طراحی و پیاده‌سازی فرم الکترونیکی ساختارمند برای گزارش‌های پاتولوژی بیماری سلیاک: رویکرد متن‌کاوی*

آزاده کامل قالیباف[1]، فرزانه خادم ثامنی[2]، مجید جنگی[1]، محمدرضا مظاهری حبیبی[1]، کبری اطمینانی[3]

مقاله پژوهشی

چکیده

مقدمه: گزارش پاتولوژی به صورت متن باز تهیه می‌شود و شامل شبکه‌ای از روابط بین مفاهیم پزشکی است که پزشک از آن برای استدلال و تشخیص استفاده می‌کند. این مطالعه با هدف، طراحی و ارزیابی مدلی جهت استخراج خودکار این مفاهیم و تبدیل آن به فرم ساختار یافته و قابل تحلیل توسط کامپیوتر انجام شد.

روش بررسی: تحقیق حاضر از نوع کاربردی و اجرایی بود و بر روی ۲۵۸ گزارش پاتولوژی با تشخیص بیماری سلیاک که به صورت تصادفی از دو آزمایشگاه پاتوبیولوژی جمع‌آوری شد، صورت گرفت. سیستم پیشنهاد شده شامل سه فاز اصلی بود. فاز اول به طراحی یک فرم استاندارد و ساختارمند برای گزارش بیوپسی بیماری سلیاک با استفاده از روش Delphi ارتباط داشت. در فاز دوم با به کارگیری ابزارهای متن‌کاوی ارایه شده توسط مرکز زبان‌شناسی دانشگاه استنفورد و برنامه واسط طراحی شده به منظور تفسیر قطعات معنایی، اطلاعات مورد نظر از متن گزارش استخراج و در قالب فرم استاندارد ذخیره گردید. در فاز سوم، کلاس Marsh مربوط به هر گزارش با استفاده از الگوریتم یادگیری درخت تصمیم J۴۸، به صورت خودکار تعیین شد.

یافته‌ها: عملکرد سیستم در فاز استخراج اطلاعات و انتساب مقادیر به فیلدهای فرم استاندارد، صحت ۷۶ درصدی را نشان داد. صحت سیستم در تعیین خودکار طبقه‌بندی Marsh بر اساس خروجی مرحله قبل، ۶۲ درصد به دست آمد که در صورت ارایه داده‌های تصحیح شده و بدون خطا، صحت الگوریتم دسته‌بندی تا ۸۴ درصد افزایش می‌یابد.

نتیجه‌گیری: در مطالعه حاضر با طراحی و پیاده‌سازی مدلی برای ساختارمند کردن گزارش‌های پاتولوژی بیماری سلیاک، علاوه بر تسهیل و تسریع در ورود و بازیابی اطلاعات و افزایش خوانایی گزارش، امکان پردازش کامپیوتری داده‌ها و پیدا کردن روابط و الگوها نیز میسر گردید.

واژه‌های کلیدی: متن‌کاوی؛ بیماری سلیاک؛ سیستم پشتیبان تصمیم بالینی؛ روش Delphi؛ درخت تصمیم

دریافت مقاله: ۱۳۹۳/۱۲/۲۳ اصلاح نهایی: ۱۳۹۴/۱۲/۱۵ پذیرش مقاله: ۱۳۹۵/۰۱/۱۵

ارجاع: کامل قالیباف آزاده، خادم ثامنی فرزانه، جنگی مجید، مظاهری حبیبی محمدرضا، اطمینانی کبری. **طراحی و پیاده‌سازی فرم الکترونیکی ساختارمند برای گزارش‌های پاتولوژی بیماری سلیاک: رویکرد متن‌کاوی.** مدیریت اطلاعات سلامت ۱۳۹۵؛ ۱۳ (۱): ۲۷-۱۹

مقدمه

بیماری سلیاک CD (Celiac disease) یک بیماری خود ایمنی است که از مشخصات آن آسیب بافت مخاطی روده کوچک، به دنبال مصرف غذاهای حاوی گلوتن می‌باشد (۱). اکثر بیماران مبتلا به سلیاک، یا به طور کامل بدون علامت هستند و یا علایم گوارشی غیر اختصاصی مانند سوء هاضمه، درد شکمی، نفخ و اختلال در حرکات روده را نشان می‌دهند که همین امر تشخیص این بیماری را مشکل می‌سازد (۲). تشخیص بیماری سلیاک تنها بر اساس علایم بالینی امکان‌پذیر نیست و اغلب ترکیبی از علایم بالینی، نتایج آزمایش‌های سرولوژیک و بیوپسی (نمونه‌برداری از بافت) از روده کوچک، در کنار هم به پزشک در رسیدن به تشخیص درست کمک می‌کنند.

به دنبال تأیید بیماری سلیاک توسط نتایج آزمایشگاهی و علایم موجود، به منظور تشخیص نهایی و تعیین میزان آسیب به پرزهای روده، بخش کوچکی از روده کوچک از طریق نمونه‌برداری مورد بررسی قرار می‌گیرد. نمونه‌برداری به عنوان معیار طلایی در تشخیص بیماری سلیاک در نظر گرفته می‌شود که طی آن پزشک پاتولوژیست مشاهدات خود از خصوصیات بافت بیوپسی را در قالب یک متن تهیه و گزارش می‌کند (۱). در سال ۱۹۹۲ یک پزشک انگلیسی به نام Marsh یک سیستم طبقه‌بندی چهار سطحی، برای استانداردسازی میزان آسیب به بافت روده معرفی نمود (۳). چند سال بعد، فردی به نام Oberhuber با ایجاد اصلاحاتی در طبقه‌بندی Marsh، آن را در قالب سه سطح با عناوین Marsh I، Marsh II و Marsh III تقسیم‌بندی نمود که Marsh III خود سه زیررده a، b و c را شامل می‌شود (۴). این دسته‌بندی در حال حاضر به عنوان مبنای تشخیص در متون تخصصی مرجع رشته پاتولوژی ذکر شده است و به صورت معمول در گزارش‌ها استفاده می‌شود (۷-۵).

گزارش پاتولوژی به صورت متن باز (Free text) است که یافته‌های مشاهده شده از سلول‌های بافت را شرح می‌دهد. متن گزارش شامل شبکه‌ای از روابط بین مفاهیم پزشکی است که پزشک برای استدلال و تشخیص از آن استفاده می‌کند. اگر از کامپیوتر برای تحلیل این گزارش‌ها کمک گرفته شود،

* این مقاله حاصل تحقیق مستقل بدون حمایت مالی و سازمانی است.

