

اصلاح خودکار رانش عملکرد تحت تغییر در فرآیند اخذ داده در طبقه بندی تصاویر پزشکی

دریافت: ۲۴ فوریه ۲۰۲۳

پذیرفته شده: ۱۰ اکتبر ۲۰۲۳

Published online: 19 October 2023

بررسی به روزرسانی ها

ملانی روشپتز^{۱,۲} ✉، گالوین خارا، جو پرزلی، نیشا شارما^۳،
جانانا جی. جیمز^۴، اوا آمبروزای^۵، آدام هروکس^۱، پیتر کچکمتی-توبیاس ریکن^{۱,۲} ✉
او بن گلاکر^{۱,۲} ✉

مدل های پیش بینی مبتنی بر تصویر برای تشخیص بیماری، به تغییرات در اکتساب داده ها مانند تعویض سخت افزار اسکنر یا به روزرسانی نرم افزار پردازش تصویر حساس هستند. تفاوت های حاصل در ویژگی های تصویر ممکن است منجر به انحراف در معیارهای عملکرد بالینی مرتبط شود که می تواند در تصمیم گیری بالینی، حتی برای مدل هایی که از نظر مساحت زیرمنحنی مشخصه عملکرد گیرنده تعمیم می یابند، آسیب ایجاد کند. ما هم ترازای پیش بینی بدون نظارت را پیشنهاد می کنیم، یک روش کالیبراسیون مجدد خودکار عمومی که نیازی به حاشیه نویسی حقایق زمینی ندارد و فقط به مقادیر محدودی از تصاویر نمونه بدون برچسب از توزیع داده های تغییر یافته نیاز دارد. ما اثربخشی روش پیشنهادی را برای تشخیص و اصلاح انحراف عملکرد در غربالگری سرطان پستان مبتنی بر ماموگرافی و روی داده های هیستوپاتولوژی در دسترس عموم نشان می دهیم. ما نشان می دهیم که روش پیشنهادی می تواند عملکرد مورد انتظار را از نظر حساسیت/اختصاصیت تحت سناریوهای مختلف واقع بینانه تغییر اکتساب تصویر حفظ کند، بنابراین یک محافظ مهم برای استقرار بالینی ارائه می دهد.

نقاطی اساس یک بده پستان معنادار بالینی خاص SEN/SPC. در اینجا، یک نقطه عملیاتی با یک آستانه کالیبره شده در امتیاز پیش بینی پیوسته هوش مصنوعی مرتبط است. آستانه ها اغلب از پیش تعیین شده اند، در حالت ایده آل به عنوان بخشی از مطالعات اعتبارسنجی در مقیاس بزرگ. آستانه ها همچنین ممکن است به عنوان بخشی از بهینه سازی محلی با استفاده از داده های تاریخی از یک سایت استقرار جدید تنظیم شوند. در اینجا، فرض بر این است که چنین داده های اعتبارسنجی، نماینده داده های جدید از بیماران نادریده هستند و همچنان باقی می ماند و نقطه عملیاتی مرتبط در محدوده مقادیر تأیید شده قرار دارد.

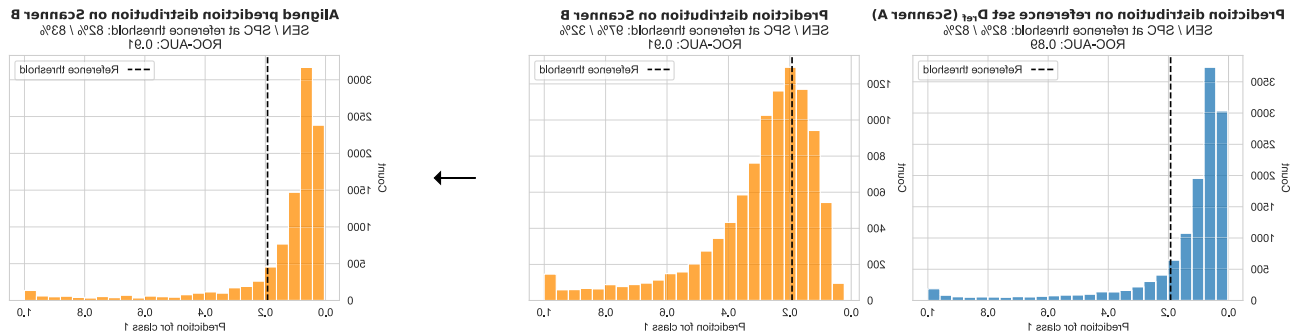
تغییراتی که پس از استقرار رخ می دهند، چالشی اساسی برای هوش مصنوعی تصویربرداری پزشکی ایجاد می کنند. سیستم های فعلی مستعد تفاوت در ویژگی های تصویر هستند. تغییرات در تصاویر ورودی، به عنوان مثال، به دلیل تعویض سخت افزار اسکنر یا به روزرسانی نرم افزار بازسازی و پردازش تصویر، می تواند باعث ایجاد تفاوت در پیش بینی های خروجی هوش مصنوعی شود (شکل ۱ را ببینید). این تغییرات ممکن است در هر زمانی پس از استقرار رخ دهند و اغلب خارج از کنترل هستند.

هوش مصنوعی (AI) نویدبخش تجزیه و تحلیل عینی تر، دقیق تر و مقرون به صرفه تر داده های تصویربرداری است و می تواند اساساً گردش کار بالینی را در تشخیص های مبتنی بر تصویر و غربالگری جمعیت متحول کند. استفاده از هوش مصنوعی می تواند به کاهش فشار بر خدمات بهداشتی، به عنوان مثال، از طریق اولویت بندی خودکار موارد بحرانی، کمک کند. با این حال، نظرات دوم در غربالگری تشخیصی^{۱,۲}.

موارد استفاده مختلف، بسته به نقش سیستم هوش مصنوعی در گردش کار بالینی، الزامات متفاوتی در مورد عملکرد هوش مصنوعی خواهند داشت. انتظار می رود یک سیستم هوش مصنوعی که برای اولویت بندی یا اولویت بندی استفاده می شود، حساسیت (SEN) تقریباً کاملی داشته باشد، در حالی که اختصاصیت (SPC) پایین ممکن است قابل قبول باشد. از سوی دیگر، انتظار می رود هوش مصنوعی به عنوان یک خواننده دوم در غربالگری سرطان سینه با خوانش دوگانه، در سطوح SEN/SPC مشابه یک خواننده انسانی عمل کند. بسیاری از سیستم های هوش مصنوعی، مدل های پیش بینی همه کاره ای هستند که می توانند در عملیات مختلف مورد استفاده قرار گیرند.

اشتراک فناوری های پزشکی خیرین، لندن، انگلستان؛ کالج امپریال لندن، گروه کامپیوتر، لندن، انگلستان؛ بیمارستان آموزشی لیدز، سازمان خدمات درمانی ملی (NHS Trust)، بیمارستان رادیولوژی، لیدز، انگلستان؛ بیمارستان های دانشگاه ناتینگهام، NHS Trust، بیمارستان شهر ناتینگهام، موسسه پستان ناتینگهام، ناتینگهام، انگلستان.

مجاورستان. ایمیل MaMMA Egészségügyi Zrt.، b.glocker@imperial.ac.uk؛ mb121@imperial.ac.uk



دریک اسکتر کاملاً نامرئی، می توانیم مشاهده کنیم که علیرغم تعمیم ROC-AUC، تغییر واضحی در پیش بینی هاو کالبراسیون مدل وجود دارد. ج. تأثیر اعمال روش UPA پیشنهادی بر حسب توزیع پیش بینی هاو موازنه SEN/SPC. داده های منبع به صورت یک فایل داده منبع ارائه شده اند.

شکل ۱ | نمایش تغییر در پیش بینی مدل هنگام اعمال یک مدل تشخیص بدخیمی ماموگرافی به یک اسکندرجید. از چپ به راست: الف. توزیع پیش بینی های مشاهده شده در مجموعه مرجع که از یک اسکتر نماینده اکثریت مجموعه توسعه حاصل شده و برای تعیین آستانه طبقه بندی استفاده شده است، ب. توزیع پیش بینی های مشاهده شده هنگام اعمال این مدل

ایمنی و قابلیت اطمینان هوش مصنوعی تصویربرداری پزشکی، شامل توسعه دهندگان هوش مصنوعی، متخصصان مراقبت های بهداشتی، بیماران، تنظیم کنندگان مقررات و سیاست گذاران.

نتایج

مجموعه داده ها

ماز چهار مجموعه داده ماموگرافی پستان با برچسب های بدخیمی تأیید شده توسط بیوپسی جانبی استفاده می کنیم. برای هر شرکت کننده، چهار تصویر ثبت می شود (دو نمایری هر پستان). مجموعه داده آموزشی دی. قطاراز OPTIMAM است^{۲۱} که یک مجموعه داده غنی از بریتانیا است، در حالی که تمام مجموعه های ارزیابی دیگر، مجموعه داده های غربالگری از سه سایت در بریتانیا و چهار سایت در مجارستان هستند که نماینده جمعیت های واقعی در برنامه های ملی غربالگری سرطان پستان مربوطه می باشند. برای همه مجموعه های ارزیابی، اگر بیش از یک مطالعه برای هر شرکت کننده وجود داشت، ما فقط یکی را (به صورت تصادفی نمونه گیری شده) نگه داشتیم. ما علاوه بر این، اثربخشی UPA را بر روی WILDS Camelyon17، یک مجموعه داده هیستوپاتولوژی در دسترس عموم، ارزیابی می کنیم.^{۲۲} این مجموعه داده برای مطالعه تغییرات مجموعه داده ناشی از تغییرات در پروتکل های رنگ آمیزی از یک بیمارستان به بیمارستان دیگر طراحی شده است. این مجموعه داده شامل تکه هایی از تصاویر کل اسلاید با برچسب های تکه ای است که نشان می دهد بافت سرطانی است یا خیر. مجموعه داده آموزشی از داده های سه بیمارستان تشکیل شده است و مجموعه مرجع از همان مکان ها نمونه برداری شده است. ارزیابی روی دو مجموعه داده دیده نشده از دو مکان جدید بیمارستان انجام می شود. خلاصه ای از ویژگی های مجموعه داده را می توانید در جداول بیابید. ^{۲۳}

تنظیمات آزمایشی

ما چندین سناریوی واقع بینانه از تغییر اکتساب را برای ارزیابی اثربخشی UPA در زمینه تشخیص سرطان پستان در ماموگرافی دیجیتال و طبقه بندی یافت در هیستوپاتولوژی طراحی کردیم. UPA روشی ساده است که پیش بینی های مدل را از یک دامنه اکتساب نامرئی (مثلاً داده های به دست آمده با یک اسکتر جدید) با توزیع پیش بینی مرجع ثبت شده در یک دامنه شناخته شده (مثلاً داده های مجموعه اعتبارسنجی که در طول توسعه روش استفاده شده است) هم تراز می کند. هم تراز کردن پیش بینی با استفاده از تجمع خطی تکه ای حاصل می شود.

