مقاله مروری

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره چهارم شماره دوم. ص۱۶۷-۱۵۴، ۱۳۹۶

شناسه‌زدایی پرونده الکترونیک سلامت با استفاده از الگوریتم‌های یادگیری ماشین: یک مرور نظام‌مند

مصطفی لنگری‌زاده[۱]، اعظم اروجی[۲]*

• **دریافت مقاله:** ۹۶/۴/۲۶ • **پذیرش مقاله:** ۹۶/۶/۱۱

مقدمه: پرونده الکترونیک سلامت حاوی اطلاعات بالینی زیادی است که برای فعالیت‌هایی چون پایش بهداشت عمومی، بهبود کیفیت و تحقیقات مورد استفاده قرار می‌گیرد. همچنین پرونده الکترونیک سلامت شامل اطلاعات سلامت قابل شناسایی است و همین موضوع اشتراک و استفاده ثانویه از پرونده‌ها را محدود می‌کند. شناسه‌زدایی یکی از رایج‌ترین روش‌های حفظ محرمانگی اطلاعات بیماران است. این مقاله مروری نظام‌مند بر تحقیقات اخیر می‌باشد، که به حذف تمامی شناسه‌ها از پرونده الکترونیک سلامت با استفاده از انواع روش‌های شناسه‌زدایی مبتنی بر یادگیری ماشین پرداخته‌اند.

روش: این مقاله به صورت مروری نظام‌مند در بازه زمانی ۲۰۱۶ – ۲۰۰۶ در پایگاه‌های PubMed و Science direct انجام شد. مقالات با استفاده از چک‌لیست CASP و سپس توسط دو ارزیاب به‌طور مستقل بررسی و ارزشیابی شدند. در نهایت ۱۲ مقاله با معیارهای ورود مطالعه همخوانی داشتند.

نتایج: مقالات منتخب بر اساس روش و منابع دانش مورد استفاده، انواع شناسه‌ها، نوع اسناد بالینی، چالش‌ها و نتایج حاصل بررسی شده‌اند. نتایج نشان داد که در زمان انتشار داده‌های بالینی برای اهداف ثانویه شناسه‌زدایی مبتنی بر یادگیری ماشین راهکاری مناسب برای حفظ حریم خصوصی بیماران است. همچنین ترکیب الگوریتم‌های یادگیری ماشین و روش‌هایی چون تطابق الگو و عبارات منظم می‌تواند نیاز به داده آموزش را کاهش دهد.

نتیجه‌گیری: در پرونده‌های پزشکی اطلاعات شناسایی زیادی وجود دارد. این مطالعه نشان داد که روش‌های شناسه‌زدایی مبتنی بر یادگیری ماشین می‌توانند به طرز چشمگیری خطر افشای این اطلاعات را کاهش دهند.

کلید واژه‌ها: محرمانگی، حریم خصوصی، شناسه‌زدایی، یادگیری ماشین، پرونده الکترونیک سلامت

• **ارجاع:** لنگری‌زاده مصطفی، اروجی اعظم. شناسه‌زدایی پرونده الکترونیک سلامت با استفاده از الگوریتم‌های یادگیری ماشین: یک مرور نظام‌مند. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۶؛ ۴(۲): ۱۶۷-۱۵۴.

۱. دکترای انفورماتیک پزشکی، گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران

۲. دانشجوی دکترای انفورماتیک پزشکی، گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع‌رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران

* **نویسنده مسئول:** تهران، میدان ونک، خیابان ولیعصر، خیابان رشید یاسمی، پلاک ۶

• **شماره تماس:** ۰۲۱۸۸۷۹۴۳۰۱

• **Email:** orooji.a@tak.iums.ac.ir

[Downloaded from jhbmi.com on 2026-11-15]

مقدمه

حریم خصوصی یکی از حقوق مورد تأیید در قانون اساسی است که به موجب آن افراد حق دارند خلوت شخصی داشته و تنها بمانند، از شهرت بی دلیل در امان باشند و بدون اینکه سر زبان‌ها بیفتند زندگی کنند [۱]. در حوزه مراقبت سلامت، حریم خصوصی به معنی حق افراد برای محدود کردن دسترسی غیرمجاز به اطلاعات مراقبت سلامت آن‌ها است [۲]. پرونده الکترونیک سلامت سیستمی است که شامل اطلاعات پرونده بیمار است که دسترسی به آن‌ها از طریق شبکه‌ای از کامپیوترها میان ارائه‌دهندگان خدمات مراقبتی سازمان‌های مختلف امکان‌پذیر است [۳،۴]. پرونده الکترونیک سلامت EHR (Electronic Health Record) حاوی اطلاعات شخصی زیادی است برای مثال، متخصصین حوزه‌های مختلف در یک بیمارستان برای درمان فرد بایستی به EHR دسترسی داشته باشند همچنین محققین برای انجام کارآزمایی‌ها و مطالعات اپیدمیولوژیک از اطلاعات پرونده استفاده می‌کنند [۶،۵]. مستندات مدارک پزشکی محرمانه است و طبق اصل رازداری مدارک پزشکی، بیمار حق دارد انتظار داشته باشد که مدارک پزشکی مربوط به مراقبت درمانی‌اش محرمانه تلقی شود و بیمارستان از اطلاعات پرونده‌اش در برابر افشای غیرمجاز محافظت نماید [۷،۸]. اهمیت این موضوع سبب شده مطالعات زیادی به تأثیر افشای غیرمجاز اطلاعات بر افراد و سازمان‌ها بپردازند [۹،۱۰] و اکثر کشورها در این خصوص قوانین وضع کنند. از این جمله می‌توان به لایحه حریم شخصی اطلاعات و مدارک بهداشتی–درمانی (Health HRIP(Records and Information Privacy Act در استرالیا و قانون حفاظت داده (Data Protection Act) و لایحه مراقبت اجتماعی و بهداشتی- درمانی (The Health and Social Care Act) در انگلستان اشاره نمود [۱۱]. در آمریکا نیز قانون پاسخگویی و قابلیت انتقال بیمه بهداشتی- درمانی (Health Insurance Portability and HIPAA(Accountability Act در سال ۱۹۹۶ تصویب شد که دربردارنده شرایطی برای محافظت از اطلاعات سلامت بیماران در برابر افشای غیرمجاز است. در این قانون واژه اطلاعات سلامت حفاظت‌شده (Protected Health PHI(Information با عنوان" تمام اطلاعات سلامتی که باعث شناسایی بیمار می‌شود،" تعریف شده است [۱۲-۱۴]. HIPAA اطلاعات سلامت حفاظت شده را در ۱۸ طبقه تعریف می‌کند [۱۵]. جدول ۱ شناسه‌های تعریف شده توسط HIPAA را نمایش می‌دهد.

جدول ۱: مجموعه ۱۸ دسته PHI که توسط HIPAA معرفی شده است

- نام‌ها
- تمامی تقسیمات جغرافیایی کوچکتر از ایالت شامل خیابان، شهر، استان، کد پستی
- تمامی اجزای تاریخ‌ها (به‌جز سال) برای تاریخ‌هایی که مربوط به یک فرد خاص است مثل تاریخ تولد، تاریخ پذیرش و ترخیص، تاریخ فوت و تمامی اجزای تاریخ‌ها (به همراه سال) برای سن‌های بالای ۸۹
- شماره تلفن‌ها
- شماره فکس‌ها
- آدرس پست الکترونیک
- شماره امنیت اجتماعی
- شماره پرونده پزشکی
- شماره بیمه درمانی
- شماره حساب‌ها
- شماره گواهینامه/مدارک
- شماره وسایل نقلیه
- شماره ابزارها و شماره سریال‌ها
- آدرس صفحات وب
- شماره آدرس پروتکل اینترنت
- شناسه‌های بیومتریک مثل اثر انگشت
- تصویر کامل صورت یا تصاویر دیگری قابل شناسایی
- هر شماره، خصوصیت یا کدی که باعث شناسایی شود

در ایران نیز اصول ۳۳، ۳۲، ۲۸، ۲۵، ۲۳، ۲۲، ۲ و ۳۹ قانون اساسی جمهوری اسلامی تأکید بر حفظ حریم خصوصی افراد مشاهده می‌شود. در مورد حفاظت از داده‌های سلامت، ماده ۶۴۸ قانون مجازات اسلامی، افشای اطلاعات بیماران را جرم دانسته است [۱]. همچنین طبق منشور حقوق بیمار علاوه بر بیمار تنها گروه درمانی، افراد مجاز از طرف بیمار و افرادی که به حکم قانون مجاز تلقی می‌شوند می‌توانند به اطلاعات دسترسی داشته باشند [۱۶].

اساسی‌ترین تکنیک‌ها به منظور حفظ امنیت اطلاعات پزشکی و حریم خصوصی بیمار شامل محافظت فیزیکی، فنی و مدیریتی است. حفاظت فیزیکی به سیاست‌هایی چون نگهداری دستگاه‌ها در مکان‌های ایزوله، ایجاد نسخه کپی و پشتیبان از داده‌ها، جلوگیری کردن از دسترسی فیزیکی افراد ناشناخته به رایانه‌ها و ایجاد یک سیستم مناسب برای امحا اطلاق می‌شود. حفاظت فنی به راه‌اندازی دیوار آتش و روش‌های امن برای انتقال اطلاعات چون شبکه خصوصی مجازی و تکنیک‌های شناسایی‌زدایی De-ID(De-Identification) می‌پردازد. حفاظت مدیریتی شامل مواردی چون الزاماتی برای سیاست‌گذاری‌های امنیتی برای بخش مستندسازی، آموزش کارکنان، استفاده از رد ممیزی به منظور پیگیری همه ورود و

[Downloaded from jhbmi.com on 2026-11-15]

خروج‌ها و تغییرات در داده‌ها می‌باشد [۳،۱۳].

