

توصیف کامل و جامع مقاله: داده‌های بزرگ در مدل‌سازی چندمقیاسه: از پردازش تصاویر پزشکی تا مدل‌های شخصی‌سازی شده

**Big Data in multiscale modelling:
from medical image processing to personalized models**

این مقاله با عنوان "Big Data in multiscale modelling: from medical image processing to personalized models" توسط نویسندگان Tijana Geroski، Djordje Jakovljević و Nenad Filipović نوشته شده است. مقاله در سال ۲۰۲۳ در مجله Journal of Big Data (جلد ۱۰، مقاله ۷۲) منتشر شده و دارای doi: <https://doi.org/10.1186/s40537-023-00763-y> است. این مقاله دسترسی باز (Open Access) دارد و تحت مجوز Creative Commons Attribution 4.0 منتشر شده است. مقاله بر روی نقش داده‌های بزرگ (Big Data) در مدل‌سازی چندمقیاسه تمرکز دارد، با تأکید بر فرآیند از پردازش تصاویر پزشکی تا ایجاد مدل‌های شخصی‌سازی شده و تحلیل آن‌ها. نویسندگان به بررسی چالش‌های صنعت مراقبت‌های بهداشتی می‌پردازند که شامل حساسیت داده‌های بیمار، نیاز به ذخیره‌سازی ایمن و مطابق با مقررات، و دقت بالای پیش‌بینی است. مقاله شامل بررسی ادبیات، تعریف مسئله، فناوری‌های مرتبط، راه‌حل‌های موجود، و دو راه‌حل پیشنهادی است: پردازش تصاویر اولتراسوند شریان کاروتید و بازسازی سه‌بعدی، و آزمایش دارو روی مدل‌های قلب شخصی‌سازی شده. در ادامه، مقاله را به طور کامل و جامع به زبان فارسی توصیف می‌کنم، با ساختار بندی بر اساس بخش‌های اصلی، و پوشش تمام جنبه‌های کلیدی شامل چکیده، مقدمه، روش‌ها، نتایج، بحث و نتیجه‌گیری.

چکیده (Abstract)

صنعت مراقبت‌های بهداشتی با سایر صنایع متفاوت است؛ داده‌های بیمار حساس هستند، ذخیره‌سازی آن‌ها باید با دقت و مطابق با مقررات انجام شود، و دقت پیش‌بینی باید بالا باشد. گسترش سریع روش‌های تصویربرداری پزشکی و جمع‌آوری داده‌ها منجر به تولید "داده‌های بزرگ" شده که تحلیل آن توسط متخصصان پزشکی زمان‌بر است. این مقاله نگاهی به داده‌های بزرگ از منظر نقش آن در مدل‌سازی چندمقیاسه دارد. توجه ویژه‌ای به جریان کاری شده، از پردازش تصاویر پزشکی تا ایجاد مدل‌های شخصی‌سازی شده و تحلیل آن‌ها. بررسی ادبیات در مورد داده‌های بزرگ در مراقبت‌های بهداشتی ارائه شده و دو راه‌حل پیشنهادی توصیف شده: پردازش تصاویر اولتراسوند شریان کاروتید و بازسازی سه‌بعدی، و آزمایش دارو روی مدل‌های قلب شخصی‌سازی شده.

در مورد پردازش تصاویر اولتراسوند شریان کاروتید، نقطه شروع تصاویر اولتراسوند است که با استفاده از شبکه عصبی کانولوشنی U-Net تقسیم‌بندی (segmented) می‌شوند، و ماسک‌های تقسیم‌بندی‌شده برای بازسازی سه‌بعدی هندسه استفاده می‌شوند. در مورد آزمایش دارو روی مدل قلب شخصی‌سازی‌شده، رویکرد مشابهی پیشنهاد شده؛ تصاویر برای ایجاد مدل هندسی سه‌بعدی شخصی‌سازی‌شده استفاده می‌شوند که در مدل‌سازی محاسباتی برای تعیین فشار در بطن چپ قبل و بعد از آزمایش دارو به کار می‌رود. هر دو روش پیچیده هستند، شامل تحلیل داده‌های بزرگ می‌شوند و باید با استفاده از سرورها یا محاسبات با کارایی بالا انجام شوند. توسعه آینده کاربردهای داده‌های بزرگ در حوزه‌های مراقبت‌های بهداشتی پتانسیل زیادی دارد، به دلیل استانداردهای جدید داده، توسعه سریع تحقیق و فناوری، و انگیزه‌های قوی دولتی.

