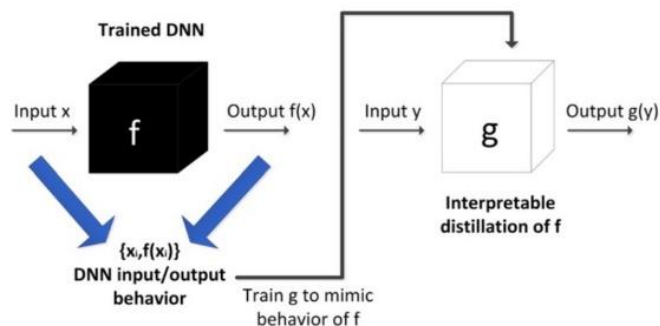
روش‌های تفسیرپذیری جایگزین

پایه‌سازی روش‌های تفسیرپذیری جایگزین نسبتاً آسان است. این روش‌ها در پی آموزش دادن مدل‌های تفسیرپذیر (مانند درخت‌های تصمیم‌گیری کم‌عمق، مدل‌های خطی و غیره) هستند تا رفتار مدل‌های «جعبه سیاه» را تقلید کنند. در این بخش، به بررسی دو رویکرد اصلی تفسیرپذیری جایگزین متقابل پرداخته می‌شود.

اگنوستیک تفسیرپذیر محلی (LIME) یک شیوه تفسیری موضعی (محلی) پرطرفدار است که می‌تواند شبکه‌های پیچیده را با تقریب زدن توسط مدل‌های ساده‌تر، توضیح دهد. به طور مشخص، محققان یک مجموعه همسایه از ورودی‌ها را انتخاب می‌کنند، سپس این داده‌ها را به شبکه‌های ساده‌تر خورانده و تلاش می‌کنند نتایج واقعی «جعبه سیاه» را تا حد امکان شبیه‌سازی کنند. در نهایت، با تحلیل شبکه‌های ساده تفسیرپذیر، تفسیری برای مدل‌های یادگیری عمیق فراهم می‌گردد [۴].

$$\text{LIME}(x) = \arg \min (L(f.g.\Pi_x) + \Omega(g)) \quad (1)$$

با وجود اینکه روش LIME شهودی، انعطاف‌پذیر و قابل فهم است، اما تنها تقریب موضعی از مدل اصلی ارائه می‌دهد. علاوه بر این، برای هر نمونه ورودی، این روش نیازمند آموزش مجدد مدلی قابل تفسیر است تا با نتایج تصمیم‌گیری مدل اصلی انطباق یابد. در نتیجه، کارایی محاسباتی این روش نسبتاً پایین است.



شکل ۲. تصویر روش‌های تفسیر جایگزین

روش‌های تفسیر ذاتی

به‌طور کلی، روش‌های تفسیرپذیری ذاتی این مفهوم را منتقل می‌کنند که مدل می‌تواند به‌خودی‌خود تفسیر شود، که غیر بحث‌برانگیز، ساده و قابل فهم است [۵]. به همین دلیل، در اینجا به بررسی چندین روش تفسیرپذیری ذاتی رایج پرداخته شده است. این روش را می‌توان به دو دسته تقسیم کرد: توجه نرم و توجه سخت. به‌صورت شهودی، مدل توجه نرم قابل اشتقاق (قابل تمایز) است؛ این روش به هر بخش از تصویر توجه بیشتری نشان می‌دهد و وزن میانگین تمام بخش‌ها را به‌دست می‌آورد. در مقابل، توجه سخت تنها بر بخش‌های خاصی از تصویر متمرکز می‌شود و هر بخش یا نادیده گرفته می‌شود یا برای دستیابی به بردار زمینه مورد استفاده قرار می‌گیرد.

علاوه بر این، توجه به خود نیز یک مکانیسم توجه داغ است که در ابتدا برای رسیدگی به وظایف پردازش زبان طبیعی طراحی شد، اما به تدریج برای انجام وظایف بینایی رایانه‌ای با عملکرد عالی نیز به کار گرفته شد. اخیراً، بحث‌هایی در مورد اینکه آیا مکانیسم توجه ذاتاً قابل تفسیر است یا خیر، مطرح شده است [۶]. بنابراین، تفسیرپذیری مکانیسم توجه ممکن است ارزش بررسی بیشتر در آینده را داشته باشد.

روش استخراج مبتنی بر قانون [۷] می‌تواند با استخراج قوانین قابل تفسیر از مدل آموزش‌دیده، قابلیت تفسیر را فراهم کند. این روش، مدل پیچیده را به‌عنوان نقطه آغازین در نظر می‌گیرد و با استفاده از مجموعه‌ای از قوانین فهم‌پذیر، توصیفات نمادین قابل تفسیر تولید می‌کند، یا مدل‌های قابل تفسیر (مانند درخت‌های تصمیم‌گیری، مدل‌های خطی و غیره) را استخراج می‌نماید تا قابلیت‌های تصمیم‌گیری مشابهی با مدل اصلی داشته باشد. با این حال، از آنجا که قوانین قابل تفسیر از یک مدل بزرگ استخراج می‌شوند، دقت مدل استخراج‌شده اغلب به‌اندازه کافی بالا نیست و در نتیجه، این روش تنها می‌تواند یک توضیح تقریبی ارائه دهد. اطلاعات بیشتر در مورد الگوریتم‌های تفسیر مبتنی بر قانون، مانند RuleFit، Node Harvest و غیره، در [۸] مورد بحث قرار گرفته است.