پروسه حذف یا تغییر تمام شناسه‌های موجود به منظور حداقل کردن خطر بازشناسایی افراد را شناسه‌زدایی گویند [۳،۱۷]. شناسه‌زدایی در گذشته اغلب به شکل دستی انجام می‌گرفت؛ اما Dorr و همکاران زمان موردنیاز برای شناسه زدایی دستی را مورد ارزیابی قرار داده (میانگین ۸۷/۲±۶۱ ثانیه برای هر یادداشت) و نتیجه‌گیری کردند که این کار بسیار زمان بر است [۱۸]. همچنین از آنجا که یکی از معیارهای ارزیابی سازمان جهانی بهداشت از سیستم سلامت هر کشور استفاده از فناوری اطلاعات در مراقبت بهداشتی است [۱۹]، اغلب پژوهشگران بر روش‌های خودکار شناسه‌زدایی متمرکز شدند [۲۰]. بعدها اثر ترکیبی این دو نیز بررسی گشت. برای مثال South و همکاران یک رابط تعاملی برای شناسه‌زدایی ارائه کردند که از یک روش شناسه‌زدایی خودکار به عنوان پیش پردازشگر برای شناسه‌زدایی دستی استفاده کردند [۲۱]. برای شناسه‌زدایی خودکار از تکنیک‌های یادگیری ماشین استفاده می‌شود، زیرا این روش‌ها به کرات در مدیریت متون پزشکی استفاده شده‌اند و به نتایج مطلوبی دست یافته‌اند [۲۲].

با توجه به اهمیت حفظ محرمانگی اطلاعات بیماران و توجه ویژه به شناسه‌زدایی خودکار در سال‌های اخیر، این مقاله به‌طور نظام‌مند به مطالعه روش‌های مختلف شناسه‌زدایی خودکار از پرونده‌های الکترونیک بیماران پرداخته است. اگرچه مطالعات مروری دیگری نیز به بررسی روش‌های مختلف شناسه‌زدایی پرداخته‌اند [۲۰،۲۲–۲۹]؛ اما هدف این مقاله آن است که مقالات ده سال اخیر که از الگوریتم‌های یادگیری ماشین (Machine Learning)ML برای شناسه‌زدایی بهره برده‌اند به صورت نظام‌مند جستجو شده و از جنبه‌های مختلف به تفصیل مورد ارزیابی قرار گیرند. بدین منظور مقالات بر حسب روش مورد استفاده، منبع دانش، نوع متنی که برای ارزیابی سیستم استفاده شده، مجموعه شناسه در نظر گرفته شده و نتایج حاصل تقسیم‌بندی، بررسی و مقایسه شده‌اند.

روش

این مطالعه از مرور نظام‌مند جهت اطمینان از دقت و جامعیت فرآیند جستجو و بازیابی استفاده کرده است. برای این پژوهش مقالات پایگاه‌های PubMed و Sciencedirect که در بازه زمانی ۲۰۰۶/۱/۱ تا ۲۰۱۶/۱/۱ و به زبان انگلیسی منتشر شده‌اند با ترکیب کلیدواژه‌های مختلف بررسی شد. جستجوی موردنظر در شکل ۱ نمایش داده شده است. همچنین معیارهای ورود و خروج مطالعه در جدول ۲ آمده است. از میان مقالات، تنها مواردی که برای شناسه‌زدایی از الگوریتم‌های ML استفاده کرده‌اند و ارزیابی سیستم خود را بر اساس EHR نوشته شده به زبان انگلیسی انجام داده‌اند وارد مطالعه شدند.

شکل ۱: جستجوی مورد استفاده برای بازیابی مطالعات مورد نظر

Science Direct	Pub-date > 2005 and pub-date < 2016 and (TITLE-ABSTR-KEY(de-identif*) or TITLE-ABSTR-KEY(deidentif*) or TITLE-ABSTR-KEY(Anonymization) or TITLE-ABSTR-KEY(De-personalization) or TITLE-ABSTR-KEY(Depersonalization) or TITLE-ABSTR-KEY(Pseudonymization)) and ("electronic health record" or "electronic medical record").
PubMed	(Electronic health record [All Fields] OR Electronic medical record [All Fields]) AND (de-identif*[Title/Abstract] OR deidentif* [Title/Abstract] OR Anonymization[Title/Abstract] OR De-personalization [Title/Abstract] OR Depersonalization [Title/Abstract] OR Pseudonymization [Title/Abstract]) AND ("2006/01/01"[PDAT] : "2016/01/01"[PDAT])

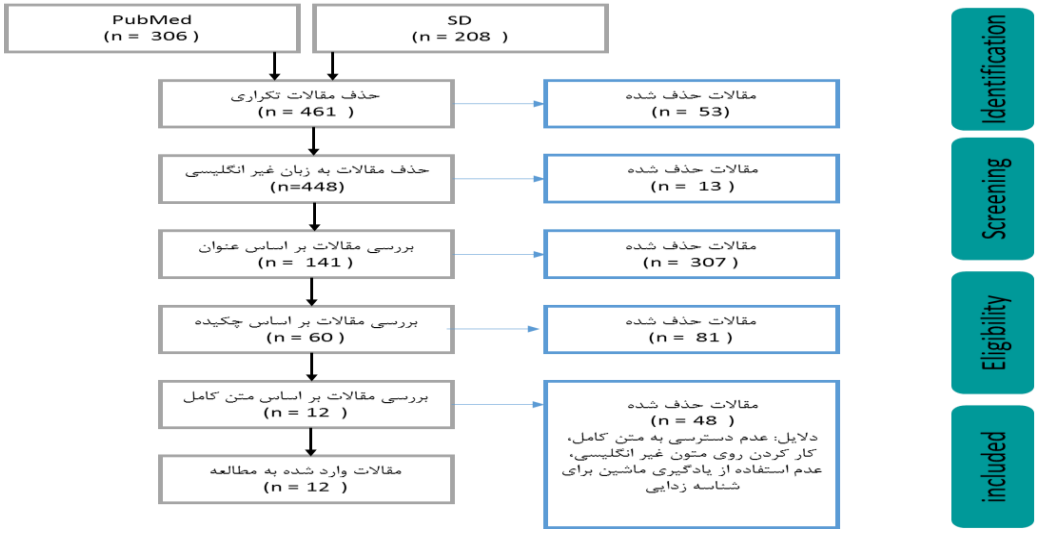
جدول ۲: معیارهای ورود و خروج مقالات

معیارهای ورود	معیارهای خروج
• مقالاتی که روی پرونده‌های پزشکی به زبان انگلیسی کار کرده‌اند.	• حذف مقالاتی که به متن کامل آن‌ها دسترسی نبود.
• استفاده از الگوریتم‌های یادگیری ماشین برای شناسه‌زدایی	• حذف روزنامه، نامه به سردبیر، کارگاه، پوستر، گزارش کوتاه، کتاب و پایان‌نامه

به منظور انتخاب مقالات از نمودار جریان کار PRISMA استفاده شد. با استفاده از جستجوی شکل ۱ تعداد ۵۱۴ مقاله که پس از سازمان‌دهی آن‌ها در نرم‌افزار اندنوت (Endnote) و بررسی مستقل دو ناظر به ترتیب عنوان، چکیده و متن کامل با در نظر گرفتن معیارهای ورود و خروج یافت شد، در مرحله بعد سنجش کیفی مقالات بر اساس چک‌لیست ۱۲ سؤالی مطالعات تست تشخیصی Critical Appraisal Skills Programme) CASP [۳۰] انجام شد. نتایج حاصل از

[Downloaded from jhbmi.com on 2026-11-15]

چک‌لیست با استفاده از نرم‌افزار اکسل ۲۰۱۳ سازمان‌دهی شـد. پروتکل جستجو در شکل ۲ نمایش داده شده است.



شکل ۲: پروتکل جستجو بر اساس نمودار جریان کاری انتخاب مقالات PRISMA

نتایج

طی ارزیابی معیارهای ورود و خروج توسط ناظرین، موارد تناقض میان دو ناظر با برگزاری یک جلسه مشترک رفع گشت همچنین میزان توافق میان دو ناظر با استفاده از آزمون آماری کاپا محاسبه و برابر ۰/۸۱ شد که از نظر آماری معنی‌دار بود (P<۰/۰۰۱). در نهایت، تعداد ۱۲ مقاله وارد مطالعه شدند که به همراه نمره CASP، سال چاپ و نام مجله‌ای که آن را چاپ کرده در جدول ۳ نشان داده شده‌اند. مقالات منتخب در دو بخش مورد بررسی قرار گرفته‌اند.

جدول ۳: نمره CASP برای مقالاتی که وارد مطالعه شدند. به هر آیتم موجود در چک‌لیست با واژه‌های بله (نمره =۲)، خیر (نمره=۱) و نمی‌توان گفت (نمره=۰) امتیاز تعلق گرفت.