تطبیق توزیع نسبی. برای دو کاربرد

ماموگرافی و هیستوپاتولوژی، ما مدل های استاندارد شبکه عصبی کانولوشن عمیق را برای طبقه بندی تصویر آموزش می دهیم و بر روی داده های اعتبارسنجی کالبره می کنیم تا یک نقطه عملیاتی ایجاد کنیم که در آن حساسیت برابر با ویژگی است. جزئیات بیشتر در مورد UPA، مدل های هوش مصنوعی و تنظیمات مربوط به هر سناریو را می توانید در بخش روش ها بیابید.

سناریوی اول: استقرار در یک سایت جدید

در این مجموعه اول از آزمایش ها، ما سناریوی را شبیه سازی می کنیم که در آن می خواهیم مدلی را که بر اساس داده های اعتبارسنجی از یک سایت مرجع بهینه شده است، به سایت های جدید دیده نشده منتقل کنیم. در اینجا، ممکن است به دلیل استفاده از اسکترهای مختلف در سایت های جدید (ماموگرافی دیجیتال) یا به طور گسترده تر به دلیل تفاوت در پروتکل تصویربرداری، تغییر در فرآیند جمع آوری داده ها رخ دهد.

از سازنده هوش مصنوعی. اگرچه تغییرات در ویژگی های تصویر ممکن است لزوماً بر قدرت پیش بینی کلی (یعنی توانایی تمایز بین موارد مثبت و منفی) تأثیر نگذارد، اما هرگونه تغییر یا انحراف در توزیع خروجی ممکن است آستانه های کالبره شده را نامعتبر کند. این امر مستقیماً بر بده پستان SEN/SPC و در نتیجه، نقطه عملیاتی مورد نظر سیستم هوش مصنوعی تأثیر می گذارد. در صورت عدم تشخیص، چنین انحراف عملکرد بالینی می تواند منجر به تشخیص کمتر یا بیشتر از حد معمول شود و عواقب شدیدی بر ایمنی بیمار داشته باشد.

نیازمبری به یک روش مؤثر برای نظارت بر هوش مصنوعی در استقرارهای دنیای واقعی وجود دارد که امکان اصلاح خودکار انحراف عملکرد ناشی از تغییرات در اخذ تصویر را فراهم کند. موضوع نظارت بر عملکرد، موضوعی مبرم برای استقرار عملی مدل های یادگیری ماشینی، به ویژه در کاربردهای حیاتی مانند تشخیص بیماری است.^{۲۴،۲۵} در گذشته، روش هایی برای تشخیص چنین انحرافات در عملکرد و به روزرسانی مدل ها ارائه شده است.^{۱۸-۲۳} اما همه اینها نیاز به حاشیه نویسی انسانی از نمونه های جدید دارند. با این حال، به دست آوردن حاشیه نویسی ها در زمان واقعی و پیوند دادن آنها با تصاویر به دست آمده معمولاً در عمل بالینی امکان پذیر نیست، به ویژه در کاربردهای مانند غربالگری سرطان پستان که در آن تشخیص تأیید شده (مثلاً بدخیمی اثبات شده با بیوپسی) در زمان تجزیه و تحلیل تصویر در دسترس نیست. برخی از روش ها برای تشخیص خودکار تغییرات بر اساس آزمون های آماری که پیش بینی های مدل را مقایسه می کنند، پیشنهاد شده اند.^{۱۹} اما همچنان نیاز برآورده نشده ای به روش هایی وجود دارد که امکان انجام خودکار را فراهم می کنند در دست رانش عملکرد در غیاب هرگونه اطلاعات تشخیصی (مانند برچسب های بیماری). این دقیقاً تمرکز این کار است.

در این کار، ما یک رویکرد ساده، عمومی و مؤثر برای هم ترازای پیش بینی بدون نظارت (UPA) را پیشنهاد و ارزیابی می کنیم که قادر به تشخیص و اصلاح انحراف عملکرد بالینی مرتبط هوش مصنوعی ناشی از تغییر در اکتساب است. آزمایش های ما اثربخشی UPA را در چندین سناریوی دنیای واقعی در زمینه غربالگری پستان و هیستوپاتولوژی نشان می دهد. در کاربرد غربالگری پستان، ابتدا نشان می دهیم که UPA قادر است خروجی های یک مدل بهینه سازی شده روی یک فروشنده سخت افزار را برای بازایی عملکرد SEN/SPC مورد نظر روی فروشندگان مختلف در سه مجموعه داده غربالگری پستان در مقیاس بزرگ بریتانیا تطبیق دهد. ما ماهیت عمومی UPA را با استفاده از WILDS Camelyon17 که در دسترس عموم است، تأیید می کنیم.^{۲۲} مجموعه داده هایی که نشان می دهد یک مدل طبقه بندی بهینه شده بر روی یک پروتکل رنگ آمیزی خاص می تواند به طور خودکار با یک پروتکل رنگ آمیزی جدید سازگار شود. نکته مهم این است که UPA به گونه ای طراحی شده است که امکان کالبراسیون مجدد مداوم را فراهم می کند، زیرا به طور خودکار تغییرات مشاهده شده در توزیع پیش بینی را در طول زمان تشخیص داده و تنظیم می کند. ما این موضوع را با شبیه سازی سناریوهای تغییر چندگانه جمع آوری، از جمله معرفی اسکترهای جدید و به روزرسانی های نرم افزار پردازش تصویر، نشان می دهیم. ما در مورد الزامات داده ها، فرضیات و محدودیت های روش پیشنهادی بحث می کنیم. ما معتقدیم که این کار برای هر کسی که نگران ...

جدول 1 | آمار مجموعه داده های غربالگری سرطان پستان

| مجموعه داده ها | دستگاه | کشور | نوع | نموارد | ن تصاویر | درصد تصاویر مثبت |
|----------------|-----------------------------|----------|---------------------|----------|----------|------------------|
| اسکندر A | هولووژیک | بریتانیا | مجموعه مرجع دی مرجع | ۳۲۲۱ | ۱۲,۸۸۴ | ۲.۵٪ |
| اسکندر ب | چوتو IMS | مجارستان | نادیده | ۴۱۵۲ | ۱۶,۶۰۸ | ۱.۸٪ |
| اسکندر C | زیمسن | بریتانیا | نادیده | ۳۹۰۴ عدد | ۱۵,۶۱۶ | ۱.۹٪ |
| اسکندر D | بهداشتو درمان جنرال الکتریک | بریتانیا | نادیده | ۳۱۱۵ عدد | ۱۶,۴۶۰ | ۱.۸٪ |

داده های منبع به صورت یک فایل داده منبع ارائه می شوند.

جدول ۲ | آمار مجموعه داده ها برای مجموعه داده های هیستوپاتولوژی WILDS Camelyon17

| مجموعه داده ها | نوع | ن تصاویر | درصد مثبت |
|-------------------|---------------------|----------|-----------|
| سایت های S1 تا S3 | مجموعه مرجع دی مرجع | ۳۳,۵۶۰ | ۵۰٪ |
| سایت S4 | نادیده | ۳۴,۹۰۴ | ۵۰٪ |
| سایت S5 | نادیده | ۸۵,۰۵۴ | ۵۰٪ |

داده های منبع به صورت یک فایل داده منبع ارائه می شوند.

اسکندر B به تدریج معرفی می شود، به این معنی که تعداد کل اسکن ها با گذشت زمان افزایش می یابد تا زمانی که نسبت مساوی از هر دو اسکنر وجود داشته باشد، که در نتیجه تعداد کل موارد دو برابر می شود. آستانه های طبقه بندی برای اسکندر A بهینه شدند. نتایج رامی توانید در شکل 1 بیابید.^۴(ردیف پایین)، دوباره UPA قادر است در طول زمان تعادل SEN/SPC را حفظ کند.

سناریو ۴: به روزرسانی پردازش تصویر
در این سناریو، ما تأثیر یک به روزرسانی OEM اعمال شده بر الگوریتم پردازش تصویر برای یک اسکندر موجود را شبیه سازی می کنیم. فرض می شود که محل استقرار، یک اسکندر واحد را اداره می کند و در زمان استقرار هوش مصنوعی، آستانه های طبقه بندی برای نسخه اولیه نرم افزار پردازش تصویر بهینه شده اند. در نقطه زمانی بعدی T1، فرض می کنیم که سازنده اسکندر، به روزرسانی نرم افزاری را بر الگوریتم پردازش تصویر خود اعمال کرده است که منجر به افزایش وضوح تصویر می شود. چنین به روزرسانی هایی تا حد زیادی خارج از کنترل توسعه دهندگان مدل هوش مصنوعی هستند. نتایج این شبیه سازی را می توانید در شکل 1 بیابید.^۵افزایش وضوح تصویر باعث تغییر ناگهانی و قابل توجه در تعادل SEN/SPC می شود. اعمال UPA منجر به سازگاری سریع با ویژگی های جدید تصویر می شود و مدل به سرعت در بده بستان مورد نظر SEN/SPC عمل می کند.