۱- دانشجوی دکتری، انفورماتیک پزشکی، گروه انفورماتیک پزشکی، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران

۲- استادیار، متخصص پاتولوژی، دانشکده پزشکی، دانشگاه آزاد اسلامی، واحد زاهدان، ایران

۳- استادیار، انفورماتیک پزشکی، گروه انفورماتیک پزشکی، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران(نویسنده مسؤول)

Email: etminanik@mums.ac.ir

باید روابط بین مفاهیم در قالب یک فرم ساخت یافته و استاندارد ارایه گردد (۸).

تاکنون سیستم‌های مختلفی برای تحلیل خودکار متون پاتولوژی از روش‌های پردازش متن استفاده کرده‌اند که از آن جمله می‌توان به مطالعه انجام شده توسط MCCOWAN و همکاران (۹) اشاره کرد که سیستمی برای تعیین خودکار مرحله سرطان ریه طراحی نمودند. این سیستم با استفاده از روش‌های پردازش زبان طبیعی، متن گزارش را با عبارات استاندارد اصطلاحنامه UMLS (Unified Medical Language System) جایگزین کرده، سپس با روش وزن‌دهی LTC (Log TF-IDF cosine) اطلاعات متن را به یک بردار عددی تبدیل می‌کند. این بردار که فرم فشرده شده متن گزارش است، به عنوان ورودی به الگوریتم دسته‌بندی کننده SVM (Support vector machine) داده می‌شود تا مرحله پیشرفت سرطان در آن مشخص گردد. صحت عملکرد کلی این سیستم ۷۴ درصد گزارش شده است. در پژوهش‌ها نیز به طور مشابه از روش‌های پردازش متن برای تحلیل گزارش‌های پاتولوژی سرطان ریه استفاده نموده‌اند (۱۱، ۱۰).

مطالعه دیگری که روی ۱۰۳۸ نمونه از گزارش‌های پاتولوژی سرطان لنفوم در بیمارستان عمومی ماساچوست انجام شد (۱۲)، نشان داد که طبقه‌بندی خودکار نوع لنفوم بر اساس گزارش پاتولوژی، در مدلی که جملات متن در قالب گراف وابستگی تجزیه شود و ارتباطات بین اجزای جمله تعیین گردد، با نرخ صحت ۸۵ درصد بهترین عملکرد را نسبت به سایر مجموعه ویژگی‌ها خواهد داشت. همچنین، در مطالعه‌ای دیگر که در فرانسه صورت گرفت (۱۳)، محققان تعداد ۵۱۲۱ گزارش پاتولوژی متن باز، مربوط به ۳۵ پاتولوژیست را با استفاده از روش‌های یادگیری بردار پشتیبان تصمیم و طبقه‌بندی کننده ساده Bayes (Naive Bayes) و بر اساس اندام درگیر سرطان، دسته‌بندی کردند. مبنای طبقه‌بندی گزارش‌ها در این مقاله، روش معرفی شده توسط آژانس بین‌المللی تحقیقات روی سرطان بوده است که نتایج به دست آمده ۹۶ درصد را برای معیار F1 نشان داده است. در این مطالعه به منظور بازنمایی متن گزارش، جهت ارایه به الگوریتم‌های دسته‌بندی کننده، از روش فراوانی وزنی TF-IDF (Term frequency inverse document frequency) استفاده شد که متن را به یک بردار عددی، از نسبت تکرار کلمات تشکیل دهنده آن تبدیل می‌کند.

علاوه بر موارد فوق تعداد زیادی سیستم پردازش زبان طبیعی پزشکی با هدف، شناسایی بیماران با مشخصات بالینی خاص به منظور شرکت در مطالعات هم گروه انجام شده است (۱۷-۱۴).

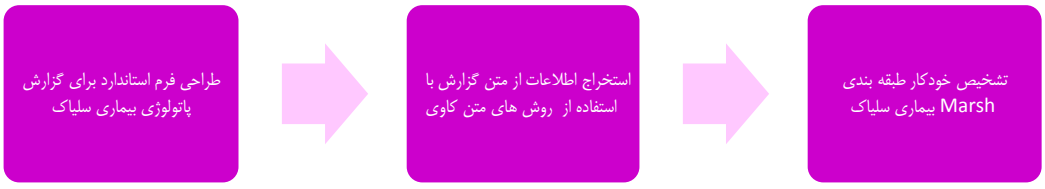
همان طور که قبل اشاره شد، علایم متعدد و گاهی نامرتبط در بیماری سلیاک، تشخیص این بیماری را برای پزشک دشوار کرده است و پزشکان اغلب با درخواست آزمایشات مختلف، به گردآوری اطلاعات تکمیلی می‌پردازند و در نهایت، از کنار هم قرار دادن این مجموعه اطلاعات، به تشخیص درست خواهند رسید. ضرورت ساختارمند کردن فرم گزارش پاتولوژی بیماری سلیاک از آن جهت حایز اهمیت است که استدلال بر روی این حجم از اطلاعات متنوع، در حالی که بدون فرمت و نظم مشخص باشد، علاوه بر زمان‌بر بودن و تحمیل بار فکری و کار شناختی مضاعف، احتمال بروز خطا در تشخیص را نیز افزایش می‌دهد. یکی از استانداردهای اعلام شده توسط کمیسیون سرطان کالج جراحان آمریکا برای اعتباربخشی آزمایشگاه‌ها، استفاده از عناصر داده‌ای معتبر علمی است که پاتولوژیست‌ها را ملزم می‌کند با استفاده از مجموعه‌ای از داده‌های ساختارمند (Structured data)، گزارش‌های اختصاری (Synoptic reporting) تهیه نمایند (۱۸). طی بررسی‌های انجام شده، تاکنون هیچ مطالعه‌ای از روش‌های متن‌کاوی برای تحلیل و استانداردسازی گزارش‌های پاتولوژی بیماری سلیاک استفاده نکرده است. هدف از مطالعه حاضر، طراحی یک فرم استاندارد برای گزارش‌های پاتولوژی بیماری سلیاک، به منظور اطمینان از ثبت اطلاعات ضروری توسط پاتولوژیست و بازیابی سریع‌تر و دقیق‌تر اطلاعات از متن گزارش، توسط پزشک و اتخاذ تصمیم درست بود. سپس، جهت ارزیابی کارایی فرم طراحی شده، طبقه Marsh گزارش و به صورت خودکار پیش‌بینی گردید.