نام نویسنده	سال انتشار	نمره CASP (٪)	محل انتشار
[۳۴]Szarvas	۲۰۰۷	۵۸/۳	Journal of the American Medical Informatics Association
[۴۱]Wellner	۲۰۰۷	۶۲/۵	Journal of the American Medical Informatics Association
[۳۶]Gardner	۲۰۰۸	۴۵/۸	IEEE International Symposium on Computer-Based Medical Systems
[۳۳]Uzuner	۲۰۰۸	۶۶/۷	Artificial Intelligence in Medicine
[۳۷]Aberdeen	۲۰۱۰	۷۵/۰	International Journal of Medical Informatics
[۳۵]McMurry	۲۰۱۳	۶۶/۷	BMC Medical Informatics and Decision Making
[۴۲]Ferrández	۲۰۱۳	۷۵/۰	Journal of the American Medical Informatics Association
[۳۹]He	۲۰۱۵	۶۶/۷	Journal of Biomedical Informatics
[۴۸]Zuccon	۲۰۱۴	۸۷/۵	Artificial Intelligence in Medicine
[۴۰]Dehghan	۲۰۱۵	۶۲/۵	Journal of Biomedical Informatics
[۳۲]Yang	۲۰۱۵	۶۶/۷	Journal of Biomedical Informatics
[۴۷]Liu	۲۰۱۵	۶۶/۷	Journal of Biomedical Informatics

دسته اول، روش‌های شناسه‌زدایی مبتنی بر داده، مقالاتی هستند که تنها بر اساس الگوریتم‌های ML و با دو مرحله آموزش و آزمایش طراحی شده‌اند. در حالی که دسته دوم یعنی روش‌های شناسه‌زدایی ترکیبی علاوه بر الگوریتم‌های ML از لغت‌نامه‌ها نیز برای جستجوی شناسه‌ها بهره برده‌اند؛ در واقع این گروه ترکیبی از روش‌های مبتنی بر داده و روش‌های مبتنی بر دانش هستند. در این بخش سعی بر آن است که مقالات از جنبه‌های متعددی چون موارد زیر بررسی گردند:

- روش (الگوریتم(های) ML مورد استفاده در شناسه زدایی)
- پایگاه دانش (در برخی از مقالات از منابعی مانند واژه‌نامه‌ها یا لیست‌های سرشماری به عنوان لغت‌نامه در جستجوی شناسه‌ها استفاده شده است).
- مجموعه شناسه‌ها (مجموعه مقادیری که به عنوان شناسه در نظر گرفته شده‌اند که می‌تواند مطابق با مجموعه تعریف شده توسط HIPAA باشد یا خیر؟)

[Downloaded from jhbmi.com on 2026-11-15]

- نوع داده (نوع متنی که روش پیشنهادی روی آن پیاده‌سازی شده است که می‌تواند شامل گزارش پاتولوژی، خلاصه ترخیص و ... باشد).
- نتایج حاصل
- نام سیستم (در صورت وجود)

۱. روش‌های شناسه‌زدایی مبتنی بر داده

این روش‌ها از الگوریتم‌های مختلف یادگیری ماشین همچون ماشین بردار پشتیبان (Support Vector Machine) SVM، درخت تصمیم و میدان تصادفی شرطی CRF(Conditional Random Fields) استفاده می‌کنند. تمامی این الگوریتم‌ها برای ساخت مدلی که بتواند کلمات را دریکی از دو دسته PHI , not PHI قرار دهد نیاز دارند که توسط یک مجموعه داده آموزش ببینند. این مجموعه اصولاً به شکل متونی است که در آن‌ها PHIهای مختلف برچسب خورده‌اند. این سیستم‌ها از مجموعه‌های متفاوتی از ویژگی‌ها استفاده می‌کنند: ۱) ویژگی‌های در سطح کلمه (شامل خصوصیات حرفی/لغوی و ریخت‌شناسی کلمات) این روش که بر اساس n-gram کار می‌کند رایج‌ترین و بهترین روش برای شناسه‌زدایی و طبقه‌بندی کلمات است. ۲) ویژگی‌های نحوی (شامل برچسب‌زنی ادات سخن (Part-of-Speech POS(Tagging، ۳) ویژگی‌های معنایی (طبقه‌بندی معناگرای کلمات)، ۴) ویژگی‌های در سطح سند (مانند واژگان رایج در یک سند) و ۵) ویژگی‌های مبتنی بر متن (شامل میزان تکرار واژگان) [۲۰،۳۱،۳۲]. جدول ۴ خلاصه‌ای از مقالات مربوط به شناسه‌زدایی مبتنی بر داده پرونده الکترونیک سلامت که تنها از روش‌های ML بهره برده‌اند را نشان می‌دهد.

جدول ۴ : مشخصات سیستم‌های شناسه‌زدایی مبتنی بر داده (یادگیری ماشین)

نویسنده	الگوریتم ML	نوع متن	منبع دانش	نتایج (درصد)	نام	شناسه‌ها
Szarvas [۳۴]	Decision tree	خلاصه ترخیص	۵ لیست از اینترنت شامل اسامی، مکان‌های جغرافیایی آمریکا، نام کشورهای جهان، نام شهرهای بزرگ جهان، نام بیماری‌ها لیستی از واژه‌های non-PHI مستخرج از داده‌های آموزش لیستی از واژه‌های non name entity از پایگاه داده کنفرانس یادگیری زبان طبیعی سال ۲۰۰۳ CoNLL (http://www.cnts.ua.ac.be/conll2003/)	دقت: ۹۸/۷۹ فراخوانی: ۹۴/۷۳ معیار F: ۹۶/۷۱	-	۸ شناسه معرفی شده در i2b2NLP 2006 شامل نام بیمار، پزشک، مکان، بیمارستان، تاریخ، شماره شناسایی، شماره تلفن، سن
Gardner [۳۶]	CRF	گزارش پاتولوژی	ندارد	صحت: ۹۸/۲	HIDE	شناسه‌های معرفی شده توسط HIPAA
Uzuner[۳۳]	SVM	خلاصه ترخیص	واژگان سرعنوان‌های موضوعی پزشکی(Medical Subject Headings)، لیستی از نام‌ها، مکان-ها و بیمارستان‌ها	دقت : ۹۹ فراخوانی: ۹۷ معیار F: ۹۹	Stat De-id	۷ گروه شناسه شامل: بیمار (نام و نام خانوادگی بیمار و خانواده‌اش)، دکتر (نام و نام خانوادگی پزشکان و سایر ارائه‌دهندگان مراقبت)، نام بیمارستان، شماره‌های شناسایی (شامل هر ترکیبی از اعداد و حروف)، تاریخ‌ها، آدرس (ایالت، شهر، نام خیابان، کد پستی، نام و شماره ساختمان‌ها)، شماره تلفن (تلفن و فکس)
McMurry [۳۵]	Decision tree (J84)	خلاصه ترخیص	نشریات پزشکی حاشیه گذاری شده توسط cTAKES (۱۳) لیست سرشماری ۱۹۹۰ آمریکا ۱۰ لغتنامه پزشکی شامل...,UMLS,ICD10, MESH,SNOMED, LOINC	دقت: ۶۲ فراخوانی: ۹۸ معیار F1: ۷۶ معیار F10: ۹۸	-	۸ گروه عمده از شناسه‌های معرفی‌شده توسط HIPAA(شامل نام بیمارستان، پزشگ، سن، تاریخ، سن، مکان، نام بیمار، شماره شناسایی و تلفن)
Aberdeen [۳۷]	CRF	اولین ارزیابی: توسط ۴ نوع پرونده از مرکز پزشکی وندربیت: خلاصه ترخیص‌ها، گزارش‌های آزمایشگاهی، نامه‌ها و خلاصه دستورها دومین ارزیابی بر اساس خلاصه ترخیص	لیست سرشماری ۱۹۹۰ آمریکا، لیست نام بیمارستان‌های سرشناس	ارزیابی اول: دقت: ۹۴/۳ فراخوانی: ۹۷/۸ معیار F: ۹۶ ارزیابی دوم: دقت: ۹۷/۸ فراخوانی: ۹۵/۱ معیار F : ۹۶/۵	MIST	شناسه‌های معرفی شده توسط HIPAA به علاوه نام مؤسسات
He [۳۹]	CRF	کل پرونده پزشکی	لغت‌نامه‌ای مستخرج از مجموعه داده و صفحات وب مربوط به شهر، کشور و ایالات (http://en.wikipedia.org/wiki/List_of_United_States_cities_by_population; http:// en.wikipedia.org/wiki/List_of_U.S._state_abbreviations; http://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area.)	ارزیابی مبتنی بر نشانه: ۱. با شناسه‌های i2b2 دقت: ۹۵/۷۱ فراخوانی: ۹۰/۵۱ معیار F: ۹۳/۰۴ ۲. با شناسه‌های HIPAA دقت: ۹۷/۰۸ فراخوانی: ۹۳/۷۱ معیار F: ۹۵/۳۶	WI-deId	۸ شناسه معرفی شده در i2b2NLP 2006 و شناسه‌های HIPAA

Uzuner و همکاران از SVM به منظور شناسه‌زدایی استفاده کرده است بدین صورت که سیستم با در نظر گرفتن دو کلمه قبل و بعد از واژه هدف به تعیین دسته مرتبط با آن (PHI یا non PHI) می‌پردازد. طبقه‌بندی در این مقاله بر اساس دو

[Downloaded from jhbmi.com on 2026-11-15]

دسته ویژگی املایی (شامل طول کلمه، حروف کوچک و بزرگ و ...) و نحوی (شامل POS خود کلمه و کلمات مجاورش) صورت می‌گیرد. این سیستم نقاط ضعفی دارد از جمله این که در آن به نحوه شناسایی انواع PHI به صورت مجزا پرداخته نشده است [۳۳].