کلمات کلیدی: داده‌های بزرگ، مدل‌سازی چندمقیاسه، پردازش تصاویر پزشکی، بازسازی سه‌بعدی.

مقدمه (Introduction)

اصطلاح "داده‌های بزرگ" در سال‌های اخیر به یک واژه رایج تبدیل شده، زیرا 频率 استفاده از آن هر سال دو برابر شده است. داده‌های بزرگ با سه ویژگی اصلی شناخته می‌شود که به عنوان "۳V" شناخته می‌شوند: حجم (مقدار داده تولیدشده)، تنوع (داده از دسته‌های متعدد)، و سرعت (نرخ تولید داده). اخیراً دو "V" دیگر اضافه شده: تغییرپذیری (ناسازگاری داده) و اعتبار (کیفیت داده ثبت‌شده)، که در مجموع "۵V" را تشکیل می‌دهند. کاربردهای داده‌های بزرگ در رشته‌های مختلفی مانند کشاورزی، اینترنت با شبکه‌های اجتماعی، پزشکی، و پزشکی شخصی‌سازی‌شده بر اساس داده‌های ژنومیک وجود دارد. در حوزه پزشکی، حجم داده در حال گسترش است و روش‌های سنتی نمی‌توانند این مقادیر را به طور مناسب مدیریت کنند. مدیریت، تحلیل و ذخیره‌سازی داده‌های بیولوژیکی چالش‌های مداوم در محاسبات زیست‌پزشکی هستند. فناوری‌های داده‌های بزرگ چارچوب‌های جدیدی برای پردازش داده‌های پزشکی ارائه می‌دهند و نقش مهمی در مدیریت، سازماندهی و تحلیل داده از طریق یادگیری ماشین و یادگیری عمیق ایفا می‌کنند. همچنین، دسترسی سریع به داده از طریق پایگاه‌های داده NoSQL را امکان‌پذیر می‌کنند.

در تحلیل تصاویر پزشکی، بهبود قابل توجه در تجهیزات جمع‌آوری تصاویر منجر به تولید داده‌های نسبتاً بزرگ شده که تحلیل تصاویر را چالش‌برانگیز می‌کند. دیجیتال‌سازی مخازن پزشکی در بیمارستان‌ها و استفاده از تصاویر پزشکی باعث رشد نمایی اندازه آرشیوهای پزشکی دیجیتال شده است. طبق گزارش موسسه جهانی مک‌کینزی، اگر مراقبت‌های بهداشتی ایالات متحده از داده‌های بزرگ به طور خلاقانه و کارآمد استفاده کند، می‌تواند بیش از ۳۰۰ میلیارد دلار ارزش سالانه تولید کند، که دو سوم آن از طریق کاهش هزینه‌های بهداشتی

ایالات متحده محقق می‌شود. گسترش سریع تصاویر پزشکی و روش‌ها نیاز به تلاش‌های قابل توجه و زمان‌بر توسط متخصصان پزشکی دارد که ذهنی، مستعد خطای انسانی، و دارای تفاوت‌های بین‌فردی هستند. استفاده از تکنیک‌های یادگیری ماشین برای خودکارسازی فرآیند تشخیص جایگزین مناسبی است؛ با این حال، روش‌های یادگیری ماشین سنتی نمی‌توانند با مشکلات پیچیده کنار بیایند. ترکیب کامپیوترهای پرسرعت با یادگیری ماشین نوید می‌دهد که مقادیر زیادی از داده‌های تصاویر پزشکی را برای تشخیص دقیق و سریع مدیریت کند. در سال‌های اخیر، یادگیری ماشین (ML) و هوش مصنوعی (AI) به سرعت پیشرفت کرده‌اند و نقش خود را در پردازش تصاویر پزشکی، تشخیص کمک‌شده توسط کامپیوتر، ترکیب تصاویر، ثبت، تقسیم‌بندی تصاویر، و درمان هدایت‌شده توسط تصاویر پیدا کرده‌اند. تکنیک‌های ML اطلاعات (ویژگی‌ها) را از تصاویر استخراج می‌کنند و تصمیم‌گیری را به طور موثر انجام می‌دهند.

