مقاله پژوهشی

مدلی برای تشخیص بیماری‌های کبدی با استفاده از روش‌های یادگیری ماشینی

حمیدرضا طهماسبی*[1]، رضا بشارتی[2]، محمد علیشاهی[3]

تاریخ دریافت ۱۴۰۱/۰۵/۰۹ تاریخ پذیرش ۱۴۰۲/۰۱/۳۰

چکیده

پیش‌زمینه و هدف: تشخیص به‌موقع بیماری‌های کبدی تأثیر قابل‌توجهی در پیشگیری از عوارض آن‌ها و همچنین کنترل و درمان بیماری دارد. یادگیری ماشینی یکی از شاخه‌های هوش مصنوعی است که کاربردهای زیادی در زمینه تشخیص پزشکی دارد. این مطالعه باهدف ارائه‌ی مدلی با دقت و اعتماد بالاتر برای تشخیص بیماری‌های کبدی با استفاده از روش‌های یادگیری ماشینی انجام شد که بتواند به متخصصان پزشکی در تشخیص و کنترل به‌موقع بیماری‌های کمک کند.

مواد و روش کار: این مطالعه از نوع کاربردی-توسعه‌ای بوده و از مجموعه داده‌های ۵۸۳ بیمار کبدی استفاده شده است. برای تشخیص دقیق‌تر افراد مبتلا به بیماری‌های کبدی، نتایج سه روش یادگیری ماشینی پرکاربرد در تشخیص پزشکی شامل ماشین بردار پشتیبان، جنگل تصادفی و شبکه‌های عصبی مصنوعی با استفاده از نظریه‌ی ترکیب دمپستر-شافر با هم ترکیب شده است. از نرم‌افزار داده‌کاوی Weka و همچنین زبان برنامه‌نویسی پایتون برای پیاده‌سازی مدل استفاده شد. برای ارزیابی کارایی، روش ارزیابی متقابل k تکه برابر بکار برده شد.

یافته‌ها: نتایج نشان داد که دقت، حساسیت و ویژگی در مدل پیشنهادی به ترتیب ۹۱/۴۷ درصد، ۸۹/۵۲ درصد و ۹۳/۰۳ درصد بوده و در مقایسه با مطالعات مشابه، عملکرد بهتری دارد.

بحث و نتیجه‌گیری: مدل پیشنهادی در جامعه‌ی آماری موردمطالعه، عملکرد بهتری در تشخیص بیماری‌های کبدی داشته و می‌تواند به پزشکان در تشخیص زودهنگام این بیماری‌ها و انجام درمان مناسب در مراحل اولیه کمک کرده و درنتیجه مانع از پیشرفت بیماری شود.

کلیدواژه‌ها: طبقه‌بندی، تشخیص، بیماری‌های کبدی، یادگیری ماشینی

مجله مطالعات علوم پزشکی، دوره سی و سوم، شماره یازدهم، ص ۸۲۲-۸۱۴، بهمن ۱۴۰۱

آدرس مکاتبه: خراسان رضوی، کاشمر، دانشگاه آزاد اسلامی واحد کاشمر، تلفن: ۰۹۱۵۱۰۴۶۱۱۷

Email: htahma@gmail.com

مقدمه

کبد دومین عضو داخلی مهم بدن انسان هست که نقش بسزایی در متابولیسم ایفا می‌کند (۱) و چندین عملکرد حیاتی ازجمله دفع مواد زائد، سم‌زدایی از مواد شیمیایی، تنظیم سوخت‌وساز قند و چربی را به عهده دارد. عواملی از قبیل عفونت ویروسی، مصرف بیش‌ازحد دارو، مسمومیت، سوءمصرف الکل، و چاقی می‌توانند باعث بروز بیماری‌های مختلف کبدی شوند. بیماری‌های کبدی با ایجاد نارسایی کبد، به بدن آسیب جدی رسانده و منجر به عملکرد نامناسب بدن و حتی مرگ بیمار می‌شوند (۲،۳). امروزه افراد مبتلا به این بیماری‌ها به‌طور مداوم در حال افزایش هستند (۴، ۵). بیماری‌های کبدی یکی از چالش‌های مهم حوزه‌ی سلامت و پزشکی در جهان به شمار می‌آیند (۵، ۶، ۷) و در سال‌های اخیر تعداد مرگ‌ومیرهای ناشی از اختلالات کبدی رشد قابل‌توجهی داشته است (۶). تشخیص علائم این بیماری‌ها در مراحل اولیه دشوار است (۶، ۸، ۹، ۱۰). زیرا بسیاری از افرادی که دچار اختلال و آسیب کبدی می‌باشند، در ظاهر احساس سلامتی می‌کنند. در این وضعیت، کبد به عملکرد طبیعی خود ادامه می‌دهد تا زمانی که به‌شدت آسیب ببیند (۱۰). تشخیص به‌موقع این بیماری‌ها تأثیر قابل‌توجهی در پیشگیری از عوارض آن و همچنین کنترل و درمان بیماری دارد (۱۱، ۱۲). علاوه بر این، تشخیص و درمان نامناسب بیماری‌های کبدی توسط متخصصان پزشکی، گاهی وقت و پول را هدر می‌دهد