تأثیر اندازه مجموعه داده ها
برای کسب بینش بیشتر در مورد میزان داده های مورد نیاز برای دستیابی به عملکرد خوب در مرحله هم تراز UPA، حساسیت روش هم تراز را نسبت به اندازه مجموعه مرجع (از دامنه منبع) و همچنین اندازه مجموعه هم تراز (از دامنه جدید) تجزیه و تحلیل می کنیم. نتایج مربوط به اسکندر ماموگرافی B را می توانید در شکل 1 بیابید.^۶، به همراه نتایج اضافی برای سایر مجموعه داده ها که در یادداشت تکمیلی ۲ ارائه شده است. این نتایج نشان می دهد که یک مجموعه هم تراز به کوچکی ۲۵۰ مورد (۱۰۰۰ تصویر) برای بازیابی تغییر در SEN/SPC در مجموعه داده های ماموگرافی کافی است. برای مجموعه مرجع، با ۵۰۰ مورد در مجموعه هم تراز، دریافتیم که ۱۰۰۰ مورد برای نتایج پایدار کافی است. در عمل، می توان از همان مجموعه مرجعی که برای اعتبارسنجی مدل و انتخاب آستانه طبقه بندی استفاده شده است، استفاده کرد.

بحث

مایک روش مؤثر برای اصلاح انحراف عملکرد بالینی مرتبط با تغییرات در خط لوله اخذ تصویر (مثلاً تعویض اسکنرها، به روزرسانی نرم افزار پردازش تصویر، استفاده از پروتکل های رنگ آمیزی مختلف) پیشنهاد کرده ایم. این موضوع در سناریوهای مختلف، از تغییرات یکباره هنگام استقرار در یک سایت جدید گرفته تا انحرافات تدریجی و پیشرونده در ویژگی های اخذ تصویر در طول زمان، نشان داده شده است. آزمایش های ما نشان می دهد که یک ROC-AUC قابل تعمیم لزوماً به این معنی نیست که آستانه های طبقه بندی انتخاب شده (یعنی نقاط عملیاتی بالینی) به داده های دیده نشده تعمیم داده می شوند. همانطور که در شکل ها نشان داده شده است،^۳ و^۴ مدل های مورد مطالعه از نظر ROC-AUC در تمام مجموعه داده ها تعمیم داده شدند، در حالی که تعادل SEN/SPC هنگام اعمال یک آستانه طبقه بندی از پیش تعریف شده، به طور قابل توجهی در بین سایت ها متفاوت بود. نتایج ما نشان می دهد که روش هم تراز پیش بینی بدون نظارت پیشنهادی می تواند با موفقیت تغییرات آستانه را در مجموعه داده های دیده نشده در دو وظیفه طبقه بندی تصویر بسیار متفاوت کاهش دهد. UPA توانست بدون استفاده از هیچ برچسبی از دامنه هدف دیده نشده، موازنه SEN/SPC مورد نظر را بازیابی کند. نکته مهم این است که این روش ساده اما مؤثر می تواند برای به روزرسانی های مداوم مدل استفاده شود و به ... کمک کند.

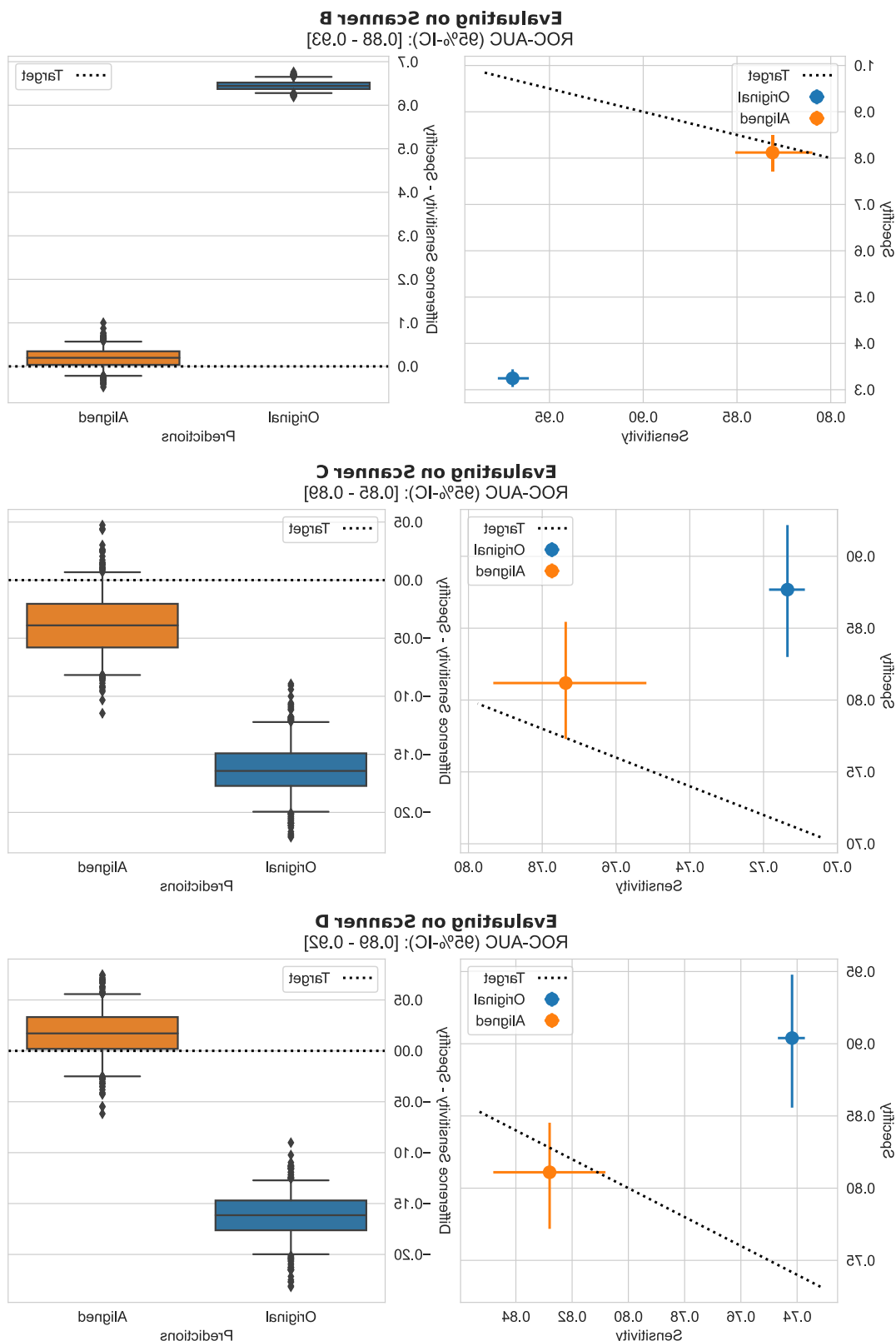
از جمله استفاده از پروتکل های رنگ آمیزی مختلف (هیستوپاتولوژی). برای وظیفه تشخیص سرطان پستان، از یک مجموعه داده آزمایشی از یک اسکندر که در طول توسعه مدل دیده نشده است، استفاده می کنیم. برای وظیفه طبقه بندی بافت هیستوپاتولوژی، داده های ارزیابی از بیمارستان دیگری با استفاده از پروتکل های رنگ آمیزی مختلف می آیند. در هر دو مورد، فرض می کنیم که مدل های طبقه بندی مربوطه بر روی داده های اعتبارسنجی مرجع (قبل از استقرار در مکان های جدید) از قبل کالیبره شده اند. آستانه طبقه بندی در نقطه عملیاتی بالینی انتخاب شده است که در آن SEN برابر با SPC در این داده های مرجع است. ما ابتدا عملکرد مدل را از نظر SEN/SPC بر روی داده های جدید قبل از استفاده از UPA ارزیابی می کنیم. سپس از تکنیک هم تراز پیشنهادی برای ارزیابی تأثیر آن بر تغییر آستانه استفاده می کنیم. معیارها بر اساس مساحت تصویر در زیر منحنی مشخصه عملکرد گیرنده (ROC-AUC) و SEN/SPC، در شکل 1 گزارش شده اند.^۲ برای تشخیص سرطان سینه و در شکل ۳، برای Camelyon17. علاوه بر این، در جداول ۳ و ۴، ما همچنین شاخص یودن را گزارش می دهیم.^۳ یک معیار ترکیبی که حساسیت و ویژگی را به طور همزمان، قبل و بعد از اعمال UPA در نظر می گیرد. ما دریافتیم که بدون UPA، تغییر قابل توجهی در تعادل SEN/SPC، در تمام وظایف و مجموعه داده های ارزیابی وجود دارد. بدون UPA، مدل دیگر در نقطه عملیاتی مورد نظر عمل نمی کند. برعکس، نتایج نشان می دهد که پس از اعمال روش هم تراز، ما تعادل SEN/SPC تا حد زیادی برای داده های دیده نشده بازیابی می شود. علاوه بر این، در یادداشت تکمیلی 1، نشان می دهیم که UPA مستقل از انتخاب خاص نقطه عملیاتی (مثلاً، ویژگی هدف از پیش تعریف شده) نیز عمل می کند.

سناریو ۲: انتقال به یک اسکندر جدید

در اینجا، ما سناریویی را شبیه سازی می کنیم که نیاز به به روزرسانی های مداوم مدل دارد، که در آن یک اسکندر جدید نصب می شود و یک اسکندر قدیمی به تدریج از رده خارج می شود. یک دوره گذار برنامه ریزی شده وجود دارد که در آن اسکنرهای قدیمی و جدید به صورت موازی با تغییر تدریجی به اسکندر جدید استفاده می شوند. در ابتدا، فقط داده ها از اسکندر A و در انتها فقط داده ها از اسکندر B وارد می شوند. آستانه های طبقه بندی مدل هوش مصنوعی برای اسکندر A بهینه شده اند. نتایج را می توانید در شکل 1 مشاهده کنید.^۴ (بالا)، که در آن سمت چپ تعداد موارد پردازش شده توسط هر اسکندر را در طول زمان برای هر سناریو نشان می دهد و سمت راست نتایج مربوطه را از نظر حساسیت و ویژگی مدل در طول زمان، با و بدون اعمال UPA به طور کلی و در سطح اسکندر نشان می دهد. از شکل، مشاهده می کنیم که معرفی اسکندر جدید منجر به تغییرات قابل توجهی در SEN/SPC در صورت عدم استفاده از UPA می شود، در حالی که اعمال UPA تعادل مطلوب SEN/SPC را در طول زمان حفظ می کند.