روش بررسی

این تحقیق از نوع کاربردی و اجرایی بود. پیکره متنی مورد مطالعه شامل ۲۵۸ گزارش پاتولوژی بیوپسی دئودنوم، مربوط به دو پزشک پاتولوژیست از دو آزمایشگاه مختلف در شهرهای زاهدان و مشهد بود که تشخیص نهایی آن‌ها بیماری سلیاک بوده است. گزارش‌ها بین سال‌های ۱۳۸۸ تا ۱۳۹۳، به صورت متن باز و به زبان انگلیسی نوشته شده‌اند. طول گزارش‌ها به طور متوسط یک صفحه بوده است و شامل سه بخش شرح ماکروسکوپی، میکروسکوپی و تشخیص نهایی است. بخش ماکروسکوپی شامل مشخصات ظاهری نمونه از قبیل اندازه، وزن، رنگ و تغییرات ظاهری است که به صورت چشمی مشاهده می‌شود. قسمت میکروسکوپی به شرح مشخصات سلول‌ها و بافت در زیر میکروسکوپ می‌پردازد که مبنای تشخیص و تعیین درجه آسیب است. به منظور رعایت اصول اخلاقی مشخصات هویتی افراد در ابتدای کار حذف گردید.

سیستم پیشنهاد شده در این مقاله شامل سه فاز اصلی بود که در شکل ۱ نشان داده شده است. در فاز اول با روش Delphi یک فرم استاندارد و ساختارمند برای گزارش پاتولوژی بیماری سلیاک تهیه شد.

روش Delphi فرایندی ساختار یافته برای جمع‌آوری و طبقه‌بندی دانش موجود در نزد گروهی از کارشناسان و خبرگان است که از طریق توزیع پرسش‌نامه‌هایی در بین این افراد و بازخورد کنترل شده پاسخ‌ها و نظرات دریافتی صورت می‌گیرد (۱۹). اعتبار روش Delphi نه به شمار شرکت کنندگان در پژوهش که به اعتبار علمی متخصصان شرکت کننده بستگی دارد. شرکت کنندگان در تحقیق Delphi از ۵ تا ۲۰ نفر را شامل می‌شوند.



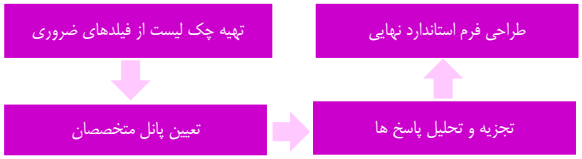
شکل ۱: مدل پیشنهادی برای خودکارسازی فرایند تشخیص گزارش پاتولوژی بیماری سلیاک

فاز دوم، اطلاعات مربوطه را از متن گزارش استخراج نموده، در قالب فرم استاندارد تهیه شده از مرحله قبل، ذخیره می‌نماید که برای این منظور از تکنیک‌های مختلف پردازش زبان طبیعی و متن‌کاوی استفاده شده است. پردازش زبان طبیعی کاربردهای بالقوه متعددی در حوزه مراقبت بهداشتی و مطالعات پژوهشی دارد. با وجود آن که بسیاری از اطلاعات بیماران از طریق پرونده‌های الکترونیک قابل بازیابی است، ولی بخشی از اطلاعات که به شکل متن باز ذخیره می‌شود، مانند گزارش پرستاری، خلاصه پرونده و گزارش‌های رادیولوژی و پاتولوژی به طور مستقیم قابل دسترس نیستند (۲۰). برای حل این محدودیت باید تکنیک‌هایی طراحی شود که بتوان اطلاعات موجود در متن را سازمان‌دهی و استخراج نمود. روش‌های پردازش زبان طبیعی با استخراج اطلاعات مرتبط در زمان مناسب، به مدیریت حجم بزرگی از متون مثل گزارش‌های بیمار کمک می‌کند (۲۱).

در نهایت برای هر گزارش، کلاس Marsh مربوط به آن با استفاده از روش‌های یادگیری ماشین و به صورت خودکار تعیین می‌گردد. یادگیری ماشین یکی از شاخه‌های پرکاربرد هوش مصنوعی است که در آن روش‌هایی برای تعلیم و یادگیری کامپیوتر ارایه می‌شود. یکی از این روش‌ها، یادگیری با سرپرستی است که در آن مجموعه‌ای از جفت‌های «ورودی- خروجی» جهت آموزش، به سیستم ارایه می‌گردد و سیستم تلاش می‌کند تا تابعی از ورودی به خروجی را فرا گیرد. در ادامه به تفصیل جزئیات پیاده‌سازی هر فاز توضیح داده شده است.

فاز اول: طراحی فرم استاندارد

مراحل طراحی فرم ساخت یافته گزارش پاتولوژی بیماری سلیاک، در شکل ۲ نشان داده شده است. در مرحله اول با بررسی متن گزارش‌های موجود و با همکاری یک نفر پزشک پاتولوژیست، لیست فیلدهای ضروری استخراج شد و در قالب یک چک‌لیست آماده گردید. سپس، لیست تهیه شده در اختیار سه نفر متخصص پاتولوژی قرار گرفت تا در مورد ضرورت وجود یا عدم وجود هر گزینه اظهار نظر نمایند. همچنین، متخصصان می‌توانستند در مورد اضافه نمودن گزینه جدیدی خارج از لیست و یا فرمت ورود اطلاعات (چک‌باکس، لیست کشویی و...)پیشنهاد دهند. پس از جمع‌آوری و تجزیه و تحلیل پاسخ‌ها، موارد با حداکثر توافق یعنی با بیش از دو رأی مثبت، در قالب یک فرم استاندارد طراحی گردید. یک نمونه متن گزارش پاتولوژی بیماری سلیاک و همچنین، نمونه فرم استاندارد طراحی شده در پیوست ۱ قابل مشاهده است.

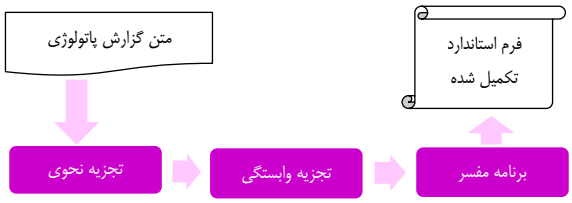


شکل ۲: طراحی فرم استاندارد گزارش پاتولوژی سلیاک

ایجاد یک فرمت استاندارد به پزشک یا کامپیوتر این امکان را می‌دهد که بتواند نوع خاصی از داده‌ها را جستجو کند یا بداند که یک عنصر اطلاعاتی خاص به کدام گروه اطلاعاتی تعلق دارد و در نتیجه امکان بازیابی، انتقال و تفسیر ساده‌تر داده‌ها را فراهم می‌کند (۲۲).