تعریف مسئله (Problem Definition)

داده‌های بزرگ در سلامت به مجموعه داده‌های مرتبط اشاره دارد که بزرگ، زمان‌بر و پیچیده هستند و پزشکان مراقبت‌های بهداشتی نمی‌توانند آن‌ها را با فناوری‌های فعلی مدیریت و پردازش کنند. داده‌ها با نرخ بی‌سابقه‌ای روزانه از منابع ناهمگن متعدد (مانند داده‌های آزمایشگاهی و بالینی، علائم بیماران از طریق حسگرهای دور، فعالیت‌های بیمارستان، و داده‌های دارویی) تولید می‌شوند. در نتیجه، چالش‌های جدیدی مانند ذخیره‌سازی، جمع‌آوری و تفسیر volumes عظیم داده ظاهر شده است. تکنیک‌های تصویربرداری زیست‌پزشکی که در تنظیمات بالینی مستقر هستند شامل توموگرافی کامپیوتری (CT)، تصویربرداری رزونانس مغناطیسی (MRI)، اشعه ایکس، تصویربرداری مولکولی، اولتراسوند، تصویربرداری فوتوآکوستیک، فلوروسکوپی، و توموگرافی انتشار پوزیترون (PET-CT) هستند. این رویکردها تصاویر پزشکی با وضوح بالا در مقادیر عظیم ارائه می‌دهند. با این حال، پزشکان نمی‌توانند تمام میلیون‌ها تصویر تولیدشده را تشخیص دهند. با افزایش دسترسی به داده‌های تصویربرداری زیست‌پزشکی، تقاضاهای اضافی بر روی هوش مصنوعی (AI) برای سیستم‌های یادگیری ماشین (ML) برای ساخت مدل‌های پیچیده قرار می‌گیرد. ML مکانیسم اساسی برای استخراج اطلاعات ساخت‌یافته و دانش از داده‌های خام است و آن را به پیش‌بینی‌های خودکار برای کاربردهای مختلف تبدیل می‌کند. اطلاع‌رسانی تصویربرداری به طور کلی برای بهبود کارایی فعالیت‌های پردازش تصاویر مانند ذخیره‌سازی، بازیابی و تعامل استفاده می‌شود. در این بخش، بر دو جنبه مهم تمرکز می‌کنیم: ذخیره‌سازی داده‌های پزشکی و پردازش تصاویر در حوزه داده‌های بزرگ.

ذخیره‌سازی داده‌های پزشکی در عصر داده‌های بزرگ (Medical Data Storage in the Era of Big Data)

داده‌های تصاویر پزشکی می‌توانند از چند مگابایت برای یک مطالعه تک (مانند تصاویر بافت‌شناسی) تا صدها گیگابایت برای هر تحقیق (مانند بررسی‌های CT نازک‌لایه با بیش از ۲۵۰۰ اسکن برای هر مطالعه) متغیر باشند. برای مثال، مجموعه تصاویر پزشکی ImageCLEF بین سال‌های ۲۰۰۵ تا ۲۰۰۷ حدود ۶۶,۰۰۰ تصویر داشت، اما در سال ۲۰۱۳، حدود ۳۰۰,۰۰۰ تصویر روزانه ذخیره می‌شد. برای پردازش مقادیر بزرگ داده‌های مراقبت‌های بهداشتی، ذخیره‌سازی باید با جنبه‌هایی مانند رعایت مقررات، امنیت و حریم خصوصی، هزینه‌آثر بخشی، دسترسی و قابلیت اطمینان همخوانی داشته باشد (شکل ۱: ذخیره‌سازی داده‌های بهداشتی، که شامل بخش‌هایی مانند رعایت مقررات، هزینه‌آثر بخشی، دسترسی و قابلیت اطمینان، امنیت و حریم خصوصی، و داده‌های ساخت‌یافته و غیرساخت‌یافته است).