کاربرد روش‌های تفسیرپذیری بر تشخیص بیماری‌ها

این یک واقعیت است که مدل‌های یادگیری عمیق با کارایی و دقت بالا، از روش‌های سنتی در حوزه پزشکی پیشی گرفته‌اند. کاربردهای متعددی برای کمک به پزشکان در تصمیم‌گیری دقیق‌تر طراحی شده‌اند.

بررسی تصاویر شبکه‌ای فوندوس با چشم غیرمسلح، به‌ویژه برای مویرگ‌های ریز چشم، کاری طولانی و خسته‌کننده است. به همین دلیل، توسعه مدل‌های یادگیری عمیق تفسیرپذیر برای کمک به پزشکان در تشخیص بیماری‌های چشم ضروری است [۹].

تشخیص بیماری‌های گلوکوم

گلوکوم یکی دیگر از بیماری‌های برگشت‌ناپذیر چشم است که می‌تواند منجر به از دست دادن بینایی شود. برای تشخیص دقیق گلوکوم، بیشتر محققان روش‌های مبتنی بر توجه را برای مکان‌یابی و شناسایی مناطق ضایعه به جای روش‌های مبتنی بر CAM اتخاذ کرده‌اند. این به این دلیل است که شواهد بالینی برای تشخیص گلوکوم عمدتاً بر روی دیسک بینایی متمرکز است و نیاز به مکان‌یابی مناطق ضایعه کوچک‌تر دارد. به‌عنوان مثال [۱۰]، یک مدل یادگیری عمیق مبتنی بر توجه به نام AG-CNN برای تشخیص گلوکوم طراحی کردند.

چالش‌ها و مسیرهای آینده در تفسیرپذیری سامانه‌های هوش مصنوعی پزشکی

با وجود پیشرفت‌های چشمگیر در توسعه روش‌های توضیح‌دهی نتایج هوش مصنوعی، به‌ویژه در حوزه تشخیص‌های پزشکی، هنوز موانع مهمی بر سر راه پذیرش عملی و فراگیر این سامانه‌ها در محیط بالینی روزمره وجود دارد. برای آنکه این سیستم‌ها علاوه بر دقت بالا، قابل اعتماد و پاسخ‌گو باشند، رفع چالش‌های زیر ضروری است:

۱. اعتبار و استحکام تفسیرها: بسیاری از روش‌های توضیحی که پس از اجرای مدل اعمال می‌شوند (نظیر روش‌های جایگزین‌ساز)، در برابر کمترین تغییرات در ورودی‌ها (مثل نویزهای کوچک یا تغییرات عمدی اندک) تفاسیر بسیار متفاوتی ارائه می‌دهند. در فضای درمانی، تغییر اندک در داده نباید دلیلی برای تغییر بنیادین در توجیه تشخیص باشد. نیازمند معیارهایی استاندارد برای سنجش استحکام این توضیحات در برابر تغییرات داده‌ای هستیم [۱۱].

۲. گسست میان زبان فنی و دانش پزشکی: تفاسیر تولید شده، اغلب به شکل نقشه‌های اهمیت پیکسل در تصاویر یا وزن‌های ویژگی، برای متخصصان غیرفنی (پزشکان) مبهم و فاقد ارزش کاربردی هستند. یک پزشک نیازمند درک مفاهیم بالینی است (مثلاً «فعالیت التهابی در ناحیه مشخصی مشاهده می‌شود») نه صرفاً فهمیدن اینکه «ویژگی الف بیشترین تأثیر را داشته است». بنابراین، ترجمه نتایج فنی به چارچوب‌های مفهومی پذیرفته شده در دانش پزشکی حیاتی است.

۳. توازن میان کارایی و توضیح‌پذیری: اغلب، پیچیده‌تر کردن مدل برای دستیابی به بالاترین سطح دقت، به بهای از دست رفتن شدید قابلیت توضیح‌دهی تمام می‌شود (و بالعکس). مدل‌های بسیار ساده نیز ممکن است برای تشخیص‌های ظریف پزشکی کافی نباشند. باید به دنبال رویکردهایی بود که مدل‌ها را در مرز بهینه بین کارایی و شفافیت قرار دهند.