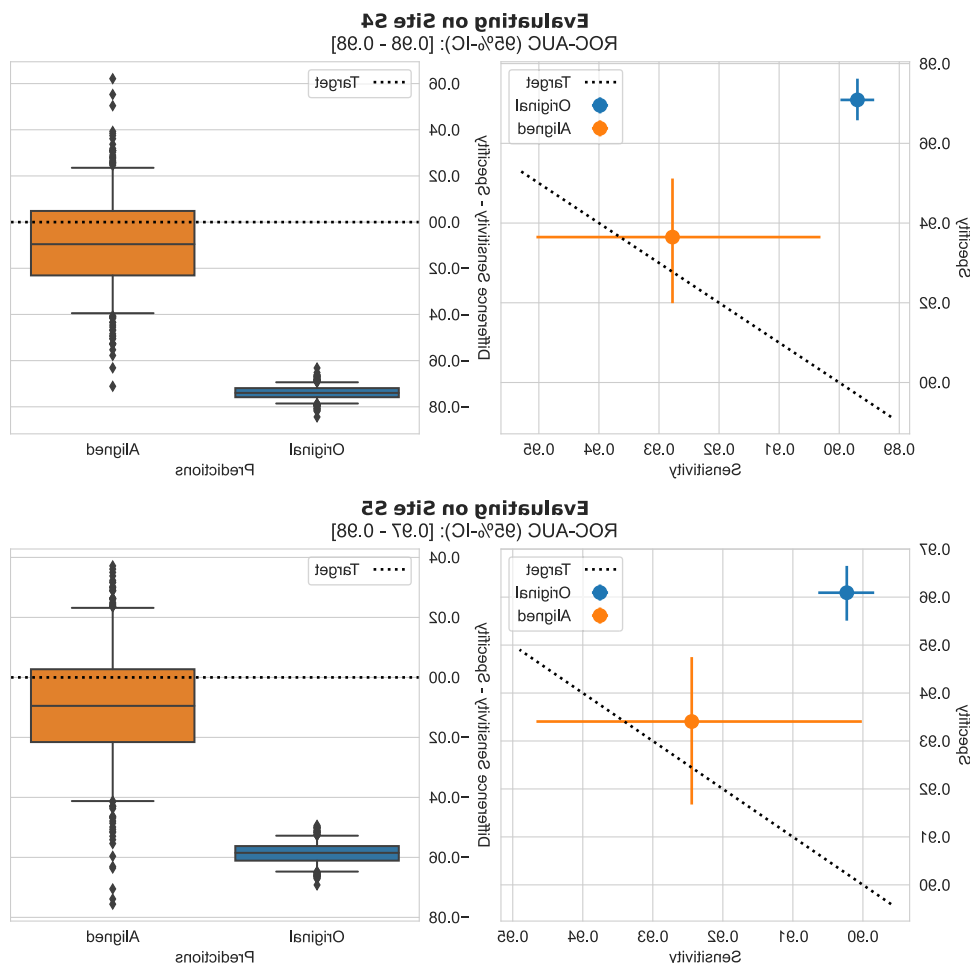
سناریو ۳: اضافه کردن یک اسکندر جدید

مشابه سناریو ۲ بالا، در اینجا یک اسکندر جدید B نصب شده است، با این حال، اسکندر قدیمی A همچنان به صورت موازی مورد استفاده قرار خواهد گرفت. اسکندر جدید



هرمعیار. ستون سمت راست: تفاوت بین حساسیت و اختصاصیت قبل و بعد از هم ترازی. نمودارهای جعبه ای از ۵۰۰ نمونه برداری مکرر از مجموعه های ارزیابی و هم ترازی ساخته شده اند؛ هر جعبه، صدک های ۲۵٪، ۵۰٪ و ۷۵٪ توزیع بوت استرپ را نشان می دهد؛ ویسکرها نشان دهنده صدک های ۵٪ و ۹۵٪ هستند و هر نقطه ای خارج از این محدوده به عنوان یک داده پرت نمایش داده می شود. UPA در بازیابی تعادل حساسیت/اختصاصیت مورد نظر در تمام مجموعه داده های خارج از توزیع مؤثر است. داده های منبع به عنوان یک فایل داده منبع ارائه می شوند.

شکل ۲ | سناریو ۱: استقرار در یک سایت خبری - وظیفه غربالگری سینه. ستون سمت چپ: ویژگی در تابع حساسیت قبل و بعد از هم ترازی پیش بینی. برای این تحلیل، ما یک مجموعه ارزیابی (از ۲۵۰۰ مورد) یک مجموعه هم ترازی مجزا (از ۱۰۰۰ مورد) را از تمام موارد موجود نمونه برداری می کنیم، این نمونه برداری ۵۰۰ بار با جایگزینی تکرار می شود. حساسیت، ویژگی، ROC-AUC روی این ۵۰۰ نمونه بوت استرپ اندازه گیری می شوند و نتایج بر اساس میانگین نتایج روی نمونه های بوت استرپ گزارش می شوند و میله های خطا، فاصله اطمینان ۹۵٪ بوت استرپ را برای



ستون: تفاوت بین حساسیت و اختصاصیت قبل و بعد از هم ترازی. نمودارهای جعبه ای از ۵۰۰ نمونه بوت استرپ از مجموعه های ارزیابی و هم ترازی ساخته شده اند؛ هر جعبه، صدک های ۲۵٪، ۵۰٪ و ۷۵٪ توزیع بوت استرپ را نشان می دهد؛ ویسکرها نشان دهنده صدک های ۵٪ و ۹۵٪ هستند و هر نقطه ای خارج از این محدوده به عنوان یک داده پرت نمایش داده می شود. UPA در بازیابی تعادل حساسیت/اختصاصیت مورد نظر در تمام مجموعه داده های خارج از توزیع مؤثر است. داده های منبع به عنوان یک فایل داده منبع ارائه می شوند.

شکل ۳ | سناریو ۱: استقرار در یک سایت جدید - وظیفه هیستوپاتولوژی WILDS Camelyon17. سمت چپ: ویژگی در تابع حساسیت قبل و بعد از هم ترازی پیش بینی. برای این تحلیل، ما یک مجموعه ارزیابی (شامل ۱۵۰۰۰ وصله) و یک مجموعه هم ترازی مجزا (شامل ۵۰۰۰ وصله) را از تمام وصله های موجود نمونه برداری می کنیم. این نمونه برداری ۵۰۰ بار با جایگزینی تکرار می شود. حساسیت، ویژگی، ROC-AUC در ۵۰۰ نمونه برداری مکرر از مجموعه های ارزیابی و هم ترازی اندازه گیری می شوند و نتایج بر اساس میانگین نتایج روی نمونه های بوت استرپ گزارش می شوند و میله های خطا، فاصله اطمینان ۹۵٪ بوت استرپ را برای هر معیار نشان می دهند. سمت راست

برای مثال، در بسیاری از مراکز غربالگری، تشخیص مجموعه ای از ۵۰۰ مورد جدید اغلب تنها چند روز طول می کشد.

در موارد کاربردی عملی که چندین اسکنر یا پروتکل جمع آوری تصویر در یک مکان استفاده می شوند (مانند سناریوهای ۲ و ۳ در بالا)، شایان ذکر است که UPA به صورت اسکنر/پروتکل محور اعمال می شود. یعنی، ما یک الگوریتم هم تراز را برای هر اسکنر به طور جداگانه بر اساس جدیدترین مجموعه اسکن های بیمار که با این اسکنر خاص گرفته شده است، تنظیم می کنیم. این امر به آنها اجازه می دهد تا کاملاً مستقل عمل کنند. این تطبیق دقیق دستگاه محور به ویژه در مواردی که نسبت اسکن های پردازش شده توسط یک نوع اسکنر خاص به صورت روزانه تغییر می کند یا یکی از آنها به طور ناگهانی برای تعمیر و نگهداری کار نمی کند، مفید است. به طور مشابه، اگر یک اسکنر به روزرسانی نرم افزاری داشته باشد که بر پیش بینی های مدل تأثیر بگذارد، فقط این اسکنر باید بدون تأثیر بر سایر اسکنرها، به طور خودکار دوباره کالیبره شود.

در مورد رانش عملکرد پیش رونده، تحلیل حساسیت ما نشان می دهد که UPA فقط به مقادیر محدودی از داده ها از توزیع جایجا شده نیاز دارد و می تواند از داده های اعتبارسنجی موجود به عنوان مجموعه مرجع استفاده مجدد کند. فقط بان =۵۰۰ مورد از حوزه هدف و ن =۱۰۰۰ مورد از دامنه منبع، نتایج تقریباً بهینه و پایداری را برای کالیبراسیون مجدد خروجی های مدل مشاهده کردیم. این واقعیت که فقط به مقدار محدودی از داده ها از دامنه هدف نیاز است، برای پیاده سازی عملی UPA مهم است. UPA که به عنوان ابزاری برای نظارت مداوم بر عملکرد و کالیبراسیون مجدد استفاده می شود، می تواند در یک بازه زمانی نسبتاً کوتاه از داده های ورودی جدید اجرا شود. در غربالگری سرطان سینه،

جدول ۳ | شاخص یودن برای سناریوی ۱، استقرار در یک سایت جدید، غربالگری پستان

| مجموعه داده ها | قبل از اصلاح | پس از اصلاح |
|----------------|--------------|--------------|
| اسکنر ب | ۰.۳۹۵ (۰.۱۱) | ۰.۶۵۱ (۰.۲۱) |
| اسکنر ج | ۰.۵۹۹ (۰.۳۳) | ۰.۵۹۴ (۰.۲۰) |
| اسکنر D | ۰.۶۴۴ (۰.۳۴) | ۰.۶۴۰ (۰.۲۱) |

به صورت میانگین روی ۵۰۰ نمونه بوت استرپ گزارش شده است، و انحراف معیار داخل پراکنش قرار دارد. داده های منبع به صورت فایل داده های منبع ارائه شده اند.