Szarvas و همکاران [۳۴] از یک روش طبقه‌بندی با یادگیری تکراری مبتنی بر درخت تصمیم برای شناسه‌زدایی استفاده می‌کند. در این مقاله برای ترکیب سه طبقه‌بند که همگی C4.5 هستند از الگوریتم ترکیبی (Boosting) استفاده شده است. به علاوه برای طبقه‌بندی از ویژگی‌های مختلف کلمات استفاده شده است، برای مثال شناسنامه‌ها از عنوان گزارش‌ها استخراج شده‌اند تا کارایی روش را بهبود دهند، یا مثلاً از عبارات منظم برای توصیف خصوصیات مشترک کلاس‌های داده‌های خوش فرم مثل اعداد و تاریخ‌ها استفاده شده است و ویژگی‌هایی چون طول کلمه، تکرر آن و یکسری لغت‌نامه برای شناسایی اسامی و محل‌های جغرافیایی مورداستفاده قرار گرفته‌اند. لغت‌نامه‌هایی که در اینجا استفاده شده است شامل پنج لیست مستخرج از اینترنت برای نام افراد، مکان‌های جغرافیایی آمریکا، کشورها، شهرهای بزرگ جهان و بیماری‌ها و دو لیست از داده‌های non PHI و non NER است. نتایج حاصل از اجرای این سیستم روی مجموعه داده مربوط به چالش پردازش زبان طبیعی (Natural Language Processing) i2b2 NLP در سال ۲۰۰۶ که شامل ۸۸۹ خلاصه ترخیص بی‌نام شده می‌باشد [۲۳] که نشان دهنده عملکرد ضعیف آن در تشخیص برخی شناسه‌ها چون محل، شماره تلفن، نام بیمارستان و تاریخ است.

McMurry و همکاران روشی مبتنی بر J84 ارائه کرده‌اند که برعکس سایر رویکردها، هم‌زمان با شناسایی PHIs به دنبال کشف کلماتی است که PHI نیستند. در این مقاله دو موضوع مهم مطرح می‌شود، اول اینکه روش ارائه شده برای آموزش به مجموعه داده شناسه‌زدایی شده نیاز ندارد، زیرا از لغت‌نامه‌ها و مراجعی که برای عموم در دسترس است استفاده می‌کند. دوم اینکه با فرض این که کلماتی در مقالات زیاد تکرار شده‌اند یا در لغت‌نامه‌های پزشکی آمده‌اند احتمال اینکه PHI باشند بسیار کم است، توانسته هم‌زمان به تعیین PHI و non-PHI بپردازد. در این مقاله برای هر کلمه چهار گروه ویژگی استخراج شده است: ویژگی‌های لغوی (POS، حروف بزرگ موجود در کلمه، طول کلمه، حرف یا عدد بودن آن)، تکرار (تکرار نسبی آن در متون پزشکی عمومی و تخصصی)، لغت‌نامه‌های پزشکی (مقایسه کلمه با ۱۰ لغت‌نامه معروف پزشکی و تعیین حضور و عدم حضور آن) و در نهایت PHIs رایج شامل مقایسه کلمه با لیست‌های مربوط به سرشماری ایالات متحده (1990 Census Gazetteer [http://www.census.gov/geo/www/gazetteer/gazette.html]) و عبارات منظمی که برای تشخیص انواع PHIs معرفی شده توسط HIPAA طراحی شده اند [۳۵].

Gardner و Xiong در مقاله خود سیستمی یکپارچه برای شناسه‌زدایی از هر دو گروه داده پزشکی چه ساختارمند و چه غیر ساختارمند ارائه کرده که بر اساس CRF طراحی شده است. در این سیستم نیز همچون تمامی روش‌های مبتنی بر یادگیری ماشین در مرحله اول به صورت دستی یا خودکار به برچسب‌زنی کلمات اسناد پرداخته می‌شود تا به عنوان مجموعه داده آموزش در اختیار طبقه‌بند قرار گیرد؛ اما وجه تمایز این سیستم با سایر روش‌ها در این است که در طول توسعه سیستم دو عمل طبقه‌بندی و برچسب‌زنی بر اساس نتایج طبقه‌بندی مرتباً تکرار می‌شوند تا دقت سیستم افزایش یابد [۳۶].

یکی از مهم‌ترین ابتکارات در زمینه شناسه‌زدایی بسته نرم‌افزاری حذف شناسه (MITRE Identification MIST(Scrubber Toolkit است که محیطی تحت وب را برای توسعه شناسه‌زدایی خودکار از انواع متفاوتی از اسناد فراهم می‌کند. MIST به شناسه‌زدایی به صورت یک مسئله برچسب‌زنی دنباله می‌نگرد و با استفاده از CRF به طبقه‌بندی کلمات به دو گروه جزئی از PHI و عدم ارتباط با PHI پرداخته است. ساختار منعطف این سیستم این اجازه را به کاربران می‌دهد تا ویژگی‌های یادگیری جدیدی را به منظور بالا بردن دقت طبقه‌بند به آن اضافه کنند، در واقع عملکرد این سیستم به مجموعه آموزش آن وابسته است. یکی از مزایای این سیستم این است که برای تیم‌های مختلف پزشکی که قصد شناسه‌زدایی نوع خاصی از متون پزشکی را دارند این امکان را فراهم می‌کند تا سیستم را سفارشی کنند بدون این که نیاز باشد به اصل مدارک پزشکی دسترسی داشته باشند و همین امر باعث شده که در مقالات زیادی برای ارزیابی نکات مختلف شناسه‌زدایی از این سیستم استفاده شود [۳۷]. برای مثال Li و همکاران [۳۸] موضوع تعمیم‌پذیری شناسه‌زداهای مبتنی بر ML را مطرح کرده‌اند و به انتخاب داده آموزش مناسب اشاره دارند. در واقع زمانی که یک مدل بر اساس نوع خاصی از متون بالینی آموزش ببیند بهترین پاسخ‌ها را در مرحله آزمون نیز برای همان نوع خاص متون خواهد داشت. این مقاله

[Downloaded from jhbmi.com on 2026-11-15]

پیشنهاد می‌دهد که ابتدا متون بر اساس ویژگی‌های مربوط به پیچیدگی نوشتاری، خوشه‌بندی شوند و سپس از هر خوشه خوانایی (که بر اساس تعداد بخش‌های یک کلمه و تعداد کلمات هر جمله برای هر متن محاسبه می‌شود) و غنای متن (هر چه یک متن کلمات منحصربه فرد بیشتری داشته باشد غنای بیشتری دارد) تأثیر این رویکرد بر شناسه‌زدای MIST بررسی شده و نشان داده شد که میزان معیار F-Measure) F-) از ۸۸٪ (در حالتی که نمونه‌ها تصادفی انتخاب شدند) به ۹۲٪ (انتخاب نمونه‌ها از خوشه‌های مختلف) افزایش یافت.

He و همکاران [۳۹] رویکردی را بر اساس CRF معرفی می‌کنند به نام WI-deId که دارای چهار ماژول است: ۱) پیش‌پردازش متن: در این مرحله کارهایی از قبیل مرزبندی جملات

و نشانه‌گذاری و سرهم کردن کلمات چند بخشی توسط یک مجموعه عبارات منظم انجام می‌شود، ۲) تولید ویژگی: سه گروه ویژگی به منظور آموزش مدل معرفی می‌شوند شامل ویژگی‌های لغوی، ریخت‌شناسی و لغت نامه‌ای، ۳) آموزش بر تعدادی نمونه برای آموزش و سپس آزمایش مدل استفاده شود. ویژگی‌هایی که در این کار در نظر گرفته شدند عبارت‌اند از: اساس ویژگی‌های گام دوم و سپس ۴) طبقه‌بندی با CRF.

این مقاله سیستم پیشنهادی را از دو جهت ارزیابی می‌کند. اول تأثیر مرحله پیش پردازش بررسی می‌شود و سپس به تأثیر هر دسته از ویژگی‌ها در کارایی سیستم پرداخته می‌شود.

۲. روش‌های شناسه‌زدایی ترکیبی

روش‌های شناسه‌زدایی مبتنی بر داده برای یادگیری مدل خود به داده آموزش کامل و جامعی نیاز دارند تا دقت خوبی داشته باشند و این موضوع بزرگ‌ترین ایرادی است که به آن‌ها وارد می‌شود. ازاین‌رو برخی از مقالات علاوه بر الگوریتم‌های یادگیری ماشین، با استفاده از تکنیک‌هایی چون تطبیق الگو، جستجو در لغت‌نامه‌ها و یا تعریف عبارات منظم به تشخیص گروهی از شناسه‌ها پرداخته‌اند. نتیجه نهایی درواقع ترکیبی از خروجی حاصل از دو مدل (مبتنی بر دانش و مبتنی بر داده) می‌باشد [۴۰]. جدول ۵ خلاصه‌ای از مقالات مربوط به شناسه زدایی متون بالینی که از روش‌های ترکیبی بهره برده‌اند را نشان می‌دهد.