۴. مسئله پاسخ‌گویی و مسئولیت: در صورت بروز خطا یا تشخیص اشتباه ناشی از سامانه‌های هوش مصنوعی، تعیین مرجع مسئول دشوار است. نبود شفافیت در فرایند تصمیم‌گیری، بازرسی دقیق و ردیابی ریشه خطا را مختل می‌کند.

بحث و نتیجه‌گیری

قابلیت تفسیر، به‌ویژه در حوزه‌های حساس و حیاتی مانند پزشکی، به کانون توجه پژوهشگران و متخصصان تبدیل شده است. این قابلیت نه تنها به عنوان ابزاری فنی به منظور شفاف‌سازی ساختار درونی مدل‌های پیچیده «جعبه سیاه» عمل می‌کند، بلکه عنصری حیاتی برای جلب اعتماد کاربر نهایی (پزشکان و بیماران) به سیستم‌های مبتنی بر هوش مصنوعی است. هنگامی که تصمیمات بر سلامت انسان تأثیر می‌گذارند، درک چرایی یک پیش‌بینی، به اندازه خودِ پیش‌بینی اهمیت دارد.

در این مقاله، مروری جامع بر روش‌های قابلیت تفسیر که به‌طور گسترده در حوزه پزشکی به کار گرفته شده‌اند، ارائه گردید. کاربردهای تفسیری براساس دسته‌بندی‌های مختلف بیماری‌ها (مانند تصویربرداری، تشخیص، پیش‌آگهی و غیره) به‌صورت ساختاریافته دسته‌بندی شد. این بررسی نشان داد اگرچه پیشرفت‌های چشمگیری حاصل شده است، اما همچنان چالش‌های قابل توجهی در مسیر وجود دارد که نیازمند توجه فوری هستند [۱۲].

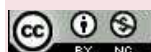
ازجمله چالش‌های اصلی می‌توان به انتقال‌پذیری روش‌های تفسیری از یک مدل یا مجموعه داده به مجموعه دیگر، پایداری تفاسیر در برابر تغییرات جزئی ورودی و حفظ دقت بالا هم‌زمان با افزایش قابلیت تفسیر، اشاره کرد. همچنین، نیاز به استانداردسازی معیارهایی برای ارزیابی کمی «کیفیت» یک تفسیر در زمینه بالینی یک مسئله حل‌نشده باقی مانده است.

با نگاه به آینده، جهت‌گیری‌های تحقیقاتی بالقوه شامل توسعه روش‌های تفسیری ذاتی که از ابتدا در معماری مدل تعبیه شده‌اند و روش‌های تفسیری تعاملی است که به پزشک اجازه می‌دهد تا با فرضیه‌سازی و آزمایش، درک خود را از تصمیم مدل عمیق‌تر کند. امید است که این مقاله به خوانندگان کمک کند تا تصویر واضح‌تری از وضعیت فعلی قابلیت تفسیر در حوزه پزشکی به‌دست آورند و با شناسایی شکاف‌های موجود، بتوانند تحقیقات آتی را به‌سمتی هدایت کنند که کاربردهای عملیاتی این فناوری‌های قدرتمند را در عمل بالینی در اسرع وقت ارتقا دهند.

منابع

1. Messalas A, Kanellopoulos Y, Makris C. Model-agnostic interpretability with Shapley values. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA); 2019. p. 1-7.
2. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci. 2019;116(44):22071-80.
3. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, Tengg-Kobligk HV, et al. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiol Artif Intell. 2020;2(3):190043.
4. Wang L, Yoon KJ. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Trans Pattern Anal Mach Intell. 2021;21-26.
5. Pintelas E, Livieris IE, Pintelas P. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms. 2020;13(1):17.
6. Serrano S, Smith NA. Is attention interpretable? arXiv preprint arXiv:1906.03731; 2019.
7. Margot V, Luta G. A new method to compare the interpretability of rule-based algorithms. AI. 2021;2(4):621-35.
8. Kumar D, Taylor GW, Wong A. Discovery radiomics with CLEAR-DR: Interpretable computer-aided diagnosis of diabetic retinopathy. IEEE Access. 2019;7:25891-6.
9. Li L, Xu M, Wang X, Jiang L, Liu H. Attention-based glaucoma detection: A large-scale database and CNN model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 10571-80.
10. Jones OT, Ranmuthu CK, Hall PN, Funston G, Walter FM. Recognising skin cancer in primary care. Adv Ther. 2020;37(1):603-16.
11. Thomas SM, Lefevre JG, Baxter G, Hamilton NA. Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. Med Image Anal. 2021;68:101915.
12. Liu F, Wu X, Ge S, Fan W, Zou Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 13753-62.

استناد به این مقاله: فیروزی، مریم، و بختیاری بهادر، نادیا. (۱۴۰۵). بررسی قابلیت تفسیر یادگیری عمیق در تشخیص پزشکی. فصلنامه پیشرفت‌های مهندسی در حوزه‌ی پزشکی و مواد، ۲(۲)، ۶۷-۷۳.



Journal of Recent Advancements in Material Science and Biomedical Engineering is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.