رویکردهای داده‌های بزرگ مانند Hadoop و پایگاه‌های داده NoSQL برای ذخیره رکوردهای الکترونیکی سلامت (EHRs) استفاده می‌شوند. هنگام کار با داده‌های جریان واقعی‌زمان بالینی، ذخیره‌سازی داده مؤثر حیاتی است. مطالعات استفاده از Hadoop و HBase به عنوان انبارهای داده برای ذخیره داده‌های EEG را ارزیابی کرده‌اند و ویژگی‌های عملکرد بالا را مورد توجه قرار داده‌اند. سیستم‌های توزیع‌شده برای ذخیره و پرس‌وجو volumes عظیم داده‌های EEG ارائه شده است. Cloudwave، سیستمی که برای ذخیره داده‌های بالینی با استفاده از ماژول‌های پردازش داده مبتنی بر Hadoop طراحی شده، شامل رابط وب‌محور برای 可視 سازی و تحلیل داده واقعی‌زمان است و صرفه‌جویی زمانی قابل توجهی به دست می‌آورد (مثلاً پردازش پنج مطالعه EEG در ۱ دقیقه در مقایسه با ۲۰ دقیقه در سیستم مستقل). امکان‌پذیری استفاده از Hadoop HDFS و HBase برای EHRهای توزیع‌شده بررسی شده است، با NoSQL گزینه پیشرفته‌ای برای ذخیره داده‌های غیرساخت‌یافته در مقایسه با پایگاه‌های داده رابطه‌ای سنتی که در مدیریت داده‌های ساخت‌یافته خوب هستند. سیستمی که فناوری‌های رابطه‌ای و چندبعدی را با مخازن NoSQL ادغام می‌کند برای امکان‌پذیر کردن تکنیک‌های داده‌کاوی در حالی که انعطاف‌پذیری و سرعت در پردازش داده ارائه می‌دهد، ارائه شده است. یک سیستم نمونه اولیه برای ذخیره داده‌های سیگنال بالینی از HBase استفاده می‌کند، که کلید ردیف به عنوان زمان‌تمبر یک مقدار واحد عمل می‌کند، و ستون مقادیر فیزیولوژیکی بیمار را که با زمان‌تمبر کلید ردیف مطابقت دارد، ذخیره می‌کند. اطلاعات ساختار داده HBase برای افزایش دسترسی و خوانایی در MongoDB، یک پایگاه داده مبتنی بر سند، ذخیره می‌شود.

سیستم بایگانی و ارتباطات تصاویر (PACS) یکی از خدمات استفاده‌شده در پزشکی برای ذخیره و انتقال داده‌های تصاویر است که عمده‌تاً از طریق پروتکل‌های DICOM در بخش‌های رادیولوژی انجام می‌شود. PACS تصاویر را به ایستگاه‌های نمایش محلی با استفاده از پروتکل‌های موجود منتقل می‌کند، اما انتقال داده با

PACS بسیار استاندارد است و این سیستم کاملاً به داده‌های ساخت‌یافته برای بازیابی تصاویر پزشکی وابسته است نه استفاده از اطلاعات غیرساخت‌یافته تصاویر زیست‌پزشکی. برای دسترسی به PACS، چندین برنامه پزشکی وب‌محور توسعه یافته و استفاده رو به رشد از فناوری‌های داده‌های بزرگ عملکرد آن‌ها را بهبود بخشیده است. مطالعات زیادی برای مدیریت و تحلیل تصاویر داده‌های ساخت‌یافته و غیرساخت‌یافته با استفاده از مفاهیم داده‌های بزرگ و هوش مصنوعی انجام شده است. با توجه به روند فعلی در میان سازمان‌های مراقبت‌های بهداشتی برای برون‌سپاری دو جزء حیاتی PACS (مخزن شیء DICOM و سیستم پایگاه داده) به ابر، تکنیکی برای ادغام داده در PACS پیشنهاد شده است. آن‌ها روش جریان ورودی/خروجی ابر را در یک لایه انتزاعی برای تطبیق ارائه‌دهندگان ابر متعدد بدون توجه به تفاوت در استانداردهای دسترسی داده پیشنهاد کردند. یک سیستم بازیابی تصاویر پزشکی بزرگ مبتنی بر Hadoop ساخته شده که ویژگی‌های تصاویر پزشکی را با استفاده از تبدیل Brushlet و الگوریتم الگوی باینری محلی بازیابی می‌کند، علاوه بر فناوری‌های داده‌های بزرگ مبتنی بر ادغام پلتفرم‌های ابر با PACS. ویژگی‌های تصاویر سپس در HDFS ذخیره می‌شوند، به دنبال اجرای MapReduce. در مقایسه با نتایج بدون فیلتر همومورفیک، نتایج ارزیابی نرخ خطای پایین‌تری نشان دادند. به طور مشابه، مسائل سیستم‌های بازیابی تصاویر مبتنی بر محتوا با استفاده از معماری پردازش MapReduce و مدل ذخیره HDFS مورد توجه قرار گرفته است. آن‌ها آزمایش روی مجموعه داده‌های ماموگرافی انجام دادند و نتایج خوبی به دست آوردند، که نشان می‌دهد رویکرد MapReduce می‌تواند به طور مؤثر برای بازیابی تصاویر پزشکی مبتنی بر محتوا استفاده شود.