اگر فرض نقض شود، اما شیوع هدف مشخص باشد، یا امکان جمع آوری برچسب ها روی مجموعه داده های دیده نشده وجود داشته باشد، UPA همچنان می تواند با اطمینان از اینکه شیوع در مجموعه مرجع انتخاب شده با شیوع هدف مطابقت دارد (مثلاً با استفاده از تکنیک های نمونه گیری مجدد) اعمال شود. با این حال، اگر هیچ اطلاعات یا اطلاعات کافی در مورد علل تغییر توزیع بین مجموعه مرجع و دامنه هدف جدید در دسترس نباشد، UPA نباید بدون بررسی بیشتر برای کالیبراسیون مجدد خودکار استفاده شود. در عمل، ما پیش بینی می کنیم که UPA در کنار نظارت جامع بر داده های ورودی، از جمله متا اطلاعات در مورد جمعیت بیمار، استفاده شود. این امر امکان تشخیص و علامت گذاری تغییرات غیرمنتظره در جمعیت بیمار را فراهم می کند. نظارت بر داده های جمعیتی پایه در حال حاضر یک روش استاندارد در برنامه های غربالگری است. این می تواند با ابزارهای نظارت خودکار، به عنوان مثال با استفاده از مدل های کمکی هوش مصنوعی، برای پیش بینی ویژگی های بیمار و داده ها از تصاویر ورودی (مثلاً سن، تراکم سینه، کیفیت تصویر غیره) و روش هایی که به طور خاص برای تشخیص انحراف توسعه یافته اند، تکمیل شود.^۹ تشخیص تغییرات، جنبه ی مهمی از نظارت مستمر بر عملکرد است. توجه داشته باشید که UPAها لزوماً نباید به صورت بلادرنگ اعمال شوند، به این معنی که برای استقرار، پیاده سازی یک تأخیر زمانی (مثلاً چند روز) بین تشخیص و اصلاح تغییر توزیع ممکن است منطقی باشد. این امر، با بررسی انسانی در حلقه را در مورد اینکه آیا منابع دیگری ممکن است در تغییر شناسایی شده نقش داشته باشند، تسهیل می کند. در اینجا، توجه داریم که اثرات ترکیبی منابع مختلف بایاس بالقوه در مجموعه داده ها می تواند تجزیه و تحلیل علت ریشه ای تغییر توزیع داده ها را پیچیده کند.^۹ با این حال، حتی در چنین شرایطی، UPA می تواند برای تشخیص تغییر توزیع در طول استقرار به کار گرفته شود، همانطور که در مورد تغییر شیوع در یادداشت تکمیلی ۴ نشان داده ایم.

اینکه شکل توزیع پیش بینی حفظ شده است، که ممکن است تفسیر مقدار مطلق پیش بینی را غیرممکن کند. این مهم است، زیرا نشان داده شده است که کالیبراسیون مدل لزوماً هنگام انتقال از اعتبارسنجی به حوزه های آزمایش دیده نشده قابل انتقال نیست.^{۲۵} در شکل ۱، دیدیم که چگونه شکل توزیع قبل و بعد از هم تراز می شود و چگونه هم تراز می شود، شکل مرجع را حفظ می کند. هنگام مقایسه خطای کالیبراسیون مورد انتظار (ECE)، ما دریافتیم که پس از هم تراز می، ECE بین مرجع ($ECE = 0.13$) و توزیع هم تراز شده ($ECE = 0.14$) حفظ می شود، در حالی که در پیش بینی های اولیه قبل از هم تراز می ($ECE = 0.29$) دو برابر بزرگتر بود. حفظ شکل توزیع پیش بینی ممکن است در کاربردهایی که از مقدار مطلق پیش بینی های مدل مانند اولویت بندی آگاه از عدم قطعیت استفاده می کنند، از اهمیت ویژه ای برخوردار باشد.^{۲۶} و وقتی در مدل های پیش بینی ریسک استفاده می شوند.^{۲۷}

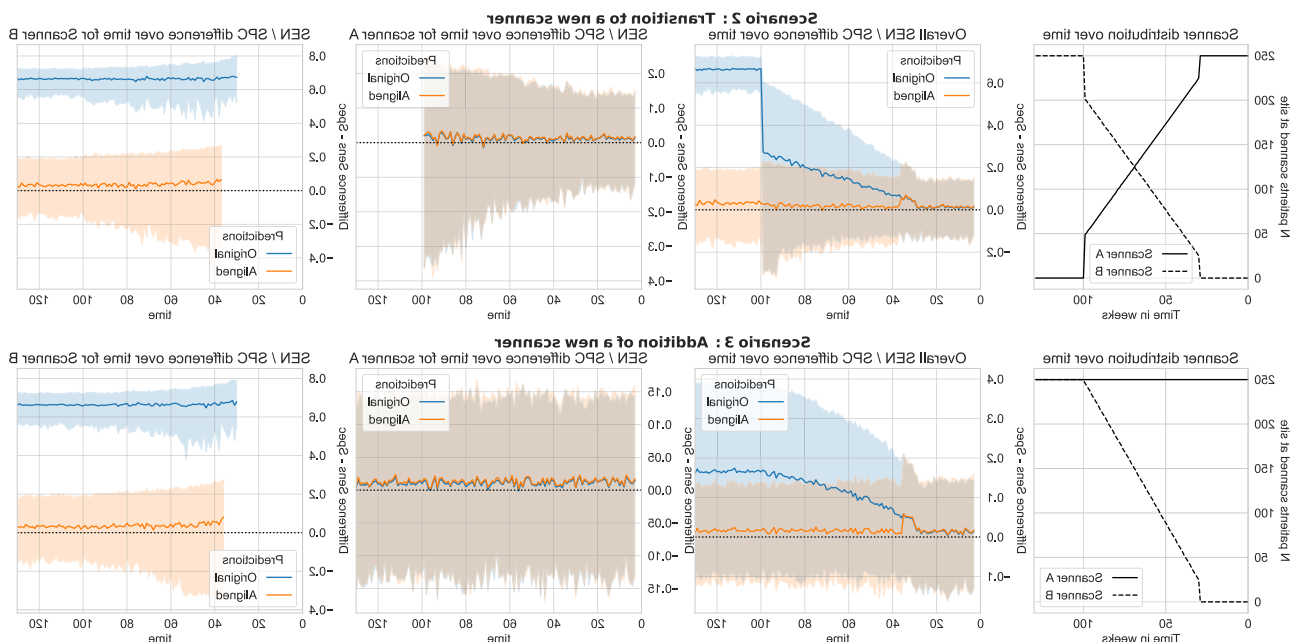
از نظر محدودیت ها، لازم به ذکر است که UPA فقط بر روی پرداختن به اثرات تغییر در اکسپانز تمرکز دارد و برای مقابله با انواع دیگر تغییرات توزیع که ممکن است باعث رانش عملکرد شوند، مانند تغییر در جمعیت، شیوع یا حاشیه نویسی، طراحی نشده است.^{۲۸} نکته مهم این است که در UPA فرض می کنیم که شیوع (تقریباً) بین دامنه های مرجع و هدف حفظ می شود. این یک فرض منطقی در بسیاری از سناریوهای بالینی عملی است، به عنوان مثال، در برنامه های غربالگری که شیوع مورد انتظار مشخص است و معمولاً در مکان هایی با جمعیت های مشابه مشابه است. توجه داشته باشید که اگر شیوع

جدول ۴ | شاخص یودن برای سناریوی ۱، استقرار در یک سایت جدید، هیستوپاتولوژی

| مجموعه داده ها | قبل از اصلاح | پس از اصلاح |
|----------------|--------------|--------------|
| سایت S4 | (۰.۰۳۰)۰.۸۶۸ | (۰.۰۰۵)۰.۸۶۴ |
| سایت S5 | (۰.۰۰۶)۰.۸۵۹ | (۰.۰۰۴)۰.۸۶۳ |

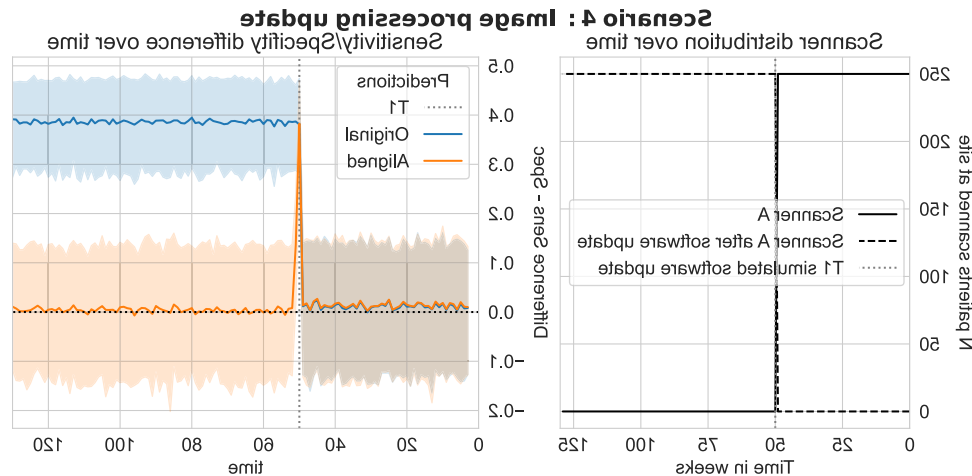
به صورت میانگین روی ۵۰۰ نمونه بوت استرپ گزارش شده است، و انحراف معیار داخل پرانتز قرار دارد. داده های منبع به صورت فایل داده های منبع ارائه شده اند.

ثانیاً، باید توجه داشته باشیم که UPA برای مقابله با تغییر آستانه طراحی شده است، نه تعمیم کلی مدل. نکته مهم این است که در اینجا فرض می کنیم که تغییر اکسپانز در درجه اول باعث تغییر در پیش بینی های مدل می شود که نقطه عملیاتی انتخاب شده را بی اعتبار می کند در حالی که ROC-AUC را در دامنه های مختلف حفظ می کند. همه مدل های بررسی شده در این مطالعه، شرایط لازم را برآورده می کنند.