[1] استادیار، گروه مهندسی کامپیوتر، واحد کاشمر، دانشگاه آزاد اسلامی، کاشمر، ایران (نویسنده مسئول)

[2] استادیار، گروه پرستاری، واحد کاشمر، دانشگاه آزاد اسلامی، کاشمر، ایران

[3] استادیار، گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

و حتی ممکن است باعث مرگ بیمار شود (۱۰). بنابراین ایجاد مدلی با دقت بالا که بتواند بیماری‌های کبدی را در مراحل اولیه تشخیص دهد، می‌تواند به پزشکان در تشخیص به‌موقع و انجام درمان مناسب در مراحل اولیه کمک کرده و درنتیجه مانع از رشد بیماری گردد.

امروزه داده‌کاوی و فن‌های یادگیری ماشینی به‌عنوان یک ابزار مهم برای تشخیص بیماری‌ها موردتوجه می‌باشند (۱۳، ۱۴). مطالعات انجام‌شده نشان می‌دهند که فن‌های طبقه‌بندی در داده‌کاوی و یادگیری ماشینی، برای تشخیص بیماری‌ها مفید بوده (۶، ۱۳، ۱۴) و همواره توسعه‌ی این الگوریتم‌ها به‌منظور پیش‌بینی دقیق‌تر و واضح‌تر بیماری‌ها از اهمیت خاصی برخوردار بوده است. به‌طورکلی استفاده از فقط یک طبقه‌بندی در طبقه‌بندی داده‌ها به دلیل ساختار ساده و سرعت محاسباتی سریع آن، موردتوجه زیادی قرار گرفته است (۱۵). از طرفی در طراحی مدل‌های تشخیص بیماری با کارایی بالا، بهبود دقت فن طبقه‌بندی بسیار مهم است. پژوهشگران نشان داده‌اند که ترکیب طبقه‌بندی‌ها می‌تواند به بهبود دقت در مدل تشخیص بیماری کمک می‌کند و دقت بالاتری در مقایسه با هر یک از طبقه‌بندی‌های استفاده شده، حاصل گردد (۸، ۱۵). در یک مدل ترکیبی، نتیجه‌ی نهایی طبقه‌بندی بر اساس ترکیب نتایج خروجی طبقه‌بندی‌های مختلف حاصل می‌شود.

در سال‌های اخیر پژوهش‌های متعددی درزمینه‌ی پیش‌بینی و تشخیص بیماری کبدی با استفاده از فن‌های طبقه‌بندی انجام شده است. با توجه به اهمیت دقت و اعتماد در سیستم‌های تشخیص بیماری‌های کبدی، توسعه و ارائه‌ی مدل‌هایی باهدف بهبود کارایی این سیستم‌ها بسیار موردتوجه است (۱۵، ۱۶). هدف این مطالعه، ارائه‌ی مدلی برای تشخیص بیماری کبد است که علاوه بر کمک به تصمیم‌گیری‌های بالینی و کاهش خطاها و همچنین صرفه‌جویی در هزینه‌های آزمون‌های تشخیصی، از اعتماد و دقت قابل قبولی برخوردار باشد. در مدل پیشنهادی، ابتدا داده‌های مفقودشده به روش ارائه‌شده توسط طهماسبی و همکاران (۱۷) با مقدار مناسب جایگزین می‌شوند، و سپس متوازن‌سازی داده‌ها و همچنین انتخاب ویژگی‌های مهم و تأثیرگذار داده‌ها بر روی بیماری انجام می‌شود. در مرحله‌ی بعد، با اعمال سه طبقه‌بندی متداول در تشخیص بیماری (۱، ۱۸) شامل ماشین بردار پشتیبان (Support Machine Vector)، جنگل تصادفی (Random Forest) و شبکه‌های عصبی مصنوعی روی مجموعه‌ی داده‌ها، نتایج حاصل از آن‌ها با استفاده از نظریه‌ی ترکیب شواهد دمپستر-شافر (Dempster-Shafer) (۱۹) ترکیب می‌شوند. نظریه‌ی دمپستر-شافر، تعمیم‌یافته‌ی احتمال بیزین هست که خصوصیات بازیابی صریح عدم قطعیت و قاعده‌ی ترکیب شواهد آن، باعث استفاده‌ی آن در مدل پیشنهادی گردیده است.