فاز دوم: استخراج اطلاعات از گزارش‌های متن باز

در این فاز سعی شده است با استفاده از مجموعه‌ای از روش‌ها و ابزارهای متن‌کاوی، اطلاعات موجود در متن گزارش‌ها استخراج و در قالب استاندارد تهیه شود و در فاز اول سازمان‌دهی گردد. نمای کلی مدل در شکل ۳ مشخص شده است که در ادامه به شرح بیشتر در مورد آن پرداخته خواهد شد.



شکل ۳: مراحل استخراج اطلاعات از متن گزارش

هدف اصلی از پردازش متن، تبدیل آن به فرمی است که برای کامپیوتر قابل درک و تحلیل‌پذیر باشد. برای این منظور از تئوری‌های محاسباتی، الگوریتم‌ها و ساختارهای داده‌ای موجود در علوم کامپیوتر بهره گرفته شد. اغلب اولین مرحله از متن‌کاوی، پیش‌پردازش است که طی آن اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی ذخیره می‌شود. از جمله کارهایی که در این مرحله انجام می‌شود، شناسایی محدوده کلمات، تعیین مرز جمله و تعیین نقش واژه‌ها POS (Part of speech) می‌باشد.

در تجزیه نحوی، جملات متن به سازه‌های نحوی تشکیل دهنده آن مانند گروه اسمی (NP)، گروه فعلی (VP)، گروه صفتی (ADJP)، گروه قید (ADVP) و... تقسیم می‌شوند. همچنین، تجزیه‌گر نحوی (Syntactic parser) در یک سطح پایین‌تر به هر واژه یک تگ، متناسب با نقش آن در جمله تخصیص می‌دهد (POS). برخی تگ‌های اشاره شده در شکل شامل حرف اضافه (DT)، اسم (NN)، صفت (JJ) و... می‌باشد. لیست کامل تگ‌های POS و نقش دستوری مربوط به آن در پژوهش Santorini (۲۳) آمده است. دستور وابستگی یکی از مباحث رشته زبان‌شناسی است و بر مبنای نظریه ظرفیت واژگانی شکل گرفته است. این نظریه بیان می‌کند که هر واژه بر اساس ظرفیت و محل قرار گرفتن آن در جمله، وابسته‌هایی دارد. تجزیه‌گر وابستگی این روابط نحوی/ معنایی بین واژه‌های درون جمله را مشخص می‌کند.

به کمک تگ‌های POS، گروه‌های نحوی و توابع وابستگی در کنار یکدیگر می‌توان به اطلاعات کاملی از ساختار معنایی جمله دست یافت. پس از تجزیه متن به گروه‌های نحوی و روابط وابستگی، باید قطعات مناسب جهت تکمیل فیلدهای فرم استاندارد مشخص شود و در محل مربوط به خود قرار گیرد. به همین منظور یک برنامه مفسر واسط با زبان برنامه‌نویسی Visual basic و در محیط Visual studio 2013 طراحی گردید. این برنامه با استفاده از اطلاعات تگ‌های POS، گروه‌های نحوی و توابع وابستگی، معنای متن را تحلیل می‌کند و فرم استاندارد را با اطلاعات مناسب تکمیل و در خروجی ارایه می‌دهد. با توجه به این که مقادیر مربوط به فیلدها به طور عمومی در ساختار جمله در نقش صفت یا قید ظاهر می‌شوند، برنامه ابتدا کلید واژه عنوان هر فیلد را در تجزیه وابستگی متن جستجو می‌کند و کلیه روابط قیدی- توصیفی که این کلید واژه در جایگاه هسته آمده است را استخراج می‌نماید. اگر یافته‌ها بیش از یک مورد

باشد، برای تشخیص مرتبط‌ترین زوج در تجزیه نحوی به دنبال گروه اسمی، صفتی یا قیدی می‌گردیم که هر دو واژه هسته و وابسته در آن وجود داشته باشد. در این حالت وابسته به عنوان مقدار توصیف‌گر، در مقابل فیلد مربوطه در فرم وارد می‌شود. وجود خطاهای تایپی در گزارش و همچنین، تنوع ساختاری جملات را می‌توان از جمله علل خطا در این سیستم برشمرد. همچنین، انتظار می‌رود که با به کارگیری تکنیک‌های پیشرفته‌تر متن‌کاوی مانند تحلیل محتوایی و شبکه‌های معنایی درصد صحت بالاتری در این مرحله به دست آید.

فاز سوم: تعیین طبقه Marsh

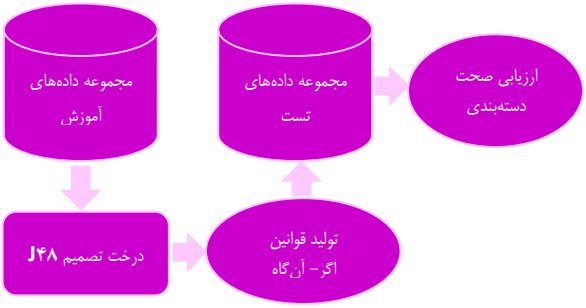
پس از این که اطلاعات متن به قالب فرم استاندارد منتقل گردید، در این فاز به منظور تسهیل کار پاتولوژیست در فرایند تهیه گزارش، طبقه Marsh مناسب برای مشخصات ذکر شده در فرم، به صورت خودکار توسط سیستم تعیین می‌شود. طبقه Marsh بر اساس سیستم اصلاح شده Oberhuber که در کتب مرجع پاتولوژی آمده است (۵)، بر مبنای سه ویژگی تعداد لنفوسیت‌های اینتراپیتلیال، کریپت هیپرپلازی و میزان تحلیل پرزهای روده، طبق جدول ۱ مشخص می‌گردد.

جدول ۱: طبقه‌بندی اصلاح شده Marsh

Marsh type	ILE*	Crypts	Villi
۰	۴۰<	Normal	Normal
۱	۴۰>	Normal	Normal
۲	۴۰>	Increased	Normal
a۳	۴۰>	Increased	Mild atrophy
b۳	۴۰>	Increased	Moderate
c۳	۴۰>	Increased	Severe

*Intraepithelial lymphocytes (per 100 Enterocytes)

الگوریتم یادگیری که در این مطالعه استفاده شده است، درخت تصمیم J۴۸ است و مقادیر ورودی، مجموعه ویژگی‌های مشخص شده در جدول ۱ می‌باشد و کلاس خروجی، طبقه Marsh مناسب خواهد بود. فرایند کار به این شکل است که ابتدا مجموعه داده‌هایی که طبقه Marsh آن‌ها به صورت صحیح توسط فرد متخصص تعیین شده است، به عنوان مجموعه داده آموزش به الگوریتم داده می‌شود. سپس، درخت تصمیم با ایجاد مجموعه‌ای از قواعد اگر- آن‌گاه، در مورد داده جدیدی که طبقه Marsh آن مشخص نیست، تصمیم‌گیری می‌نماید. روال و مراحل کار در شکل ۴ نشان داده شده است.