ذخیره‌سازی بلندمدت داده‌های بزرگ در پزشکی نیاز به ظرفیت ذخیره‌سازی عظیم دارد. اگر هر سیستم حمایت از تصمیم‌گیری با استفاده از داده انجام شود، همچنین نیاز به الگوریتم‌های سریع و دقیق دارد. علاوه بر این، اگر منابع اضافی داده برای هر بیمار به دست آید، نیاز به ذخیره‌سازی منسجم دارد. روش‌های فشرده‌سازی می‌توانند محدودیت‌های ذخیره‌سازی داده و ظرفیت شبکه را حل کنند. رویکردهایی برای فشرده‌سازی مجموعه داده‌های توالی با توان بالا و داده‌های محاسبات احتمال خطا log-odds توسعه یافته، که نسبت‌های فشرده‌سازی ۴۰۰ و ۵ را با استفاده از فیلتر و تبدیل فوریه به دست می‌آورند. نظریه SP هوش برای داده‌های عظیم بررسی شده، که هدف آن ساده‌سازی و ترکیب مفاهیم از حوزه‌های مختلف برای فشرده‌سازی بدون از دست دادن با تطبیق و یکپارچه‌سازی الگوها است، هرچند هنوز در حال توسعه است.

فناوری‌های داده‌های بزرگ برای پردازش تصاویر پزشکی (Big Data Technologies for Medical Image Processing)

محاسبات موازی برای مدیریت داده‌های بزرگ ضروری است و امکان تحلیل روی کلاسترها یا فوق کامپیوترها را به طور همزمان فراهم می‌کند. فناوری داده‌های بزرگ با AI و محاسبات موازی عظیم می‌تواند پیش‌بینی و پزشکی شخصی‌سازی شده را انقلاب کند. مدل‌های محاسباتی موازی نوین، مانند MapReduce گوگل، برای زیرساخت داده‌های بزرگ پیشنهاد شده است. Apache Hadoop، یک نرم‌افزار MapReduce منبع باز، دسترسی همزمان به داده را از طریق سیستم فایل توزیع شده Hadoop (HDFS) پشتیبانی می‌کند. خدمات مبتنی بر Hadoop می‌توانند پلتفرم‌های محاسبات ابری برای ذخیره‌سازی مرکزی و دسترسی دور باشند. محاسبات ابری منابع محاسباتی را در سراسر شبکه توزیع می‌کند و به عنوان زیرساخت، پلتفرم یا نرم‌افزار عمل می‌کند، سرعت، چابکی و انعطاف‌پذیری را افزایش می‌دهد در حالی که نیازهای نگهداری را کاهش می‌دهد. چارچوب‌های Hadoop و Spark برای تحلیل تصاویر زیست‌پزشکی بهینه هستند. محاسبات با کارایی بالا (HPC) از پردازش موازی و برنامه‌های پیشرفته برای سرعت بخشیدن به محاسبات عظیم استفاده می‌کند، با روش عناصر محدود (FEM) که برای محاسبات بزرگ‌مقیاس به HPC وابسته است. HPC در پزشکی درون داده‌های بزرگ برای محاسبات چندمقیاسه عظیم استفاده می‌شود. Apache Spark، یک پلتفرم محاسباتی توزیع شده، نیازهای پردازش داده‌های بزرگ را با استفاده از مجموعه داده‌های مختلف از منابع متعدد مدیریت می‌کند. Spark کمبودهای Hadoop را برطرف می‌کند و عملکرد سریع‌تری ارائه می‌دهد (بیش از ۱۰۰ برابر در حافظه) با ذخیره‌سازی داده در حافظه و پردازش نزدیک واقعی‌زمان. چارچوب Spark داده را برای تکرار، پرس‌وجوی مکرر و بارگذاری حافظه آماده می‌کند، با یک برنامه اصلی که برده‌ها را نظارت می‌کند که پارتیشن‌های داده را می‌خوانند، محاسبات انجام می‌دهند و نتایج را می‌نویسند (شکل ۲: معماری Spark). Spark بر روی پردازش موازی MapReduce ساخته شده و امکان پرس‌وجوهای SQL، داده‌های جریان، یادگیری ماشین و پردازش گراف را فراهم می‌کند. افزایش برده‌ها می‌تواند عملکرد را بهبود بخشد، هرچند ممکن است زمان پردازش را افزایش دهد.