شکل ۴ | سناریوهای ۲ و ۳: هم تراز می پیش بینی در طول زمان تحت تغییر مداوم در جمع آوری داده ها. هر شبیه سازی ۲۵۰ بار تکرار می شود، خطوط ممتد میانگین تفاوت بین حساسیت و اختصاصیت را در تمام نمونه های بوت استرپ نشان می دهند و نواحی سایه دار نشان دهنده فاصله اطمینان بوت استرپ صدک ۵٪ تا ۹۵٪ هستند. نمودارهای ستون سمت چپ تعداد اسکن های پردازش شده توسط اسکنر A و اسکنر B را در طول زمان برای هر سناریو نشان می دهند. نمودارهای ستون وسط، تکامل تعادل حساسیت-اختصاصیت را در طول زمان با و بدون اعمال مقایسه می کنند.

راز نظر اسکنر مقایسه می کنند. هدف، جلوگیری از انحراف بین حساسیت و اختصاصیت در حضور تغییر تدریجی در فرآیند اخذ تصویر است. روش پیشنهادی با موفقیت اختلاف صفر بین حساسیت و اختصاصیت را در طول زمان حفظ می کند، در حالی که مدل تطبیق نیافته می تواند منجر به تغییرات چشمگیر در تعادل حساسیت و اختصاصیت شود. داده های منبع به عنوان یک فایل داده منبع ارائه می شوند SEN/SPC در تمام اسکن ها. دو ستون سمت راست، تکامل تعادل UPA



شکل ۵ | سناریو ۴: هم ترازوی پیش بینی در طول زمان در مورد به روزرسانی نرم افزار هر شبیه سازی ۲۵۰ بار تکرار می شود، خطوط ممتد میانگین تفاوت بین حساسیت و ویژگی را در تمام نمونه های بوت استرپ نشان می دهند و نواحی سایه دار نشان دهنده فاصله اطمینان بوت استرپ درصدی ۵ تا ۹۵ هستند. نمودارهای ستون سمت چپ تعداد اسکن های پردازش شده توسط اسکنر A و اسکنر B را در طول زمان برای هر سناریو نشان می دهند. نمودارهای ستون سمت راست، سیر تکاملی را مقایسه می کنند.

تعداد حساسیت-ویژگی در طول زمان با و بدون اعمال UPA. هدف، جلوگیری از تغییر بین حساسیت و ویژگی در حضور تغییر تدریجی در اکتساب است. روش پیشنهادی با موفقیت اختلاف صفر SEN/SPC را در طول زمان حفظ می کند، در حالی که مدل تطبیق نیافته می تواند منجر به تغییرات چشمگیر در تعداد حساسیت-ویژگی شود. داده های منبع به عنوان یک فایل داده منبع ارائه می شوند.

اما باید با تنظیم کننده ها مذاکره و توافق شود و نظارت انسانی مناسبی را در استقرار لحاظ کند.^{۳۵} اعتبارسنجی کامل فرآیندهای مؤثر بر عملکرد یک دستگاه پزشکی، و همچنین نظارت جامع پس از ورود به بازار و پایش مداوم مدل، برای موفقیت این رویکرد در یک محیط قانونمند بسیار مهم است.

روش ها

هم ترازوی پیش بینی بدون نظارت شامل اعمال تطبیق توزیع تجمعی تکه ای خطی بین توزیع پیش بینی روی مجموعه داده دیده نشده و توزیع پیش بینی مرجع است. تطبیق خطی تکه ای که به عنوان «تطبیق هیستوگرام» نیز شناخته می شود، یک تکنیک شناخته شده در زمینه پردازش تصویر است که در آن به طور سنتی برای استانداردسازی شدت های تصویر اعمال می شود UPA^{۳۶} در تنظیمات ما، این الگوریتم تطبیق بر پیش بینی های خروجی مدل اعمال می شود. برای برازش الگوریتم تطبیق، به مجموعه ای از پیش بینی ها از دامنه مرجع («مجموعه مرجع») و همچنین مجموعه ای از پیش بینی ها از دامنه خارج از توزیع، «مجموعه هم ترازوی» نیاز داریم.

الگوریتم تطبیق شامل سه مرحله ساده زیر است:

- توزیع تجمعی مشاهده شده ی پیش بینی ها را روی مجموعه های مرجع و هم ترازوی محاسبه کنید.
- یک میان یاب خطی برازش دهید تا توزیع تجمعی مجموعه هم ترازوی را با توزیع تجمعی مجموعه مرجع مطابقت دهد.
- این میان یاب خطی را روی هر پیش بینی آزمایشی جدیدی که از دامنه ی نادیده می آید، اعمال کنید.

شبه کدالگوریتم تطبیق در یادداشت تکمیلی ۳ ارائه شده است، و یک پیاده سازی کاملاً کاربردی پایتون در کد تکمیلی ارائه شده است.

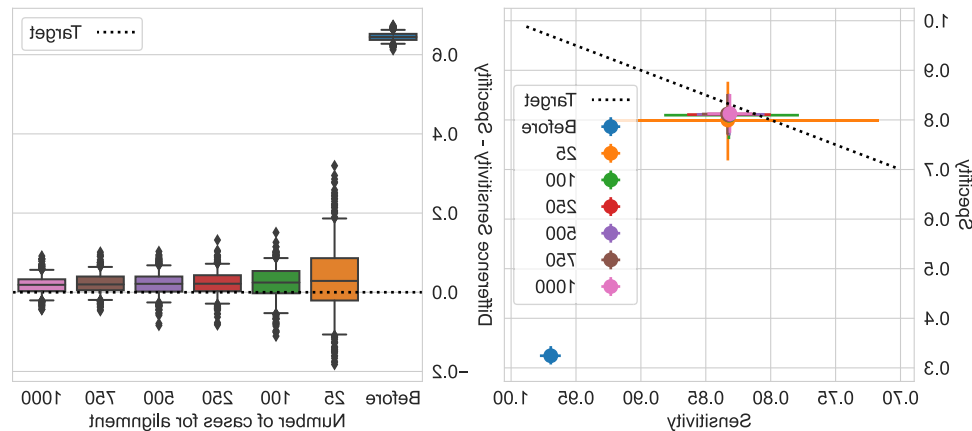
مدل هوش مصنوعی

برای اهداف این مطالعه، ما یک ResNet-50 ساده را آموزش دادیم.^{۳۷} شبکه عصبی کانولوشنی دی.قطر (به بالا مراجعه کنید) با یک زبان آنتروپی متقاطع برای هر وظیفه. برای کاربردهای سرطان سینه، قبل از پردازش تصاویر با مدل هوش مصنوعی، از ماسک کردن سینه و نرمال سازی شدت استفاده می کنیم. این مدل بر روی یک مجموعه داده غربالگری سینه غنی شده با ۱۲۲۸۵ مورد (۲۲.۷٪ تصاویر مثبت) آموزش داده می شود. برای مدل هیستوپاتولوژی، ما بر روی بخش آموزش رسمی از مجموعه داده Camelyon17 با ۳۰۲۳۶ تصویر (۵۰٪ مثبت) آموزش می دهیم. پس از

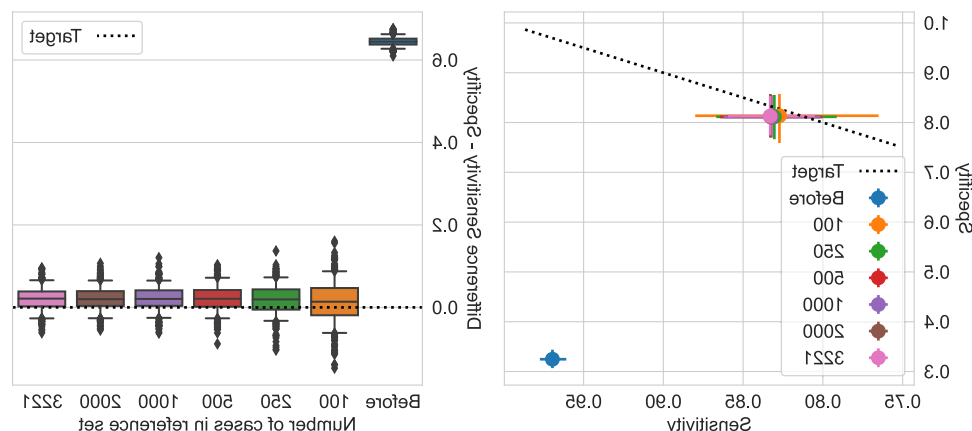
این الزام را به خوبی تعمیم می دهند و از نظر ROC-AUC به خوبی تعمیم می دهند، یعنی توانایی آنها در جداسازی کلاس ها حتی زمانی که ویژگی های تصویر ورودی تغییر می کند، تا حد زیادی حفظ می شود. همانطور که درون یابی خطی تکه ای را به خروجی های مدل اعمال می کنیم، UPA تضمین می کند که ROC-AUC قبل و بعد از ترازبندی یکسان باشد (زیرا ترتیب نسبی پیش بینی ها را تغییر نمی دهد). به این ترتیب، توانایی UPA برای بازیابی تعادل SEN/SPC مورد انتظار، تا زمانی که ROC-AUC در دامنه ها حفظ شود، به عملکرد طبقه بندی وابسته نیست. اگر ROC-AUC یک مدل از یک دامنه به دامنه دیگر به طور قابل توجهی کاهش یابد، انتظار می رود UPA عملکرد غیربهبوده داشته باشد و قادر به بازیابی عملکرد ROC-AUC در داده های دیده نشده نخواهد بود. رسیدگی به مسائل مربوط به افت عملکرد از نظر جداسازی کلی کلاس، یعنی اندازه گیری شده با افت AUC-ROC، تمرکز بخش بزرگی از کارها در تطبیق دامنه (DA) است.^{۳۸،۳۹} و تعمیم دامنه (DG).^{۳۲} این خطوط کاری مکمل یکدیگر هستند، زیرا روش های DA/DG بر ساخت مدل های قابل تعمیم بر اساس ROC-AUC تمرکز دارند و معمولاً توجه کمی به تغییر آستانه دارند. در اینجا، UPA با تمرکز بر تعمیم پذیری کالیبراسیون مدل و نقاط عملیاتی بالینی از پیش تعریف شده آن، به یک نیاز برآورده نشده می پردازد. روش پیشنهادی به هیچ وجه به عنوان جایگزینی برای تکنیک های سنتی DA/DG در نظر گرفته نشده است، بلکه باید علاوه بر آن در نظر گرفته شود. در حالی که استفاده از روش های DA/DG در طول توسعه مدل ممکن است منجر به افزایش مقاومت در برابر تغییر اکتساب شود، هیچ تضمینی برای عملکرد در حوزه های جدید دیده نشده، به ویژه از نظر معیارهای بالینی، وجود ندارد. در اینجا، UPA یک ضمانت مهم برای شناسایی رانش عملکرد در استقرارهای جدید اضافه می کند. علاوه بر این، ما معتقدیم که اعتبارسنجی محلی بر روی داده های نماینده، ممیزی های منظم هوش مصنوعی^{۳۲} و تجزیه و تحلیل دقیق موارد ناسازگار، همچنان اجزای حیاتی برای استقرار بالینی ایمن و اطمینان از ایمن بودن هوش مصنوعی در طول زمان هستند.