شکل ۴: تخصیص خودکار دسته Marsh به هر گزارش

یافته‌ها

از ۲۵۸ گزارش موجود در پیکره متنی، ۱۸۶ مورد متعلق به کلاس Marsh I، ۱۴ نمونه Marsh II و در مجموع ۵۸ مورد مربوط به کلاس Marsh III می‌باشند. همچنین، از نظر جنسیتی ۱۵۹ نفر از بیماران زن و ۹۹ مورد مرد بوده‌اند. در این مطالعه به منظور تجزیه متن گزارش‌ها از ابزارهای ارایه شده توسط مرکز زبان‌شناسی دانشگاه استنفورد استفاده شده است (۲۴).

در شکل ۵ (قسمت الف)، تجزیه نحوی برای جمله نمونه (The lamina propria is expanded with numerous lymphocytes and plasma cells and the crypts are mildly hyperplastic) نشان داده شده است و قسمت ب تجزیه وابستگی دستوری مربوط به جمله را نشان می‌دهد. همان طور که در شکل مشخص است خروجی این پارسر به شکل زوج‌های دوتایی «هسته- وابسته» است که نوع وابستگی به صورت تابعی از این زوج بیان می‌شود (اعداد کنار لغات شماره آن واژه در جمله اصلی است). به عنوان مثال رابطه بین کلمات "Hyperplastic" و "Mildly" در جمله با تابع وابستگی توصیف‌گر قیدی (advmod) مشخص شده است که نشان می‌دهد هایپرپلازی یا تکثیر کریپت‌ها متعادل بوده است.

الف) تجزیه نحوی

```
(ROOT
  (S
    (S
      (NP (DT the) (NN lamina) (NN propria))
      (VP (VBZ is)
        (VP (VBN expanded)
          (PP (IN with)
            (NP (JJ numerous) (NNS lymphocytes)
              (CC and)
              (NN plasma) (NNS cells))))))
    (CC and)
    (S
      (NP (DT the) (NNS crypts))
      (VP (VBP are)
        (ADJP (RB mildly) (JJ hyperplastic))))
    (. .)))
```

ب) تجزیه وابستگی

```
det (propria-3, the-1)
nn (propria-3, lamina-2)
nsubjpass (expanded-5, propria-3)
auxpass (expanded-5, is-4)
root (ROOT-0, expanded-5)
amod (lymphocytes-8, numerous-7)
prep_with (expanded-5, lymphocytes-8)
nn (cells-11, plasma-10)
prep_with (expanded-5, cells-11)
conj_and(lymphocytes-8, cells-11)
det (crypts-14, the-13)
nsubj (hyperplastic-17, crypts-14)
cop (hyperplastic-17, are-15)
advmod (hyperplastic-17, mildly-16)
conj_and (expanded-5, hyperplastic-17)
```

شکل ۵: خروجی تجزیه نحوی و تجزیه وابستگی برای یک جمله نمونه

به منظور ارزیابی صحت عملکرد برنامه مفسر، تعداد ۵۴ نمونه گزارش به صورت تصادفی به برنامه داده شد و از کل ۴۳۲ فیلد ورودی، سیستم ۳۲۷ مورد را به درستی مقداردهی کرده است که در کل نرخ درستی، ۷۶ درصد را نشان می‌دهد. با توجه به حجم نمونه انتخاب شده بر اساس مطالعه‌ای (۲۵)، می‌توان نتیجه را با سطح اطمینان ۹۰ درصد برای کل داده‌ها تعمیم داد.

شکل ۶ (قسمت الف) رسم گرافیکی درخت تصمیم و قسمت ب نمونه‌ای از قواعد تولید شده در درخت را نشان می‌دهد که اعداد داخل پرانتز برای هر کلاس، به ترتیب نشان دهنده تعداد کل نمونه‌هایی است که به این گره برگ رسیده‌اند و عدد دوم بیانگر تعداد نمونه‌هایی است که به اشتباه دسته‌بندی شده‌اند. جهت پیاده‌سازی الگوریتم درخت تصمیم J۴۸ از نرم‌افزار Weka نسخه ۳.۶.۱۱ استفاده شد که یک نرم‌افزار داده‌کاوی متن باز می‌باشد و بسیاری از الگوریتم‌های یادگیری ماشین را پشتیبانی می‌کند (۲۶).

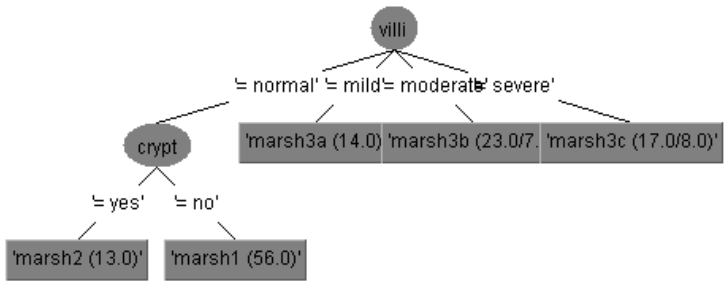
نتایج مربوط به طبقه‌بندی خودکار کلاس Marsh در دو حالت گزارش شده است. یکی در حالتی که نتایج به دست آمده از فازهای قبلی به عنوان ورودی به سیستم داده شود، در این حالت خطای مراحل قبل روی نتایج این مرحله تأثیر می‌گذارد. در حالت دیگر ارزیابی، ورودی سیستم عاری از خطا و تکمیل شده توسط فرد خبره است که در این صورت کارایی این فاز به صورت مستقل مورد سنجش قرار می‌گیرد. صحت عملکرد سیستم برای حالت اول ۶۲ درصد، و در حالت دوم ۸۴ درصد به دست آمده است. جدول ۲ جزئیات نتایج مربوط به اجرای سیستم در حالت دوم به همراه مقادیر مربوط به سایر معیارهای ارزیابی مانند معیار دقت (Precision) و بازخوانی (Recall) و معیار F را نشان می‌دهد. معیار نرخ صحت TP (True positive) در جدول ۲، بیانگر نسبت تشخیص‌های درست به کل تشخیص‌ها است که همین نسبت برای تشخیص‌های نادرست با FP (False positive) مشخص شده است. از مقایسه مقادیر این دو معیار برای کلاس‌های Marsh مختلف مشاهده می‌شود که در مورد کلاس Marsh ۱ و ۲ که در مجموع ۷۷ درصد کل نمونه‌ها را شامل می‌شوند، به طور تقریبی تمام موارد به درستی توسط سیستم تشخیص داده شده است. به این ترتیب می‌توان انتظار داشت چنان چه تعداد نمونه‌های کلاس Marsh ۳ را افزایش دهیم، عملکرد سیستم در تشخیص این موارد نیز بهبود یابد.