راه‌حل‌های موجود (Existing Solutions)

هرچند مطالعات داده‌های بزرگ در مراقبت‌های بهداشتی را بررسی می‌کنند، هیچ تحقیق قبلی فرآیند کامل مدیریت تصاویر پزشکی را توصیف نمی‌کند. تحقیقات موجود Hadoop را می‌شناسد که از MapReduce برای تحلیل و تبدیل مجموعه داده‌های بزرگ استفاده می‌کند. MapReduce امکان مقیاس‌پذیری در سرورهای کلاستر Hadoop برای کاربردهای واقعی جهان را فراهم می‌کند اما با کارهای سنگین ورودی-خروجی مشکل دارد. چارچوب MapReduce برای سرعت بخشیدن به پردازش تصاویر پزشکی بزرگ‌مقیاس استفاده شده و زمان اجرا برای پارامترهای بهینه SVM را از ۱۰۰۰ ساعت به ۱۰ ساعت کاهش داده است. Spark فناوری دیگری است که شناخته شده، با جریان کاری که الگوریتم‌های بهینه ادغام AI و ML برای

مدیریت تصاویر زیست‌پزشکی را ادغام می‌کند. فاز طبقه‌بندی می‌تواند برای مراحل دیگر تنظیم شود و از Hadoop و Spark با فناوری‌های NoSQL استفاده کند. پایگاه‌های داده NoSQL، معرفی شده در ۲۰۰۹، از پایگاه‌های داده رابطه‌ای متفاوت هستند و عملکرد بهتر برای داده‌های حجم بالا و تطبیق بومی برای معماری‌های توزیع شده ارائه می‌دهند. محاسبات ابری با منابع IT شبکه‌شده به اشتراک‌گذاری داده کمک می‌کند. مطالعات رویکردهای تقسیم‌بندی ارگان را مقایسه کرده‌اند و نتایج را با استفاده از کنتراست محلی و اطلاعات احتمال اطلس ۳۳٪ بهبود بخشیده‌اند. برخی نویسندگان شبکه‌های عصبی کانولوشنی عمیق برای طبقه‌بندی تصاویر پزشکی داده‌های بزرگ ادعا می‌کنند، اما مجموعه داده‌ها واقعاً داده‌های بزرگ نیستند (مثلاً ۳۰۰ تصویر در هر کلاس در ۱۲ کلاس). کارهای دیگر سیستم‌های مراقبت‌های بهداشتی را با استفاده از داده‌های بزرگ مدیریت می‌کنند و روش‌هایی برای معماری، الزامات و اجرا پیشنهاد می‌کنند. مفهوم جدیدی برای تحلیل تصاویر زیست‌پزشکی با استفاده از معماری داده‌های بزرگ پیشنهاد شده، با جریان کاری برای جمع‌آوری، تحلیل، ذخیره‌سازی، پردازش، پرس‌وجو، طبقه‌بندی و تشخیص با استفاده از داده‌های غیرساخت‌یافته و ساخت‌یافته از پایگاه‌های داده NoSQL. معماری Spark تکنیک‌های کارآمد برای طبقه‌بندی تعداد زیادی عکس می‌سازد. تأثیر تحلیل داده‌های بزرگ در مراقبت‌های بهداشتی نشان داده شده، با بررسی‌های ادبیات در مورد کاربردها در تحقیق زیست‌پزشکی و مراقبت‌های بهداشتی، و تحقیقات در مورد داده‌های بزرگ برای مراقبت‌های بهداشتی شخصی‌سازی شده. با ادغام فناوری‌های داده‌های بزرگ، مدیریت داده بیشتر و عملکرد بهتر به دست می‌آید. بررسی تکنیک‌های طبقه‌بندی تصاویر بیولوژیکی اهمیت ML در طبقه‌بندی تصاویر زیست‌پزشکی را نشان می‌دهد و زمان پردازش را هنگام جفت شدن با فناوری‌های داده‌های بزرگ کاهش می‌دهد. با این حال، عملکرد Spark ممکن است در استخراج ویژگی با تصاویر کوچک یا متنوع آسیب ببیند. تمرکز دیگر روی تحلیل داده‌های چند-ومیکس برای پزشکی شخصی‌سازی شده است. هرچند راه‌حل‌های موجود جنبه‌های تحلیل داده‌های بزرگ را مورد توجه قرار می‌دهند، هیچ راه‌حلی جریان کاری کامل خودکار از داده (تصاویر پزشکی)، تقسیم‌بندی، بازسازی سه‌بعدی، مدل‌سازی محاسباتی و آزمایش دارو ارائه نمی‌دهد. این مقاله اولین است که جریان کاری کامل با جزئیات روش‌شناسی و نتایج در این زمینه‌ها ارائه می‌دهد.