مواد و روش کار

این پژوهش، یک مطالعه‌ی کاربردی-توسعه‌ای برای پیش‌بینی و تشخیص بیماری کبد است. جامعه‌ی آماری، مجموعه داده‌ی بیماران کبدی هند (ILPD : Indian Liver Patient Dataset) است که در مخزن داده‌ی دانشگاه ایروین کالیفرنیا (۲۰) در دسترس بوده و در اغلب مطالعات مشابه درزمینه‌ی تشخیص بیماری کبدی، از این مجموعه داده استفاده شده است (۴، ۶، ۸، ۱۰، ۲۱، ۲۲، ۲۳، ۲۴). مجموعه داده‌ی ILPD، شامل اطلاعات ۱۱ ویژگی مربوط به ۵۸۳ نمونه است که ۴۱۶ نمونه از آن‌ها بیمار کبدی و ۱۶۷ نمونه، بیمار غیر کبدی هستند. در این مطالعه، نمونه‌های عدم مبتلا به بیماری‌های کبدی را سالم نام‌گذاری می‌کنیم. جدول (۱) این ویژگی‌ها را نشان می‌دهد. ۱۰ ویژگی بیانگر علائم بیماری بوده و یک ویژگی نتیجه‌ی تشخیص است که ابتلا یا عدم ابتلای نمونه به بیماری‌های کبدی را بر اساس علائم آن نمونه نشان می‌دهد. (۱: ابتلا به بیماری‌های کبدی، ۲: سالم یا عدم ابتلا به بیماری‌های کبدی).

در مدل پیشنهادی، نتایج طبقه‌بندی حاصل از سه الگوریتم طبقه‌بندی متداول مورداستفاده در تشخیص بیماری شامل ماشین بردار پشتیبان (SVM)، جنگل تصادفی (RF) و شبکه‌های عصبی مصنوعی چندلایه (MLP) به کمک نظریه‌ی ترکیب شواهد دمپستر-شافر با هم ترکیب می‌شوند تا تصمیم نهایی طبقه‌بندی بر اساس اطلاعات ترکیب شده گرفته شود و یک روش دقیق‌تر برای تشخیص بیماری کبدی حاصل گردد. این مدل از دو مرحله یا بخش تشکیل می‌شود. مرحله‌ی اول پیش‌پردازش داده‌ها و مرحله‌ی دوم طبقه‌بندی و تشخیص بیماری کبدی است.

در مرحله‌ی پیش‌پردازش داده‌ها، مقدار ویژگی A/G ratio در چهار نمونه از مجموعه داده، نامشخص و مفقود شده است. این مقادیر با استفاده از روش جانشینی Mean (۱۷) جایگزین شدند. همچنین در این مجموعه داده، تعداد نمونه‌های مبتلا به بیماری کبدی نسبت به تعداد نمونه‌های سالم بسیار بیشتر است و درنتیجه همانند اغلب داده‌های پزشکی با عدم توازن مواجه است. نرخ عدم توازن یا نسبت تعداد نمونه‌های طبقه‌ی اکثریت به تعداد نمونه‌های طبقه‌ی اقلیت در مجموعه داده برابر ۲/۴۹ است. این عدم توازن در داده‌ها می‌تواند دقت طبقه‌بندی را تحت تأثیر قرار دهد و معمولاً نمونه‌های متعلق به طبقه‌ی اقلیت به‌عنوان نمونه‌های طبقه‌ی اکثریت دسته‌بندی می‌شوند (۶، ۱۰، ۲۴). بدین منظور همانند اغلب مطالعات برای متوازن‌سازی داده‌ها از فن معروف بیش نمونه‌برداری اقلیت مصنوعی (SMOTE) (۱۸) استفاده شد. در این فن برای طبقه‌ی اقلیت، نمونه‌های جدیدی در همسایگی نمونه‌های موجود در این طبقه به‌صورت مصنوعی تولید می‌شود. درنتیجه تعداد