نحوه تقسیم‌بندی مجموعه‌های آموزش و تست برای الگوریتم یادگیری با روش K-fold cross-validation بوده است که در این روش مجموعه داده‌های اولیه به ۱۰ قسمت مساوی تقسیم می‌گردد. سپس، در هر بار اجرای دسته‌بندی، یکی از قسمت‌ها برای مرحله تست انتخاب شده، سایر قسمت‌ها در مرحله آموزش استفاده می‌شوند. این فرایند ۱۰ بار تکرار می‌شود؛ به طوری که در نهایت هر رکورد داده به طور دقیق یک بار در مرحله تست استفاده شده باشد.

بحث

مطالعات انجام شده در رابطه با به کارگیری روش‌های کامپیوتری در حوزه پاتولوژی را می‌توان به دو دسته مطالعات پردازش تصاویر پاتولوژی (۲۷) و مطالعات پردازش متن گزارش‌ها تقسیم کرد که مطالعات صورت گرفته بر پردازش متن گزارش به نسبت تعداد کمتری را شامل می‌شوند. این دسته مطالعات اغلب از روش‌های مبتنی بر آمار جهت تحلیل گزارش‌ها استفاده کرده‌اند که قادر به بازنمایی روابط و مفاهیم عمیق در متن نیستند.

الف) رسم گرافیکی درخت تصمیم



ب) بخشی از قوانین درخت تصمیم

```
villi = normal
|   crypt = yes: marsh2 (13.0)
|   crypt = no: marsh1 (56.0)
villi = mild: marsh3a (14.0)
villi = moderate: marsh3b (23.0/7.0)
villi = severe: marsh3c (17.0/8.0)
```

شکل ۶: الگوریتم یادگیری درخت تصمیم J۴۸

جدول ۲: نتایج مربوط به طبقه‌بندی خودکار کلاس Marsh برای ورودی فاقد خطا

Class	ROC Area	F-Measure	Recall	Precision	FP Rate	TP Rate
Marsh۱	۱	۱	۱	۱	۰	۱
Marsh۲	۱	۱	۱	۱	۰	۱
Marsh۳a	۹۵۰/۰	۸۲۰/۰	۷۰۰/۰	۱	۰	۷۰۰/۰
Marsh۳b	۹۱۰/۰	۶۲۰/۰	۶۶۰/۰	۵۹۰/۰	۱۱۰/۰	۶۶۰/۰
Marsh۳c	۸۸۰/۰	۴۳۰/۰	۵۰۰/۰	۳۸۰/۰	۰۷۱/۰	۵۰۰/۰
Weighted avg	۹۶۰/۰	۸۵۰/۰	۸۴۰/۰	۸۷۰/۰	۰۲۷/۰	۸۴۰/۰

TP: True positive; FP: False positive; ROC: Receiver operating characteristic

Jouhet و همکاران (۱۳) از بردار فرکانس اصطلاحات در متن (TF-IDF) برای بازنمایی اطلاعات استفاده کرده‌اند که این روش کلمات و مفاهیم کلیدی متن را بر اساس فرکانس تکرار آن‌ها شناسایی می‌کند، ولی قادر به تشخیص روابط معنایی بین اصطلاحات نبوده است و تنها بر مبنای ویژگی‌های نحوی و آماری محتوای متن را مورد تحلیل و پردازش قرار می‌دهد. شناسایی گروه‌های نحوی و توابع وابستگی در پژوهش حاضر درک عمیق‌تری از ساختار معنایی و ارتباطات میان مفاهیم متن در اختیار قرار می‌دهد که امکان دستیابی به تحلیل‌های دقیق‌تر و صحیح‌تر از متن را میسر می‌سازد. مطالعه انجام شده توسط Li و Martinez (۲۸)، از عبارات با قاعده و مدل Bag-of-words (BOW) به منظور استخراج اطلاعات از متون گزارش‌های پاتولوژی بهره گرفته است. به کارگیری عبارات با قاعده برای شناسایی مفاهیم مورد نظر در متن قابلیت تعمیم‌پذیری سیستم را پایین می‌آورد و کاربرد مدل طراحی شده را به یک حوزه خاص محدود می‌نماید. در مقابل روش‌های زبان‌شناسی و تکنیک‌های هوش مصنوعی و پردازش متن به کار رفته در پژوهش حاضر وابسته به نوع و ساختار خاصی در متون نبوده است و می‌توان مدل پیشنهاد شده در این مقاله را به سایر حوزه‌ها گسترش داد.

مدل طراحی شده در این مقاله با هدف، معرفی چارچوبی برای ساختمند کردن گزارش‌های متن آزاد در پزشکی ارایه شده است که به طور خاص گزارش‌های پاتولوژی بیماری سلیاک را مورد توجه قرار داده است. نتایج به دست آمده عملکرد قابل قبول سیستم در تشخیص خودکار نتیجه گزارش را نشان می‌دهد. در مقایسه با سایر مقالات، پژوهش Luo و همکاران (۱۲) با گزارش دقت ۸۷ درصد برای دسته‌بندی گزارش‌های پاتولوژی سرطان لنفوم، نتایج مشابهی با مطالعه حاضر داشته است. این در حالی است که تعداد طبقات در نظر گرفته شده برای سرطان لنفوم در مقاله ذکر شده، سه کلاس بوده است که در مقایسه با مطالعه حاضر با پنج کلاس، پیچیدگی کمتری دارد. Nguyen و همکاران نیز با نرخ صحت ۹۵ درصد، مرحله TNM سرطان ریه را به صورت خودکار از متن گزارش پاتولوژی استخراج نمودند (۱۱). مجموعه داده آموزش به حجم ۱۱۰۲ رکورد را می‌توان یکی از دلایل اصلی برای دستیابی به درصد صحت بالا در این مقاله دانست. از این‌رو، انتظار می‌رود که با افزایش تعداد گزارش‌ها در مطالعه حاضر، صحت به دست آمده از الگوریتم یادگیری بهبود یابد. در بین متون فارسی و انگلیسی بررسی شده، تاکنون هیچ مطالعه‌ای به بررسی و تحلیل متن گزارش‌های بیماری سلیاک نپرداخته است و به همین دلیل امکان مقایسه نتایج وجود ندارد.