علاوه بر این، در مواردی که UPA را در یک مورد استفاده از تصویربرداری پزشکی پیشنهاد می کنیم، باید مقررات جهانی تجهیزات پزشکی در نظر گرفته شود. کنترل دقیق یک دستگاه پزشکی تأیید شده برای اطمینان از ایمنی و اثربخشی مداوم پس از عرضه به بازار، به ویژه هنگام در نظر گرفتن تغییر بدون نظارت پارامترها، ضروری است. به این ترتیب، ما ادعا نمی کنیم که اعمال UPA کاربران را از رویه های استاندارد حسابرسی شامل مقایسه با ارزیابی انسانی بی نیاز می کند. با این حال، ادغام آن در گردش کار می تواند امکان اصلاح خودکار را قبل از دور بعدی حسابرسی فراهم کند، زیرا به روزرسانی های جمع آوری داده ها ممکن است در هر زمانی اتفاق بیفتد. نرم افزارهای مدرن هوش مصنوعی به عنوان یک دستگاه پزشکی، شامل گزینه «تغییر پروتکل ها» هستند تا به دستگاه های هوش مصنوعی اجازه دهند با توجه به معیارهای بالینی مرتبط، به مرور زمان بهبود یابند.

Evaluating the effect of alignment set size on Scanner B



Evaluating the effect of reference set size on Scanner B



هرکادر، صدک های ۲۵٪، ۵۰٪ و ۷۵٪ توزیع بوت استرپ را نشان می دهد؛ ویسکرها نشان دهنده صدک های ۵٪ و ۹۵٪ هستند و هر نقطه ای خارج از این محدوده به عنوان یک داده پرت نمایش داده می شود. این تحلیل نشان می دهد که با تعداد کمی حدود ۲۵۰ مورد در مجموعه هم تراز، ما در حال حاضر نتایج بسیار خوبی به دست می آوریم که در آزمایش های نمونه برداری مکرر پایدار است. در تحلیل زیر، عملکرد روش ها را برای اندازه های مختلف مجموعه مرجع اندازه گیری کردیم، از ۵۰۰ مورد برای هم تراز استفاده کردیم و اندازه مجموعه مرجع را تغییر دادیم، می توانیم ببینیم که این روش به اندازه مرجع خیلی حساس نیست، حتی اگر توزیع مرجع شامل داده های بیشتری باشد، نتایج بهتری حاصل می شود. داده های منبع به عنوان یک فایل داده منبع ارائه می شوند.

شکل 6 | تحلیل حساسیت روی تأثیر اندازه های مجموعه های هم تراز و مرجع بر عملکرد روش هم تراز (اسکنر B). در تحلیل حساسیت روی اندازه مجموعه هم تراز (بالا)، از مجموعه مرجع کامل (۳۲۲۱ مورد) استفاده کردیم. نتایج روی ۵۰۰ نمونه بوت استرپ گزارش شده اند. برای تحلیل اندازه هم تراز، هر نمونه بوت استرپ با نمونه برداری از یک مجموعه هم تراز با اندازه مورد نظر از تمام موارد موجود و همچنین یک مجموعه ارزیابی ایجاد می شود (ن = ۲۵۰۰ مورد). برای تحلیل اندازه مرجع، هر نمونه بوت استرپ با نمونه گیری از یک مجموعه مرجع با اندازه مورد نظر از تمام موارد موجود و همچنین یک مجموعه ارزیابی ایجاد می شود (ن = ۲۵۰۰ مورد). در سمت چپ، نقاط میانگین SEN/SPC راروی نمونه ها نشان می دهند و میله های خط نشان دهنده فاصله اطمینان بوت استرپ ۹۵٪ هستند. در سمت راست،

عملکرد مدل روی مجموعه ارزیابی پس از هم تراز برای سنجش اثربخشی روش پیشنهادی در کاهش رانش حساسیت-ویژگی. نمونه برداری از مجموعه هم تراز و ارزیابی ۵۰۰ بار تکرار می شود تا تخمین های عدم قطعیت برای معیارهای گزارش شده به دست آید. توجه داشته باشید که در مجموعه داده های غربالگری سینه، برای هر مورد، ۴ تصویر در هر مطالعه داریم.

جزئیات مربوط به «سناریوهای ۲ تا ۴: به روزرسانی های مداوم مدل» در این سناریوها، ما می خواهیم اثربخشی روش خود را برای به روزرسانی های مداوم مدل در حضور تغییر تدریجی زمانی برای کار ماموگرافی بررسی کنیم. برای این منظور، ما یک جریان داده را با تعریف یک تابع نگاشت زمان به یک توزیع اسکنر معین که هر یک از سناریوهای فوق را توصیف می کند، شبیه سازی می کنیم. برای بدست آوردن مجموعه مرجع و مجموعه های شبیه سازی برای اسکنر A، داده های اصلی را تقسیم می کنیم. دی. مرجع به دو زیرمجموعه مجزا تقسیم می شود: یک زیرمجموعه تصادفی نمونه گیری شده از 1200 مورد به عنوان مجموعه مرجع (برای تعیین آستانه طبقه بندی شروع و به عنوان هدف هم تراز در طول شبیه سازی) استفاده می شود؛ سپس بخش باقیمانده شامل 2221 مورد به عنوان منبع داده برای خود شبیه سازی استفاده می شود. در اینجا فرض می کنیم که زمان

در آموزش، آستانه طبقه بندی را در نقطه عملیاتی متعادل (که در آن حساسیت برابر با ویژگی است) روی مجموعه اعتبارسنجی درون توزیعی انتخاب می کنیم. دی. مرجع، که ما آن را با مجموعه مرجع نشان می دهیم (به جداول مراجعه کنید) ۱ و ۲. توجه داشته باشید که مدل های آموزش دیده از نظر ROC-AUC به خوبی به تنظیمات خارج از توزیع خود تعمیم داده می شوند، اما از نظر حساسیت/اختصاصیت (همانطور که در بخش نتایج نشان داده شده است) این طور نیستند.

جزئیات مربوط به «سناریو ۱: استقرار در یک سایت جدید» ما از مدل هوش مصنوعی به همراه یک آستانه طبقه بندی از پیش تعیین شده که روی ... بهینه سازی شده است، استفاده می کنیم. دی. مرجع. ما ابتدا عملکرد این مدل را هنگام استقرار در یک سایت با نمونه برداری از یک مجموعه ارزیابی شامل ۲۵۰۰ مورد از مجموعه داده های خارج از توزیع (به طور معمول ۱۵۰۰۰ وصله برای کاربرد هیستوپاتولوژی) اندازه گیری می کنیم و عملکرد مدل را از نظر حساسیت و اختصاصیت روی این مجموعه ارزیابی می کنیم. علاوه بر این، ما یک مجموعه مجزا از N = ۱۰۰۰ مورد (به طور معمول ۵۰۰۰ وصله) را برای تطبیق با الگوریتم هم تراز پیش بینی نمونه برداری می کنیم. برای کاربرد ماموگرافی، از مجموعه مرجع کامل استفاده می کنیم. دی. مرجع به عنوان توزیع مرجع ما (3221 مورد). برای کار هیستوپاتولوژی، 5000 تصویر از دی. مرجع به عنوان توزیع مرجع خود استفاده کنیم. سپس دوباره ارزیابی می کنیم

دردسترس بودن کد

پیاده سازی UPA به همراه کدی برای بازتولید تمام نتایج و ارقام در آیین نامه تکمیلی ارائه شده همچنین در دسترس عموم قرار گرفته است. biomed-mira.upa.github.com این کد با پایتون ۳.۱۰ و به همراه بسته های زیر نوشته شده است: 2.0.2, scikit_learn-1.2.2, jupyter-1.0.0, tqdm-4.65.0, notebook-6.5.4, seaborn-0.12.2, matplotlib-3.7.1, numpy-1.25, pandas