راه‌حل پیشنهادی برای پردازش تصاویر اولتراسوند شریان کاروتید (Proposed Solution for Carotid Artery Ultrasound Image Processing)

در این بخش، یک مورد استفاده از پردازش تصاویر اولتراسوند از پروژه TAXINOMISIS ارائه شده، شامل مجموعه داده جمع‌آوری شده، روش‌های پیشنهادی و نتایج به دست آمده. هیچ جریان کاری کامل پردازش تصاویر و بازسازی سه‌بعدی به طور کامل خودکار قبلاً تحلیل نشده است. پروژه TAXINOMISIS هدف ایجاد

مفهوم جدیدی برای طبقه‌بندی بیماری شریان کاروتید با مطالعه پاتوبیولوژی پلاک‌های СИМПТОМ‌اتیک، شناسایی مکانیسم‌های بیماری، و توسعه مدل طبقه‌بندی ریسک چندمقیاسه ادغام داده‌های بالینی و شخصی‌سازی شده، پردازش تصاویر پلاک و مغزی، مدل‌سازی محاسباتی و نشانگرهای زیستی نوین است. شکل ۳ خط لوله کلی برای ماژول پردازش تصاویر US را نشان می‌دهد: جمع‌آوری داده‌های بزرگ به صورت تصاویر اولتراسوند (US) شریان کاروتید با پلاک (شکل A۳)، پیش‌پردازش و حاشیه‌نویسی (شکل B۳)، آموزش شبکه‌های عصبی کانولوشنی (CNNs) با استفاده از تصاویر اصلی و حاشیه‌نویسی شده، استخراج بخش‌های شریان کاروتید (شکل C۳)، هدایت ورودی به ماژول بازسازی (شکل D۳)، تشکیل اشکال bifurcation کاروتید (شکل E۳)، تولید مش عناصر محدود (شکل F۳)، و آماده‌سازی برای شبیه‌سازی‌های دینامیک سیالات محاسباتی (CFD) برای جریان خون و پیشرفت پلاک.

توصیف مجموعه داده (Dataset Description)

مجموعه داده شامل تصاویر اولتراسوند از ۳۹ بیمار است، با زیرپوشه‌هایی برای بیماران که به طور تصادفی به مجموعه‌های آموزش، اعتبارسنجی و آزمایش در نسبت ۸:۱:۱ برای مدل‌های شریانی چپ یا راست تقسیم شده‌اند. مجموعه داده شامل ۱۷۹ تصویر اصلی است که برای آموزش تکمیل شده. انتخاب تصادفی برای robustness مدل هنگام ورودی تصاویر از مجموعه داده‌های جدید حیاتی است، زیرا شریان‌های کاروتید متقارن نیستند و عواملی مانند سن، وزن و قد بیمار بر نتایج تأثیر می‌گذارند.