جدول ۵: مشخصات سیستم‌های شناسه‌زدایی ترکیبی

نویسنده	روش	نوع متن	منبع دانش	نتایج (درصد)	نام سیستم	شناسه‌ها
Wellner [۴۱]	CRF, HMM, عبارات منظم	خلاصه ترخیص	لیست ایالت‌های آمریکا، ماه‌ها و کلمات رایج انگلیسی	دقت: ۹۹/۲۲ فراخوانی: ۹۷/۵۰ معیار F: ۹۸/۳۵	–	۸ شناسه معرفی شده در i2b2NLP 2006
Ferrández [۴۲]	تطبیق الگو، جستجو در لغت‌نامه‌ها، CRF, SVM	ارزیابی اول: خلاصه ترخیص ارزیابی دوم: انواع متفاوتی از اسناد بالینی VHA	لیست نام و نام خانوادگی از سرشماری آمریکا، لیست اسامی ایالت‌ها، شهرها، کشورها و شرکت‌ها از وب (Wikipedia, usps.com) لیست کلمات رایج و عناوین بالینی و درمانگاه‌ها از متن آموزش VHA	ارزیابی اول: دقت: ۸۷/۸ فراخوانی: ۹۲/۱ معیار F1: ۸۹/۹ معیار F2: ۹۱/۲ ارزیابی دوم: دقت: ۸۱/۵ فراخوانی: ۹۲/۱ معیار F1: ۸۶/۴ معیار F2: ۸۹/۷	BOB	۱۶ شناسه شامل: نام بیمار و بستگان وی، نام ارائه‌دهنده مراقبت، سایر نام‌ها، شهر/خیابان، ایالت/کشور، کد پستی، متفرقه، واحدهای مراقبت سلامت، سایر سازمان‌ها، تاریخ، سن بالای ۸۹ سال، شماره تلفن، آدرس الکترونیک، کد ملی، سایر شماره‌های شناسایی
Dehghan [۴۰]	قوانین اگر–آنگاه، جستجو در لغت‌نامه و CRF	کل پرونده پزشکی	لغت‌نامه‌های جمع‌آوری‌شده از ویکی‌پدیا، GATE و deid	ارزیابی اول: دقت: ۹۷/۲۲ فراخوانی: ۹۲/۵۰ معیار F1: ۹۴/۸۰ ارزیابی دوم: دقت: ۹۷/۹۷ فراخوانی: ۹۵/۴۲ معیار F1: ۹۶/۶۸	-	ارزیابی اول: ۲۵ شناسه تعریف‌شده در i2b2-2014 ارزیابی دوم: شناسه‌های معرفی‌شده توسط HIPAA
Yang [۳۲]	قوانین اگر–آنگاه، عبارات منظم، جستجو در لغت‌نامه و CRF	کل پرونده پزشکی	لغت‌نامه خودساخته	ارزیابی اول: دقت: ۹۸/۱۵ فراخوانی: ۹۴/۱۴ معیار F1: ۹۶/۱۱ ارزیابی دوم: دقت: ۹۸/۸۹ فراخوانی: ۹۶/۲۹ معیار F1: ۹۷/۵۷	-	ارزیابی اول: ۲۵ شناسه تعریف‌شده در i2b2-2014 ارزیابی دوم: شناسه‌های معرفی‌شده توسط HIPAA

[Downloaded from jhbmi.com on 2026-11-15]

جدول ۵: مشخصات سیستم‌های شناسه‌زدایی ترکیبی(ادامه)

Liu [۴۷]	قوانین اگر-آنگاه، عبارات منظم، CRF	کل پرونده پزشکی	ندارد	ارزیابی اول: دقت:۹۵/۶۴ فراخوانی:۹۳/۶۶ معیار F1: ۹۴/۶۴ ارزیابی دوم: دقت:۹۷/۴۸ فراخوانی: ۹۵/۷۸ معیار F1: ۹۶/۶۲	-	ارزیابی اول: ۲۵ شناسه تعریف‌شده در i2b2-2014 ارزیابی دوم: شناسه‌های معرفی‌شده توسط HIPAA
Zuccon [۴۸]	عبارات منظم، تطبیق الگو و CRF	ارزیابی اول: خلاصه ترخیص ارزیابی دوم: یادداشت-های بالینی ارزیابی سوم: گزارش پاتولوژی و سیتولوژی	ندارد	ارزیابی اول: دقت: ۹۸/۹۹ فراخوانی: ۸۸/۵۹ معیار F: ۹۳/۰۰ ارزیابی دوم: معیار F: ۹۸/۰۶ ارزیابی سوم: معیار F: ۸۲/۴۸	Anonym	۸ گروه شناسه شامل: بیمار (نام و نام خانوادگی بیمار و خانواده‌اش)، دکتر (نام و نام خانوادگی پزشکان و سایر ارائه‌دهندگان مراقبت)، نام بیمارستان، شماره-های شناسایی (شامل هر ترکیبی از اعداد و حروف)، تاریخ‌ها، آدرس (ایالت، شهر، نام خیابان، کد پستی، نام و شماره ساختمان‌ها)، شماره تلفن (تلفن، موبایل و فکس)، اعداد

در مقاله Wellner و همکاران [۴۱] یک سیستم شناسه‌زدایی بر مبنای وفق‌پذیری سریع دو بسته نرم‌افزاری موجود برای شناسایی موجودیت‌های نامدار (Named Entity Recognition) NER به نام‌های Carafe و LingPipe ارائه شده است. در این روش به کلمات برچسب‌های خاصی داده می‌شود که نشان دهنده این است که آن کلمه ابتدا، انتها یا بخشی از یک PHI هست یا نه (برچسب‌گذاری یک دنباله). همچنین از عبارات منظم برای شناسایی PHIs عددی استفاده می‌گردد. Carafe نرم‌افزاری است که مبتنی بر CRF بوده و قابلیت مهم آن سهولت اضافه کردن ویژگی‌های جدید به آن است، از این طریق محققان به راحتی می‌توانند با اضافه کردن یکسری ویژگی جدید دقت شناسه‌زدایی را افزایش دهند. سیستم NER دیگری که در این مقاله به کار گرفته شده LingPipe است که بر اساس مدل مخفی مارکف به برچسب زدن کلمات می‌پردازد. سیستم عملکرد ضعیفی در تشخیص برخی شناسه‌ها داشت و پژوهشگران با اضافه کردن عبارات منظم به سیستم توانستند فراخوانی را افزایش دهند. اضافه کردن ویژگی‌های خاص یک وظیفه بر اساس یک مجموعه داده آموزشی محدود و قابلیت تنظیم کردن الگوریتم برای رسیدن به اهداف متفاوت (گاهی اوقات دقت مهم‌تر از فراخوانی است و گاهی برعکس این موضوع صادق است) از مزایای این سیستم است.

سیستم ارائه شده توسط Ferrández و همکاران به نام BoB(Best-of-Breed) نیز نمونه‌ای دیگر از سیستم‌های ترکیبی است که به شناسه‌زدایی از اسناد بالینی مدیریت سلامت بازنشستگان جنگ (Veterans Health Administration) VHA می‌پردازد. اهداف این مقاله در دو بخش ارائه شد: ۱) بهبود هر چه بیشتر محرمانگی اطلاعات بیمار با حذف شناسه‌ها تا جایی که امکان دارد و ۲) سندی که شناسه‌زدایی شد حاوی حداکثر اطلاعات بالینی قابل استفاده باشد. BoB سیستمی ترکیبی است که روش‌های مختلف مبتنی بر پایگاه قانون و یادگیری ماشین را که در شناسایی نوع خاصی از PHIs به خوبی عمل کرده‌اند را یکپارچه می‌نماید. این سیستم برای شناسه‌زدایی دو جزء دارد: استخراج با حساسیت بالا (که از یک مجموعه قوانین و CRF برای تعیین تمامی PHI ممکن در متن استفاده می‌کند) و فیلترسازی مثبت کاذب (که با استفاده از SVM آن دسته از PHIs که در فاز اول به اشتباه برچسب خورده‌اند را حذف می‌کند). در واقع بخش اول باعث افزایش فراخوانی و دومی سبب بالا رفتن دقت می‌گردد. BoB این امکان را فراهم می‌آورد که کلماتی که در بخش اول به اشتباه به عنوان PHI دسته‌بندی شده‌اند مشخص شوند به عبارت دیگر خروجی بخش اول را به دو گروه مثبت‌های کاذب و مثبت‌های درست تقسیم‌بندی می‌کند بااین‌حال این سیستم نیز در تشخیص شناسه‌هایی نظیر آدرس و نام سازمان‌ها ضعیف عمل می‌کند [۴۲]. کارایی این سیستم در مقایسه با دو روش مبتنی بر داده HIDE و MIST در مقاله‌ای دیگر توسط Ferrández و همکاران بررسی شده است [۴۳].