منابع

۱. کیم، سی کی و همکاران. یک خط لوله تریاژ خودکار کووید-۱۹ با استفاده از هوش مصنوعی بر اساس رادیوگرافی قفسه سینه و داده های بالینی. *مجله پزشکی Digit* ۵۵.۵ (۲۰۲۲).
۲. لیبینگ، سی. و همکاران. ترکیب نقاط قوت رادیولوژیست ها و هوش مصنوعی برای غربالگری سرطان سینه: یک تحلیل گذشته نگری. *نست دیجیت، سلامت* ۴، e507-e519 (2022).
۳. مک کینی، اس ام و همکاران. ارزیابی بین المللی یک سیستم هوش مصنوعی برای غربالگری سرطان سینه. *طبیعت* ۸۹، ۹۴-۸۹۷ (۲۰۲۰).
۴. لارسن، م.، آگلن، سی اف، هاف، اس آر، لوند-هانسن، اچ. و هافویند، اس. استراتژی های ممکن برای استفاده از هوش مصنوعی در خواندن ماموگرافی ها بر اساس داده های گذشته نگر از ۱۲۲،۹۶۹ آزمایش غربالگری. *یورو رادیول* ۸۲۳۸، ۳۲-۸۲۴۶ (۲۰۲۲).
۵. وریس، سی. دی و همکاران. تأثیر سیستم های مختلف ماموگرافی بر عملکرد هوش مصنوعی در غربالگری سرطان سینه. *رادیو، مصنوعی، هوشمند* ۳، ۵ (۲۰۲۳).
۶. فینلیسون، اس. جی و همکاران. تغییر پزشک و مجموعه داده ها در هوش مصنوعی. *مجله پزشکی انگلستان (N. Engl. J. Med)* ۳۸۵، ۲۸۳-۲۸۶ (۲۰۲۱).
۷. Lundstrom, C., Stacke, K., Eilertsen, G., Unger, J. & IEEE Biomed. Health Inf 7، ۳۳۶-۳۴۵ (۲۰۲۱).
۸. ژو، ک.، لیو، ز.، کیائو، ی.، شیانگ، ت. و لوی، CC تعمیم دامنه: یک بررسی. <https://arxiv.org/abs/2103.02503> (۲۰۲۱).
۹. گوش، دی. و اکتا گوش، دی. یک تحقیق چند مرکزی در مقیاس بزرگ در مورد تعمیم دامنه در تشخیص توده مبتنی بر یادگیری عمیق در ماموگرافی: یک بررسی. *درانجمن زیست شناسی آکتا* ۹۰-۹۵ (طبقه تحقیقاتی، ۲۰۲۲).
۱۰. برین، جی.، زوکر، ک.، اورسی، ان ام و راویکومار، ن. ثبت تصاویر زیست پزشکی، تعمیم دامنه و تحلیل خارج از توزیع ۱۴-۲۲ (انتشارات بین المللی اشپرینگر، ۲۰۲۲).
۱۱. جنکینز، دی. ای و همکاران. به روزرسانی و نظارت مداوم بر مدل های پیش بینی بالینی: آیا زمان سیستم های پیش بینی پویا فرا رسیده است؟ *تشخیص، پیش بینی، تحقیق* ۱، ۵ (۲۰۲۱).
۱۲. بینویا، ام ای، انگلهرت، ای جی، شاتس، دلیو، اشمیت، ام کی و استایربرگ، ای دلیو. راهنمای روش شناختی برای ارزیابی و به روزرسانی مدل های پیش بینی بالینی: یک بررسی سیستماتیک. *روش شناسی پژوهش های پزشکی BMC* ۳۱۶، ۲۲ (۲۰۲۲).
۱۳. مک کورمیک، تی اچ، رفرتی، ای ای، مادیگان، دی. و برد، آر اس. رگرسیون لجستیک پویا و میانگین گیری مدل پویا برای طبقه بندی دودویی. *بیومتریک* ۲۳، ۳۰-۳۸ (۲۰۱۲).
۱۴. هیکی، جی ال و همکاران. رویکردهای مدل سازی پیش بینی پویا برای جراحی قلب. *دورقلب و عروق، کیفیت نتایج* ۶، ۶۴۹-۶۵۸ (۲۰۱۳).
۱۵. ورگوو، وای. و همکاران. یک رویه آزمایش بسته برای انتخاب روش مناسب برای به روزرسانی مدل های پیش بینی. *آمار، پزشکی* ۳۶، ۴۵۲۹-۴۵۳۹ (۲۰۱۷).
۱۶. ون کالستر، ب. و همکاران. اعتبارسنجی و به روزرسانی مدل های ریسک مبتنی بر رگرسیون لجستیک چندجمله ای. *تشخیص، پیش بینی، تحقیق* ۲، ۱۷ (۲۰۱۷).
۱۷. اشلینگر، ام ای، یانگ، دلیو. و کیم، اس ای. مقایسه استراتژی های به روزرسانی پویا برای مدل های پیش بینی بالینی. *تشخیص، پیش بینی، تحقیق* ۲، ۵ (۲۰۲۱).
۱۸. لوی، تی جی و همکاران. توسعه و اعتبارسنجی مدل های پیش آگهی خودپایشی و به روزرسانی خودکار بقا برای بیماران بستری شده مبتلا به کووید-۱۹. *نات، کمون* ۱۳، ۶۸۱۲ (۲۰۲۲).
۱۹. رابانسر، س.، گونمن، س. و لپیتون، ز. شکست یا صدای بلند: مطالعه تجربی روش های تشخیص تغییر مجموعه داده ها. *فرآیند اطلاعات عصبی پیشرفته، سیستم* ۳۲، ۱۳۹۶، ۱۴۰۸-۱۴۰۹ (۲۰۱۹).

واحد در این شبیه سازی ها بر حسب هفته است. برای هر نقطه زمانی T، داده ها را طبق توزیع اسکرنر مورد انتظار در نقطه T - همانطور که توسط هر سناریو تعریف شده است - نمونه برداری می کنیم. در هر نقطه زمانی معین T، از یک پنجره اجرایی متشکل از داده های دو هفته گذشته T-2 و T-1 برای برآزش هر دو الگوریتم هم تراز (به طور جداگانه برای اسکرنر) استفاده می کنیم. سپس عملکرد مدل را قبل و بعد از هم تراز روی مجموعه نمونه برداری شده برای زمان T، هم از نظر اسکرنر و هم به طور کلی، ارزیابی می کنیم. علاوه بر این، اطمینان حاصل می کنیم که - در هر نقطه زمانی معین T، مجموعه پنجره اجرایی (نمونه های T-1 و T-2) برای مجموعه ارزیابی (نمونه های T) از هم جدا هستند. در سناریوهای 2 و 4 فرض می کنیم که تعداد کل موارد در هر نقطه زمانی (هفته) در n = ۲۵ مورد. در سناریوی ۳، ما با n = ۲۵ مورد در ابتدای شبیه سازی و به تدریج تعداد موارد را بزرگ کرده با افزایش اسکرنر دوم افزایش می یابد تا به ... برسیم. n = ۵۰. در سناریوی ۴، ما به روزرسانی نرم افزار را با افزایش وضوح تصاویر ورودی شبیه سازی کردیم. در این سناریو نیز از تصاویر اسکرنر A استفاده کردیم.

اطلاعات اخلاقی

مجموعه داده های داخلی سرطان سینه از بریتانیا و مجارستان قبلاً با تأیید اخلاقی از سازمان تحقیقات بهداشتی خدمات ملی بهداشت بریتانیا (NHS) (HRA) (مرجع: HRA/0376/19) و شورای تحقیقات پزشکی، کمیته اخلاق علمی و پژوهشی در مجارستان (ETT-TUKEB) (مرجع: OGYÉI/46651-4/2020) جمع آوری شده بودند. مطالعه اصلی مطابق با اصول مندرج در اعلامیه هلسینکی برای همه تحقیقات تجربی انسانی انجام شد. لزوم رضایت آگاهانه برای شرکت در مطالعه توسط HRA و TUKEB-ETT بررسی و تأیید شد که نیازی به آن نیست زیرا این مطالعه شامل استفاده ثانویه از داده های گذشته نگر و با نام مستعار بود. مطالعه حاضر از نسخه کاملاً ناشناس این داده های جمع آوری شده قبلی استفاده ثانویه کرد. برای مجموعه داده های سرطان سینه OPTI-MAM، پروژه OPTIMAM در ژوئیه ۲۰۱۹ تأییدیه HRA و نظر اخلاقی مساعد را برای یک دوره ۵ ساله قابل تمدید (مرجع: SC/0284/۱۹) برای جمع آوری تصاویر و داده ها از سایت های شرکت کننده جهت ایجاد یک پایگاه داده تحقیقاتی و افزودن سایت های جمع آوری جدید، دریافت کرد. رضایت بیمار توسط کمیته اخلاق تحقیقاتی Oxford C—NHS HRA South Central طبق قانون اساسی NHS لغو شد که تصریح می کند جمع آوری داده های بدون هویت بدون رضایت بیمار مجاز است. مطالعه حاضر از یک نسخه کاملاً ناشناس از مجموعه داده های OPTIMAM استفاده کرد. استفاده از مجموعه داده های هیستوپاتولوژی از تأیید اخلاقی معاف است زیرا تجزیه و تحلیل بر اساس داده های ثانویه کاملاً ناشناس است که به صورت عمومی در دسترس هستند.

خلاصه گزارش

اطلاعات بیشتر در مورد طراحی تحقیق در خلاصه گزارش نمونه کارهای طبیعت که به این مقاله لینک شده است، موجود است.

دردسترس بودن داده ها

تمام خروجی های مدل ناشناس که از یافته های شرح داده شده در این نسخه خطی پشتیبانی می کنند، به صورت عمومی در مخزن کد ما به آدرس در دسترس هستند. <https://github.com/biomed-mira-upa> شامل داده های مورد استفاده برای رسم نمودارهای نشان داده شده در شکل ها و همچنین جداول است. دسترسی به مجموعه داده های سرطان سینه OPTIMAM را می توان در وب سایت پروژه درخواست کرد: <https://medphys.royalsurrey.nhs.uk> مجموعه داده WILDS-Camelyon تحت مجوز Creative Commons CC0 به عنوان بخشی از معیار WILDS در دسترس عموم است و به راحتی برای دانلود در دسترس همه قرار دارد. دستورالعمل های داناود را می توانید (شروع شده) https://wids.stanford.edu/get_started/. به عنوان بخشی از این مطالعه، از تقسیم بندی های رسمی داده ها استفاده شد. تصاویر خام از مجموعه داده های تصویربرداری داخلی سرطان سینه از بریتانیا و مجارستان تحت مجوزهای تجاری به دست آمده و در دسترس عموم نیستند. درخواست اطلاعات بیشتر را می توان از طریق ایمیل به نویسندگان مربوطه ارسال کرد و ظرف چهار هفته پردازش خواهد شد. داده های منبع در این مقاله ارائه شده است.