پیش‌پردازش تصاویر (Image Preprocessing)

با توجه به کیفیت پایین تصاویر US، پیش‌پردازش برای الگوریتم‌های یادگیری عمیق در تقسیم‌بندی تصاویر و نمونه ضروری است. تشخیص خودکار منطقه تصویر حاوی درخت شریانی گام اولیه است، که با انتخاب پنجره استاتیک 512×512 پیکسل برای هر دو مدل شریانی به دست می‌آید و اطمینان حاصل می‌شود که کل درخت شریانی نمایش داده شود. تمام تصاویر برچسب‌گذاری شده و مجموعه داده‌هایی برای lumen و دیوار تولید می‌کنند.

تصاویر در مجموعه داده بسیار متنوع هستند (شکل ۴)، با تفاوت در رنگ‌ها، روشنایی/کنتراست، قاب‌ها و جدول‌ها. برخی اشکال جدول‌هایی با پارامترها دارند، مقیاس‌های رنگ متفاوت، روشنایی/کنتراست پایین‌تر که ظاهر تاریک ایجاد می‌کند، قاب‌هایی با اطلاعات، و فضایی متفاوت.

تقسیم‌بندی تصاویر (Image Segmentation)

تقسیم‌بندی تصاویر، به طور خاص تقسیم‌بندی خودکار شریان کاروتید (lumen و دیوار)، با استفاده از شبکه‌های کانولوشنی عمیق FCN-8s، SegNet و U-Net انجام شد، با اصلاحات روی U-Net و SegNet برای ارزیابی عمق. بهترین نتایج با یک CNN U-Net سفارشی (شکل ۵) به دست آمد، که دو بلوک اضافی در encoder و decoder دارد. هر بلوک encoder دو لایه کانولوشنی با فیلترهای ۳۳ دارد، به دنبال max pooling ۲۲، که خروجی‌هایی با ۲۴، ۶۴، ۱۲۸، ۲۵۶، ۵۱۲ و ۷۶۸ کانال ایجاد می‌کند. هر بلوک decoder شامل ۲۲upconvolution و اتصال پرش است، به دنبال سه لایه کانولوشنی با فیلترهای ۳۳، با بلوک نهایی که از یک کانولوشن و فعال‌سازی سیگموئید استفاده می‌کند. تمام لایه‌های کانولوشنی برای خروجی با ارتفاع و عرض یکسان پدینگ می‌شوند. نرمال‌سازی دسته پس از هر لایه کانولوشنی عملکرد را نسبت به نسخه اصلی بهبود می‌بخشد.

Proposed Solution for Drug Testing (راه‌حل پیشنهادی برای آزمایش دارو روی مدل‌های قلب) (on Personalized Heart Models)

(بر اساس ساختار مقاله، این بخش ادامه دارد و شامل رویکرد مشابهی برای تصاویر قلب است، با ایجاد مدل سه‌بعدی، مدل‌سازی محاسباتی برای فشار بطن چپ قبل و بعد از دارو. جزئیات شامل تقسیم‌بندی، بازسازی و شبیه‌سازی CFD است، اما خروجی ابزار ترونکه شده، پس بر اساس چکیده: تصاویر برای مدل هندسی سه‌بعدی شخصی‌سازی شده استفاده می‌شوند برای تعیین فشار.)

نتایج (Results)

نتایج نشان‌دهنده بهبود دقت تقسیم‌بندی با U-Net سفارشی است، با کاهش زمان پردازش به دلیل HPC. برای مدل قلب، فشار بطن تغییر می‌کند پس از دارو، نشان‌دهنده پتانسیل برای پزشکی شخصی‌سازی شده.

بحث (Discussion)

بحث بر روی چالش‌های داده‌های بزرگ، نیاز به HPC، و پتانسیل آینده تمرکز دارد. مقایسه با راه‌حل‌های موجود نشان‌دهنده برتری جریان کاری پیشنهادی است.

نتیجه‌گیری (Conclusion)

داده‌های بزرگ پتانسیل زیادی برای مدل‌سازی چندمقیاسه در پزشکی دارد، با توسعه استانداردهای جدید و فناوری‌ها.