دهقان و همکاران [۴۰] روشی برای شرکت در چالش i2b2NLP سال ۲۰۱۴ [۴۴] ارائه کرده که از ترکیب روش‌های مبتنی بر دانش (پایگاه قانون و لغت‌نامه‌ها) و روش مبتنی بر داده CRF که از ویژگی‌های املایی، موقعیتی، لغوی و معنایی بهره می‌برد، استفاده می‌کند. ابتدا پیش پردازش متون توسط cTAKES [۴۵] و GATE [۴۶] شامل مرزبندی

[Downloaded from jhbmi.com on 2026-11-15]

جملات، نشانه‌گذاری و غیره انجام می‌گیرد. سپس دو مدل مبتنی بر ML و دانش اجرا می‌شوند و پاسخ آن‌ها در سه مدل یکپارچه می‌گردد. نتایج نشان داد که رویکرد پیشنهادی در این مطالعه توانایی تشخیص ابهامات را ندارد برای مثال نمی‌تواند میان ملیت و زبان یک فرد تمایز قائل شود.

Yang و همکاران [۳۲] سیستم ترکیبی دیگری را معرفی می‌کنند که مانند مقاله دهقان از CRF، پایگاه قانون و لغت‌نامه برای تشخیص PHI از متون آزاد پزشکی استفاده می‌کند و از مجموعه داده‌های i2b2 NLP 2014 برای ارزیابی سیستم پیشنهادی بهره برده‌اند. سیستم ارائه شده شامل چهار ماژول پردازشی است: ۱) پیش‌پردازش داده‌ها (مانند نشانه‌گذاری به منظور به دست آوردن ویژگی‌هایی چون POS برای آموزش الگوریتم یادگیری ماشین. همچنین در این بخش یکسری ویژگی‌های مبتنی بر سند مانند عناوین بخش‌ها و ویژگی موقعیت جمله توسط یک مجموعه قوانین دستی استخراج می‌شوند). ۲) ایجاد ویژگی (بخشی از ویژگی‌هایی که برای آموزش الگوریتم یادگیری ماشین لازم هستند در گام اول ایجاد می‌شوند و بخشی در این مرحله. در این مقاله شش دسته ویژگی تعیین شده است: ویژگی‌های در سطح کلمه و جمله، ویژگی‌های املایی، سخنی (Discourse)، متنی و خاص وظیفه (Task-specific). ویژگی‌های خاص وظیفه شامل لیستی است از نام‌ها و مخففهای ایالات آمریکا، کشورها، زبان‌ها و ... و عباراتی که نشان‌دهنده وقوع یک PHI هستند مثل "Dr." و ارتباطات میان آن‌ها). ۳) مدل ترکیبی شناسایی PHIs (نویسندگان این مقاله برای تشخیص آن دسته از شناسه‌هایی که برای آن‌ها به قدر کافی نمونه آموزش وجود دارد از الگوریتم یادگیری ماشین CRF استفاده می‌کنند. در واقع چندین CRF هر کدام برای یک زیرمجموعه از PHIs آموزش می‌بینند. برای شناسه‌های باقیمانده از پایگاه قانون و لیست کلمات و عبارات منظم استفاده می‌شود). ۴) پس پردازش (در این مرحله تکنیک‌های زیادی برای ایجاد مجموعه واژگان PHI صحیح، با دو هدف به کار گرفته می‌شود: اصلاح کردن خطاهای مرحله شناسایی واژگان یا پیدا کردن PHIs بیشتر). این مطالعه انواع خطاها را به خوبی تحلیل کرده و نشان داد که سیستم در تشخیص برخی شناسه‌ها چون مکان، واژگان کم تکرار و خاص و حروف اختصاری با مشکل روبه‌رو است.

در مقاله Liu و همکاران [۴۷] از دو نوع طبقه‌بند برای دسته‌بندی PHIs استفاده شده است: مبتنی بر پایگاه قانون و مبتنی بر یادگیری ماشین. دو طبقه‌بند مبتنی بر یادگیری ماشین از دو نوع CRF استفاده می‌کنند، اولی بر مبنای ویژگی‌هایی در سطح نشانه (مثل POS، ویژگی‌های املایی و نمایش کلمات و غیره) و دومی بر مبنای ویژگی‌های کارکتری که از ویژگی‌هایی همانند ویژگی‌های نشانه‌ای استفاده می‌کند؛ اما در این CRF جمله به بخش‌های جزئی‌تری (کاراکترها) تجزیه می‌شود. طبقه‌بند مبتنی بر قوانین از عبارات منظم به منظور استخراج PHIs استاندارد مثل شماره تلفن، شماره پرونده پزشکی و غیره بهره می‌برد. در نهایت خروجی این سه طبقه بند با استفاده از یک پایگاه قانون یکپارچه می‌شود: آن دسته از PHIs که در بین سیستم‌ها مشترک نیستند مستقیماً به عنوان خروجی کل سیستم در نظر گرفته می‌شوند؛ اما برای نمونه PHIs همپوشان اولویت اول با خروجی طبقه‌بند مبتنی بر عبارات منظم است و بعد از آن به ترتیب اولویت با خروجی طبقه بندهای مبتنی بر کاراکتر و مبتنی بر نشانه است.

Anonym نام نرم‌افزاری است برای شناسه‌زدایی خودکار EHR که توسط مرکز تحقیقات سلامت از راه دور استرالیا توسعه یافته است. Anonym دارای سه ماژول است: ماژول ایجاد ویژگی (استخراج ویژگی‌های زبانی و لغوی)، ماژول یادگیری مدل که از ویژگی‌های استخراج شده در گام اول بهره می‌برد و ماژول دسته‌بندی که مبتنی بر عبارات منظم و CRF می‌باشد. عملکرد این سیستم با سه مجموعه داده مورد ارزیابی قرار گرفته است: مجموعه داده i2b2 NLP 2006، مجموعه داده MTSpamles که شامل ۱۸۸۵ یادداشت بالینی است که به صورت دستی در دانشگاه کالیفرنیا حاشیه‌گذاری شده‌اند و مجموعه سوم که شامل ۸۵۲ گزارش پاتولوژی و سیتولوژی مرکز تحقیقات سرطان NSW (New South Wales) می‌باشد. این گزارش‌ها کاغذی بوده و با استفاده از فناوری تشخیص نوری کارکترها (Optical Character Recognition) OCR به نسخه الکترونیک تبدیل شده‌اند به همین دلیل مانند دو مجموعه قبل کامل نیستند. حاشیه‌گذاری این مجموعه توسط نویسندگان مقاله انجام شده است؛ اگرچه این سیستم در میان تیم‌های شرکت کننده در چالش i2b2NLP در سال ۲۰۰۶ بهترین عملکرد را ارائه داده است؛ اما نتایج این مقاله نشان داد که اگر مجموعه داده آموزش و آزمایش متفاوت باشند این سیستم عملکرد ضعیفی دارد [۴۸].

بحث و نتیجه‌گیری

مزیت روش‌های مبتنی بر دانش که از لغت‌نامه یا عبارات منظم استفاده می‌کنند این است که به داده‌ای برای آموزش نیاز

[Downloaded from jhbmi.com on 2026-11-15]

ندارند. همچنین به روز کردن آن‌ها به راحتی و از طریق اضافه کردن یکسری عبارت منظم/ واژه/الگو امکان‌پذیر است. مهم‌ترین ایرادی که به این الگوریتم‌ها وارد می‌شود تعمیم‌پذیری پایین آن‌ها است به‌عبارتی‌دیگر این امکان وجود دارد که فرد خبره‌ای مسلط به همه انواع PHI موجود در متن، وجود نداشته باشد. به علاوه برای اضافه کردن یک نوع PHI جدید یا برای توسعه دادن این روش‌ها برای انواع دیگری از متون کار زیادی بایستی انجام گیرد [۲۰]. در حالی که مزیت روش‌های مبتنی بر داده این است که آن‌ها می‌توانند به صورت خودکار الگوهای پیچیده تشخیص PHI را یاد بگیرند و درنتیجه نیازی نیست که توسعه‌دهندگان سیستم اطلاعات زیادی از انواع PHI داشته باشند. این سیستم‌ها تعمیم‌پذیری بیشتری نسبت به روش‌های مبتنی بر دانش دارند؛ اما بر خلاف آن‌ها توانایی کمی در تشخیص شناسه‌های کمیاب و نادر دارند. همچنین پیچیدگی محاسباتی و سرعت این روش‌ها در طول زمان افزایشی ندارد در حالی که در روش‌های مبتنی بر دانش اگر قرار بود آن‌ها برای حوزه یا نوع جدیدی از داده‌ها به کار گرفته شوند، پیچیدگی آن‌ها افزایش می‌یافت. مهم‌ترین عیب این روش‌ها نیاز به داشتن یک مجموعه به اندازه کافی بزرگ از داده‌های آموزش است. ایراد دیگری که بر این سیستم‌ها وارد است این است که در صورت وجود خطا تشخیص علت آن دشوار است. فرض کنید که الگوریتم موردنظر در یافتن اسم مربوط به یک محل خاص خوب عمل نمی‌کند نمی‌توان مطمئن بود که با اضافه کردن داده‌های آموزش این مشکل رفع می‌شود [۲۰،۳۷،۴۸]؛ با توجه به مزایا و معایب این دو رویکرد، روش‌های ترکیبی ایجاد شدند که همان‌طور که در جدول ۵ مشاهده می‌شود به نتایج خوبی دست یافته‌اند.