یکی از مزایای ساختارمند کردن گزارش‌های متن باز، امکان نظارت و ارزیابی عملکرد پزشک پاتولوژیست است. در فاز دوم مطالعه مشخص شد که طی فرایند تبدیل گزارش‌ها از حالت متنی به فرم ساختمند، به طور متوسط در ۳۰ درصد موارد حداقل یک یا دو فیلد فاقد مقدار هستند که نشان دهنده نقص اطلاعات در متن گزارش می‌باشد. دلیل این نقص می‌تواند ناشی از فراموشی پزشک برای ذکر این موارد در گزارش باشد.

نتیجه‌گیری

با توجه به تشخیص پیچیده بیماری سلیاک و خطرات ناشی از تشخیص نادرست این بیماری (۲۰ درصد بیماران مبتلا به سلیاک در صورت عدم درمان به سرطان روده کوچک دچار خواهند شد)، ضرورت طراحی سیستم‌های تسهیل‌گر و پشتیبان تصمیم در این حوزه به خوبی احساس می‌شود. در این مطالعه با طراحی و پیاده‌سازی مدلی برای ساختمند کردن گزارش‌های پاتولوژی این بیماری، علاوه بر تسهیل و تسریع در ورود و بازیابی اطلاعات، بهبود کیفیت و کامل بودن داده‌ها و افزایش خوانایی گزارش، امکان پردازش کامپیوتری داده‌ها و پیدا کردن روابط و الگوها با اهداف پژوهشی و مدیریتی نیز میسر می‌گردد. همان طور که در فاز سوم مطالعه نشان داده شد، پس از انتقال اطلاعات متون گزارش به فرم الکترونیکی ساختارمند، از الگوریتم یادگیری ماشین درخت تصمیم برای پردازش داده‌ها و طبقه‌بندی خودکار کلاس Marsh گزارش‌ها استفاده شد و نتایج قابل قبولی به دست آمد. پردازش متن و استخراج اطلاعات گام مهمی در تحلیل معنایی و اکتشاف دانش از متن به شمار می‌رود و کمک شایانی به انجام فعالیت‌های آموزشی و پژوهشی حوزه‌های آزمایشگاهی و بالینی می‌کند. از دیگر مزایای استانداردسازی فرم گزارش و ساختمند شدن اطلاعات می‌توان به امکان به اشتراک‌گذاری و تبادل داده‌ها بین مراکز درمانی مختلف و همچنین تشخیص و مشاوره از راه دور اشاره کرد.

محدودیت‌ها

تمرکز بر روی یک بیماری خاص با مجموعه لغات تخصصی به نسبت ثابت و محدود، امکان پردازش بهتر و مؤثرتر متن را فراهم می‌آورد. همچنین، پزشکان پاتولوژیست به طور معمول برای نگارش گزارش‌های روزانه از ساختار نحوی و معنایی مشابهی استفاده می‌کنند که از این امر نیز می‌توان در تحلیل هر چه دقیق‌تر متن بهره جست. این ویژگی‌ها از طرف دیگر عمومیت کار را کاهش می‌دهند و عملکرد سیستم را به صورت خاص تنها برای همان حوزه تعریف شده تقویت می‌کند که این امر را می‌توان به عنوان یکی از محدودیت‌های مطالعه در نظر گرفت. یکی دیگر از محدودیت‌های این پژوهش شیوع به نسبت پایین بیماری سلیاک بود که امکان دسترسی به مجموعه داده‌های بیشتر را دشوار می‌ساخت.

پیشنهادها

مدل پیشنهاد شده در این پژوهش را می‌توان برای ساختارمند کردن گزارش‌های مربوط به نتیجه پاتولوژی سایر بیماری‌ها و سرطان‌ها و یا هر گزارش متن باز دیگر در حوزه پزشکی مانند گزارش‌های رادیولوژی و یا گزارش شرح حال بیمار و موارد دیگر به کار برد. همان طور که ساختارمند کردن گزارش‌های متن باز موجب بهبود کارایی و صحت و سرعت در تصمیم‌گیری کادر پزشکی می‌شود، می‌توان در مطالعات آینده با بهره‌گیری از روش‌های مصورسازی داده‌ها به طراحی فرم گزارشی پرداخت که برای بیمار نیز قابلیت درک داشته و مفهوم باشد.

همچنین، پیشنهاد می‌شود در مطالعات بعدی مدل ارایه شده در این پژوهش با تعداد گزارش‌های بیشتر و متنوع‌تر به لحاظ نگارش و با تعداد و تنوع بیشتری در متخصصان تهیه کننده گزارش‌ها تکرار شود و با نتایج این پژوهش مقایسه گردد. به لحاظ روش‌های استفاده شده نیز، می‌توان عملکرد سایر روش‌های متن‌کاوی و پردازش زبان را در تشخیص مفاهیم و ارتباطات متنی بر روی گزارش‌های پزشکی پاتولوژی مورد بررسی و مقایسه قرار داد. به منظور ارزیابی میزان رضایتمندی پزشکان و بررسی کارایی فرم گزارش ساختارمند در بهبود کیفیت تصمیم‌گیری پزشک و تسریع و تسهیل امور، می‌توان مطالعه‌ای مقایسه‌ای بین گزارش‌های متن باز و خروجی به دست آمده از سیستم طراحی نمود. با توجه به تعداد محدود مطالعاتی که در زمینه ساختمند کردن گزارش‌های پاتولوژی انجام شده است، محققان این پژوهش امیدوار هستند که نتایج به دست آمده از این مطالعه مبنای پژوهش‌های جامع‌تر بعدی در این حوزه باشد.

تشکر و قدردانی

نویسندگان مقاله وظیفه خود می‌دانند که از کلیه پرسنل آزمایشگاه پاتوبیولوژی دانش زاهدان، به خصوص جناب آقای طباطبایی مسؤول فنی آزمایشگاه، به دلیل همکاری صمیمانه در زمینه جمع‌آوری داده‌ها کمال تشکر و قدردانی را داشته باشند.