همان‌گونه که در نتایج این مقاله اغلب مقالات کارایی سیستم‌های پیشنهادی خود را بر اساس معیارهایی چون دقت، فراخوانی و معیار-F ارزیابی کرده‌اند. در واقع تمامی این معیارها به حذف هر چه بیشتر و درست‌تر انواع PHI تمرکز دارند؛ اما بدیهی است که با توجه به شرایط خاصی که هر مقاله در بخش تعریف مسئله بیان کرده و تأثیر آن‌ها بر اثربخشی رویکرد پیشنهادی، مقایسه این روش‌ها زیاد منطقی نیست.

یکی از عامل‌های تأثیرگذار بر نتایج حاصل از به کارگیری سیستم‌ها منابع دانشی است که در رویکردهای مختلف مورداستفاده قرارگرفته‌اند. آیا فرد خبره‌ای که برای کار برگزیده‌اند به همه انواع PHI آگاهی داشته است؟ از چه لغت‌نامه‌هایی و برای تشخیص چه نوع شناسه‌هایی بهره برده‌اند؟ مسئله دیگری که بر کارایی سیستم‌ها تأثیر می‌گذارد روشی که برای نمایش دانش (مانند عبارات منظم) یا استدلال (روش‌های طبقه‌بندی مختلف) انتخاب کرده‌اند می‌باشد. پیامد حاصل از این روش‌ها به شدت به مجموعه ویژگی‌هایی است که برای تشخیص PHIs استفاده می‌کنند. در برخی از این مقالات نتایج حاصل از یک الگوریتم با استفاده از مجموعه ویژگی‌های متفاوتی ارزیابی شده است [۴۷].

یکی از مهم‌ترین و مؤثرترین عامل‌ها در کارایی سیستم‌ها نوع متن است. در حقیقت از منظری دیگر نیز می‌توان مطالعات این حوزه را در دو گروه دسته‌بندی کرد و آن نوع متنی که شناسه زدایی شده می‌باشد. اکثر مقالات شناسه‌زدایی بر روی یک نوع متن مثل گزارش‌های پاتولوژی [۳۶] و خلاصه ترخیص [۴۱،۳۹،۳۵-۳۳]. تمرکز داشته‌اند، اما تعدادی از مطالعات نیز روشی را برای شناسه‌زدایی از هر نوع متنی معرفی کرده‌اند [۳۲،۳۷،۴۰،۴۲،۴۷]. بدیهی است که الگوریتم‌های معرفی شده در گروه اول چون از لغت‌نامه‌های خاص منظوره استفاده کرده‌اند تعمیم‌پذیری بالایی نداشته و نمی‌توان از آن‌ها برای شناسه‌زدایی از سایر متون نیز بهره برد این مسئله در مقاله Zuccon و همکاران بررسی شده است [۴۸]. همچنین در برخی از گزارش‌هایی که در مقالات خاص منظوره استفاده می‌گردد همه انواع شناسه وجود ندارد برای مثال عموماً کد پستی و آدرس محل زندگی در یک گزارش پاتولوژی جراحی ذکر نمی‌گردد. بر اساس جداول ۴ و ۵ مشاهده می‌شود خلاصه ترخیص و پرونده پزشکی (به طور کامل) درصد بیشتری از مطالعات را به خود اختصاص داده‌اند و علت این موضوع چالش‌های i2b2NLP در سال‌های ۲۰۰۶ و ۲۰۱۴ بوده است که در ادامه معرفی می‌شود.

به طور کلی، برای ارزیابی عملکرد روش‌های مختلف شناسه-زدایی بایستی به توسعه متون استانداردی پرداخت تا بر اساس آن بتوان نتایج گزارش شده توسط مقالات مختلف را مقایسه کرد. در این راستا گروهی به این موضوع (متن مورد استفاده) پرداخته‌اند و به ایجاد استانداردی طلایی که به عنوان ورودی سیستم‌های شناسه‌زدایی استفاده شود، تأکید دارند. برای مثال کار مشترک پردازش زبان طبیعی i2b2/UTHealth (The Informatics for Integrating Biology and the Bedside (i2b2) and the University of Texas Health Science Center at Houston (UTHealth) (natural language processing (NLP) shared task در سال ۲۰۱۴ [۴۴] چهار موضوع مطرح کرد که اولین مورد به بحث شناسه‌زدایی پرونده‌های پزشکی به منظور استفاده مجدد

[Downloaded from jhbmi.com on 2026-11-15]

از آن‌ها در تحقیقات و غیره تمرکز دارد. این گروه همچنین، مجموعه داده‌ای با بیش از ۱۳۰۰ پرونده پزشکی که نام‌های بدلی برای بیماران آن‌ها استفاده شده است را در اختیار افرادی که بخواهند سیستمی با این مضمون طراحی کنند قرار می‌دهد [۱۵].

در حقیقت پیش از i2b2 مقالات زیادی روی شناسه‌زدایی از متون بالینی کار کرده‌اند؛ اما نتایج آن‌ها با هم قابل مقایسه نبود. i2b2 بستری را برای مقایسه و ارزیابی بهتر سیستم‌های ارائه شده فراهم کرده است. Uzuner که یکی از مسئولان کار مشترک i2b2 NLP است در سال‌های مختلف مروری بر سیستم‌هایی که برای حل این چالش شرکت کرده‌اند، داشته است [۲۷،۲۳].

در این میان، مقالاتی به چالش‌های ایجاد استاندارد طلایی به شکل دستی نیز پرداخته‌اند [۴۹]. برای مثال Deleger و همکاران ابتدا یک گروه کامل از انواع PHI معرفی کرده‌اند و سپس تنها از دو حاشیه‌نویس به منظور حاشیه‌گذاری متون برای کشف انواع PHI، داروها، ارتباطات علائم و بیماری استفاده کرده‌اند [۵۰] هدف Mayer و همکاران [۵۱] نیز ساخت یک مرجع استاندارد برای شناسه‌زدایی مبتنی بر NLP از مجموعه متفاوتی از اسناد بالینی است. ابتدا دو دانشجو روی ۲۰ عدد سند آموزش می‌بینند و سپس ۲۲۰ سند دیگر را برچسب می‌زنند. موضوع مهمی که در این مقاله به آن پرداخته شده است اهمیت PHIs مختلف است. در واقع برای ارزیابی یک سیستم شناسه‌زدا باید به این نکته توجه شود که سیستم برای چه گروهی از PHIs خوب یا بد کارکرده است (به عبارت دیگر چه PHIs را پیدا نکرده است). از همین رو دو حاشیه‌نویس علاوه بر مشخص کردن PHI به آن‌ها یک رتبه هم اختصاص می‌دهند. سه رتبه‌ای که در این مقاله در نظرگرفته شده عبارت‌اند از: کم، متوسط و زیاد.

اگرچه بررسی نتایج گزارش شده در مقالات دید کلی از مزایا و معایب هر گروه از روش‌های شناسه‌زدایی مبتنی بر ML ایجاد می‌کند؛ اما به منظور مقایسه عملکرد آن‌ها بایستی همگی بر یک نوع متن اعمال شوند. علاوه بر این، باید توجه داشت که تمامی مقالات مورد مطالعه بر متونی که به زبان انگلیسی نوشته شده‌اند تمرکز دارند و همین امر موجب می‌شود تا نتوان به طور قطعی نتیجه گرفت که عملکرد آن‌ها برای سایر زبان‌ها نیز مشابه باشد به خصوص برای زبان‌هایی که ازلحاظ ساختاری و نحوی با انگلیسی متفاوت است مانند زبان‌های بدون فاصله ژاپنی و چینی یا زبان راست‌چین فارسی که نحوی متفاوت دارد.

هدف این مطالعه بررسی شناسه‌زدایی به عنوان راه‌حلی برای حفظ محرمانگی اطلاعات بیماران و تسهیل تحقیقات پزشکی است. در این مقاله مروری بر شناسه‌زدایی مبتنی بر یادگیری ماشین انجام شد و چالش‌های مطرح در به کارگیری و ارزیابی هر کدام از روش‌ها بررسی گشت. اگرچه تمامی مقالات بررسی شده تنها به پرونده‌های الکترونیک پزشکی که به زبان انگلیسی بودند، پرداخته‌اند؛ اما بسیاری از مسائلی که در آن‌ها مطرح شده بود برای طراحی سیستم شناسه‌زدایی به هر زبانی مشترک است. نتایج نشان می‌دهد که شناسه‌زدایی هرگز باعث حذف تمامی شناسه‌ها نمی‌شود و بنابراین در بررسی عملکرد آن‌ها باید به نوع شناسه‌هایی که در متن باقی می‌ماند توجه شود. نکته قابل‌توجه این است که با وجود این که هدف شناسه‌زدایی حذف شناسه‌ها است؛ اما باید به موضوعاتی چون احتمال خطر بازشناسی افراد، مناسب بودن داده بی‌نام شده برای تحقیقات بالینی و میزان آسیبی که از افشای انواع شناسه به افراد وارد می‌شود نیز توجه شود که نیازمند کار بیشتر خواهد بود.

تشکر و قدردانی

این مقاله حاصل بخشی از طرح پژوهشی به شماره ۲۷۸۸۱-۹۵-۰۲-۱۳۶ می‌باشد که با حمایت دانشگاه علوم پزشکی ایران انجام شده است.

تضاد منافع

بدین‌وسیله نویسندگان تصریح می‌نمایند که هیچ‌گونه تضاد منافعی در خصوص پژوهش حاضر وجود ندارد.

References

1. Faghihi M, Memarzadeh G, Astane H. Protection of patient privacy; A prerequisite for electronic health development. Journal of Medical Ethics 2010;4(2):163-88. Persian

2. Harman LB, Flite CA, Bond K. Electronic health records: privacy, confidentiality, and security. Virtual Mentor 2012;14(9):712-9.

3. Andriole KP. Security of electronic medical information and patient privacy: what you need to know. J Am Coll Radiol 2014;11(12 Pt B):1212-6.

4. Gozali E, Langarizadeh M, Sadooghi F, Sadeghi M. Letter to Editor: Electronic Medical Record, Step toward Improving the Quality of Healthcare Services and Treatment Provided to Patients. Journal of Ardabil University of Medical Sciences. 2014;14(1):93-6. Persian

[Downloaded from jhbmi.com on 2026-11-15]

5. Abdelhak M, Grostick S, Hanken MA. Health Information: Management of a Strategic Resource. 4th ed: Louis Missouri: Saunders; 2014.
6. Gozali E, Langarizadeh M, Sadoughi F. a survey of the possibility of electronic medical records implementation in teaching hospitals affliated to Urmia University of Medical Sciences. J Urmia Nurs Midwifery Fac 2013; 11(5):391-7. Persian
7. Pozgar GD, Santucci NM. Legal Aspects of Health Care Administration. USA: Jones & Bartlett Learning; 2003.
8. Farzandipoor M. Review on policies about medical records release in university hospitals in Tehran [dissertation]. Tehran: Iran University of Medical Sciences; 1995. Persian
9. Affairs Dov. Review of Issues Related to the Loss of VA Information Involving the Identity of Millions of Veterans. 2006. Available from: https://www.va.gov/oig/pubs/VAOIG-06-02238-163.pdf
10. Fernández-Alemán JL, Señor IC, Lozoya PÁ, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013;46(3):541-62.
11. Ghaderi NL, Yarmohammadian MH, Raeisi AR, Tavakoli N. Medical record information disclosure laws and policies for purpose law enforcement among selected countries Health Info Manage 2011; 8(3):335-44. Persian.
12. Annas GJ. HIPAA regulations - a new era of medical-record privacy? N Engl J Med 2003;348(15):1486-90.
13. Wager KA, Lee FW, Glaser JP. Health care information systems: a practical approach for health care management. 3th ed. San Francisco, CA, USA: John Wiley & Sons; 2013.
14. Government Publications Office. Health Insurance Portability And Accountability Act of 1996. [cited 2017 Mar 12]. Available from:https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf.
15. Stubbs A, Uzuner O. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform 2015;58 Suppl:S20-9.
16. Khorshidi E. Patient's Bill of Rights. [cited 2016 Oct 20]. Available from: http://www.e-khorshidi-lawyer.ir/index.php?ToDo=ShowArticles&AID=13145 Persian
17. Committee IIIT. IHE IT Infrastructure Handbook De-Identification 2014 [cited 2015 Sep 24]. Available from: http://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Handbook_De-Identification_Rev1.1_2014-06-06.pdf.
18. Dorr D, Phillips W, Phansalkar S, Sims S, Hurdle J. Assessing the difficulty and time cost of de-identification in clinical narratives. Methods Inf Med 2006;45(3):246-52.
19. Gozali E, Langarizadeh M, Sadoughi F. The Ability of Educational Hospitals Affiliated to Urmia University of Medical Sciences in Establishment of Electronic Medical Records from Organizational Perspective. Iranian Journal of Medical Informatics 2013;2(3):8-12.
20. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10:70.
21. South BR, Mowery D, Suo Y, Leng J, Ferrández Ó, Meystre SM, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform 2014;50:162-72.
22. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-44.
23. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550-63.
24. Fenz S, Heurix J, Neubauer T, Rella A. De-identification of unstructured paper-based health records for privacy-preserving secondary use. J Med Eng Technol 2014;38(5):260-8.
25. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ, editors. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. AMIA Annu Symp Proc 2014; 2014: 767–76.
26. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. Yearb Med Inform 2015 ;10(1):194-8.
27. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015;58 Suppl:S11-9.
28. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Medical Research Methodology 2012;12(1):109.
29. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb Med Inform 2015;10(1):183-93.
30. Jaeschke R, Guyatt GH, Sackett DL, Guyatt G, Bass E, Brill-Edwards P, et al. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994;271(9):703-7.
31. Hanauer DA, Mei Q, Malin B, Zheng K. Location Bias of Identifiers in Clinical Narratives. AMIA Annu Symp Proc 2013; 2013: 560–9.
32. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. J Biomed Inform 2015;58 Suppl:S30-8.
33. Uzuner Ö, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine 2008;42(1):13-35.

[Downloaded from jhbmi.com on 2026-11-15]

34. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007;14(5):574-80.
35. McMurry AJ, Fitch B, Savova G, Kohane IS, Reis BY. Improved de-identification of physician notes through integrative modeling of both public and private medical text. BMC Med Inform Decis Mak 2013;13:112.
36. Gardner J, Xiong L. HIDE: An Integrated System for Health Information DE-identification. 21st IEEE International Symposium on Computer-Based Medical Systems; 2008 Jun 17-19; Jyvaskyla, Finland: IEEE; 2008.
37. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010;79(12):849-59.
38. Li M, Carrell D, Aberdeen J, Hirschman L, Malin BA. De-identification of clinical narratives through writing complexity measures. Int J Med Inform 2014;83(10):750-67.
39. He B, Guan Y, Cheng J, Cen K, Hua W. CRFs based de-identification of medical records. J Biomed Inform 2015; 58(Suppl): S39–S46.
40. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J Biomed Inform. 2015;58 Suppl:S53-9.
41. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc 2007;14(5):564-73.
42. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013;20(1):77-83.
43. Ferrández Ó1, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Generalizability and comparison of automatic clinical text de-identification methods and resources. AMIA Annu Symp Proc 2012;2012:199-208.
44. 2014 NLP Shared Task [cited 2015 Sep 22]. Available from: https://www.i2b2.org/NLP/HeartDisease/Main.php.
45. 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.
46. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications; 2002 Jul 6-12; Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA: Association for Computational Linguistics; 2002. p. 168-75.
47. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015;58 Suppl:S47-52.
48. Zuccon G, Kotzur D, Nguyen A, Bergheim A. De-identification of health records using Anonym: effectiveness and robustness across datasets. Artif Intell Med 2014;61(3):145-51.
49. Browne AC, Kayaalp M, Dodd ZA, Sagan P, McDonald CJ. The Challenges of Creating a Gold Standard for De-identification Research. AMIA Annu Symp Proc 2014;2014:353-8.
50. Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annu Symp Proc 2012;2012:144-53.
51. Mayer J, Shen S, South BR, Meystre S, Friedlin FJ, Ray WR, et al. Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes. AMIA Annu Symp Proc 2009; 2009: 416–20.

[Downloaded from jhbmi.com on 2026-11-15]

Review Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2017; 4(2): 154-167

De-identification of Electronic Health Records Using Machine Learning Algorithms

Langarizadeh Mostafa[1], Orooji Azam[2]*

• **Received:** 17 Jul, 2017 • **Accepted:** 2 Sep, 2017

Introduction: Electronic Health Record (EHR) contains valuable clinical information that can be useful for activities such as public health surveillance, quality improvement, and research. However, EHRs often contain identifiable health information that their presence limits the use of the records for sharing and secondary usages. De-identification is one of the common methods for protecting the confidentiality of patient information. This systematic review has focused on recently published studies on the usage of de-identification methods based on Machine Learning (ML) approaches for removing all identifiable information from electronic health records.

Methods: A systematic review was performed in electronic databases like PubMed and ScienceDirect between 2006 and 2016. Studies were assessed for adherence to the CASP checklists and reviewed independently by two investigators. Finally, 12 articles were matched with inclusion criteria.

Results: The selected studies have been discussed in terms of used methods and knowledge resources, types of identifiers detected, types of clinical documents, challenges and achieved results. The results showed that ML-based de-identification is a widely invoked approach to protect patient privacy when disclosing clinical data for secondary purposes, such as research. Also, the combination of the ML algorithms and some techniques such as pattern matching and regular expression matching could decrease need to train data.

Conclusion: There is a lot of identifiable information in medical records. This study showed ML-based de-identification methods can intensively reduce the disclosure risk of information.

Keywords: Confidentiality, Privacy, De-identification, Machine Learning

• **Citation:** Langarizadeh M, Orooji A. De-identification of Electronic Health Records Using Machine Learning Algorithms. Journal of Health and Biomedical Informatics 2017; 4(2): 154-167.

1. Ph.D. of Medical Informatics, Health Information Management Dept., School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.
2. Ph.D Student of Medical Informatics, Health Information Management Dept., School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.

***Correspondence**: No. 6, Rashid Yasemi Av. Vali-e-Asr St., Vanak Sq., Tehran, Iran.
• **Tel:** 02188794301 • **Email:** orooji.a@tak.iums.ac.ir

[Downloaded from jhbmi.com on 2026-11-15]