References

1. Catassi C, Fasano A. Celiac disease diagnosis: simple rules are better than complicated algorithms. Am J Med 2010; 123(8): 691-3.
2. Ensari A. Gluten-sensitive enteropathy (celiac disease): Controversies in diagnosis and classification. Arch Pathol Lab Med 2010; 134(6): 826-36.
3. Marsh MN. Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). Gastroenterology 1992; 102(1): 330-54.
4. Oberhuber G, Granditsch G, Vogelsang H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur J Gastroenterol Hepatol 1999; 11(10): 1185-94.
5. Odze RD, Goldblum JR. Surgical pathology of the gi tract, liver, biliary tract, and pancreas. Philadelphia, PA: Elsevier Health Sciences; 2009.
6. Corazza GR, Villanacci V. Coeliac disease. J Clin Pathol 2005; 58(6): 573-4.
7. Tytgat NJ, Tytgat SH. Grading and staging in gastroenterology. Stuttgart, Germany: Thieme; 2011.
8. Zhang R, Wang Y, Melton G. Natural language processing in medicine. In: Agah A, Editor. Medical applications of artificial intelligence. New York, NY: CRC Press; 2013. p. 375-89.
9. McCowan I, Moore D, Fry M. Classification of cancer stage from free-text histology reports. Engineering in medicine and biology society. Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society New York City; 2006 Aug 31-Sep 3; New York, NY.
10. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, et al. Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc 2007; 14(6): 736-45.
11. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010; 17(4): 440-5.
12. Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc 2014; 21(5): 824-32.
13. Jouhet V, Defossez G, Burgun A, Le Beux P, Levillain P, Ingrand P, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods Inf Med 2012; 51(3): 242-51.
14. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001; 17-21.
15. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010; 17(5): 507-13.
16. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010; 62(8): 1120-7.
17. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008; 15(1): 14-24.
18. Srigley JR, McGowan T, Maclean A, Raby M, Ross J, Kramer S, et al. Standardized synoptic cancer pathology reporting: a population-based approach. J Surg Oncol 2009; 99(8): 517-24.
19. Stitt-Gohdes WL, Crews TB. The DELPHI technique: a research strategy for career and technical education. Journal of Career and Technical Education 2004; 20(2): 55-67.
20. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 4th ed. Berlin, Germany: Springer; 2013.
21. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Berlin, Germany: Springer Science & Business Media; 2006.
22. Pantanowitz L, Tuthill JM, Balis U. Pathology Informatics: Theory & Practice. Chicago, IL: ASCP; 2012.
23. Santorini B. Part-of-speech tagging guidelines for the penn treebank project [Project]. Philadelphia, PA: Department of Computer and Information Science, University of Pennsylvania; 1990.
24. Marneffe M, MacCartney B, Manning Ch. Generating typed dependency parses from phrase structure parses. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC); 2006 May 24-26; Genoa, Italy.
25. Raosoft. Sample Size Calculator [Online]. [cited 2004]; Available from: URL: http://www.raosoft.com/samplesize.html
26. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten L. The WEKA data mining software: an update. ACM SIGKDD Explorations 2009; 11(1): 10-8.
27. Hegenbart S, Uhl A, Vecsei A. Survey on computer aided decision support for diagnosis of celiac disease. Comput Biol Med 2015; 65: 348-58.
28. Li Y, Martinez D. Information extraction of multiple categories from pathology reports. Proceedings of the Australasian Language Technology Workshop; 2010 Dec 9-10; Melbourne, Australia.

پیوست ۱: با فرض آن که اطلاعات دموگرافیک در زمان پذیرش ثبت شده است. در این فرم فیلدهای ضروری مورد توافق از گزارش پاتولوژی بیماری سلیاک که دارای بیشترین توافق در میان پنل متخصصان بوده‌اند، در نظر گرفته شده است.

Date __/__/__

No. of biopsies ..... Oriented ..... Non-Oriented.....

Villi: ☐ normal atrophy ☐ Mild ☐ Severe

Villus/Crypt Ratio: normal [1:3] ..... altered .....

Intraepithelial Lymphocytes: normal..... increased .....

Evaluation with CD3 .....

Glands: normal ..... hyperplastic .....

Lamina Propria .....

Diagnosis (Oberhuber-marsh):

☐ Type 1 ☐ Type 3a ☐ Type 3c
☐ Type 2 ☐ Type 3b

Note: .....

Design and Implementation of a Structured Electronic Form for Celiac Disease Pathology Reports: A Text Mining Approach*

Azadeh Kamel-Ghalibaf[1], Farzaneh Khadem-Sameni[2], Majid Jangi[1], Mohammad Reza Mazaheri-Habibi[1], Kobra Etminani[3]

Original Article

Abstract

Introduction: Pathology reports generally use an unstructured text format and contain a complex web of relations between medical concepts. In order to enable computers to understand and analyze the reports' free text, we aimed to convert these concepts and their relations into a structured format.

Methods: The training, validation, and evaluation of this implementation study was based on a corpus of 258 pathology reports with a positive diagnosis of celiac disease randomly selected from among the records of 2 pathology laboratories. Our proposed system consisted of 3 phases of standardization of celiac disease pathology reports using Delphi technique with 3 experts, information extraction from free text reports with text mining techniques using Stanford Parser, and automatic classification of celiac disease stages in marsh system using decision tree classifier J48 algorithm.

Results: We were successful in extracting information from free text pathology reports and assigning each piece of information to the associated pre-defined fields in standardized template form with an accuracy of 76%. After determining marsh stage for each report in the third phase, our system showed an average overall accuracy of 62%. Evaluation of the third phase as an independent system with manually corrected, gold-standard input achieved an accuracy of greater than 84%.

Conclusion: The benefits of standardized synoptic pathology reporting include enhanced completeness and improved consistency, avoidance of confusion and error, and facilitation of the faster and safer transmission of critical pathological data in comparison with narrative reports.

Keywords: Text Mining; Celiac disease; Decision Support Systems; Clinical; Delphi Technique; Decision Trees

Received: 25 May, 2015 Accepted: 7 Apr, 2015

Citation: Kamel-Ghalibaf A, Khadem-Sameni F, Jangi M, Mazaheri-Habibi MR, Etminani K. **Design and Implementation of a Structured Electronic Form for Celiac Disease Pathology Reports: A Text Mining Approach.** Health Inf Manage 2016; 13(1): 19-27

* This article resulted from an independent research without financial support.

1- PhD Candidate, Medical Informatics, Department of Medical Informatics, Mashhad University of Medical Sciences, Mashhad, Iran
2- Assistant Professor, Pathologist, School of Medicine, Zahedan Brunch, Islamic Azad University, Iran
3- Assistant Professor, Medical Informatics, Department of Medical Informatics, Mashhad University of Medical Sciences, Mashhad, Iran (Corresponding Author) Email: etminanik@mums.ac.ir