نمونه‌های متعلق به دوطبقه‌ی بیمار و غیر بیمار، متوازن می‌شوند. با این روش، ۱۶۷ نمونه‌ی جدید از نوع سالم در همسایگی نمونه‌های سالم موجود، تولیدشده و درنتیجه تعداد نمونه‌های سالم به ۳۳۴ نمونه افزایش یافت. برای انتخاب ویژگی‌های مهم و تأثیرگذار در تشخیص بیماری، از روش مبتنی بر فیلتر استفاده شد (۲۵). در این روش همانند مطالعه‌ی انجام شده توسط Joloudari و همکاران (۱)، با وزن‌گذاری ویژگی‌ها بر اساس بهره‌ی اطلاعاتی (Information gain) در طبقه‌بندی‌های RF و MLP، و وزن‌گذاری ویژگی‌ها بر اساس SVM برای طبقه‌بندی SVM ، به ترکیب‌های مختلف از ویژگی‌ها امتیاز داده‌شده و بر اساس آن زیرمجموعه‌ی مهم از ویژگی‌ها انتخاب شدند. بر این اساس از ۱۰ ویژگی در مجموعه داده‌ها، به‌جز ویژگی جنسیت، ۹ ویژگی دیگر برای طبقه‌بندی RF انتخاب شدند. برای طبقه‌بندی‌های MLP و SVM، همه‌ی ۱۰ ویژگی انتخاب شدند.

جدول (۱): مشخصات ویژگی‌ها در مجموعه داده‌ی بیماران کبدی هند (ILPD)

ویژگی	توضیحات	بازه‌ی مقادیر
Age	سن به سال	بین ۴ تا ۹۰ سال
Gender	جنسیت	زن/ مرد
TB	مجموع بیلی‌روبین (Total Bilirubin)	۰/۴ – ۷۵
DB	بیلی‌روبین مستقیم (Direct Bilirubin)	۰/۱ – ۱۹/۷
Alkphos	آلکالین فسفاتاز (Alkaline Phosphotase)	۶۳ - ۲۱۱۰
Sgpt	آمینوترانسفراز آلامین(Alamine Aminotransferase)	۱۰ - ۲۰۰۰
Sgot	اسپارتات ترانس آمیناز (Aspartate Aminotransferase)	۱۰ - ۴۹۲۹
TP	مجموع پروتئین‌ها (Total Protiens)	۲/۷ – ۹/۶
ALB	آلبومین (Albumin)	۰/۹ – ۵/۵
A/G ratio	نسبت آلبومین و نسبت گلوبولین (Albumin and Globulin Ratio)	۰/۳ – ۲/۸
Selector field	تشخیص (۱: بیمار کبدی، ۲: غیر بیمار کبدی)	۱ و ۲

پس از پیش‌پردازش داده‌ها، در مرحله‌ی طبقه‌بندی و تشخیص بیماری، ابتدا هر یک از سه طبقه‌بندی، به‌صورت مجزا طبقه‌بندی داده‌ها را انجام داده و سپس نتایج آن‌ها با هم ترکیب می‌شوند. خروجی هر طبقه‌بندی یکی از دو کلاس ابتلا به بیماری کبد و سالم (عدم ابتدا به بیماری کبد) است. برای اعمال طبقه‌بندی‌های SVM، RF و MLP روی مجموعه داده‌ی کبدی ILPD، از ماژول‌های نرم‌افزار داده‌کاوی متن‌باز Weka نسخه‌ی ۳،۷،۸ (۲۶) با همان پارامترهای پیش‌فرض استفاده شد.

با توجه به خصوصیات بازیابی صریح عدم قطعیت و قاعده‌ی ترکیب شواهد در نظریه‌ی ترکیب شواهد دمپستر–شافر، از این نظریه برای ترکیب نتایج خروجی طبقه‌بندی‌ها استفاده شد. به‌طوری‌که نتایج خروجی طبقه‌بندی‌های مستقل را به‌عنوان شواهد در نظر گرفته و با روی هم‌گذاری آن‌ها، یک تابع باور به دست می‌آورد که متناظر با خروجی‌های طبقه‌بندی‌ها است. در نظریه‌ی شواهد دمپستر - شافر، باور (Belief) مقداری است که برای بیان قطعیت یک گزاره یا رویداد به کار می‌رود. قاعده‌ی ترکیب شواهد دمپستر، دو بدنه شواهد مستقل تعریف‌شده در یک چارچوب مشاهدات را با هم ترکیب کرده و به یک بدنه شواهد تبدیل می‌کند (۱۳).

در مدل ترکیبی پیشنهادی با سه طبقه‌بندی و M طبقه، خروجی طبقه‌بندی‌ها به‌عنوان شواهد در نظر گرفته‌شده و طبقه‌ها به‌عنوان چارچوب مشاهدات محسوب می‌شوند. در صورت وجود عدم قطعیت برای تعیین طبقه‌ی نمونه X ، این نمونه به هیچ طبقه‌ای تعلق نمی‌یابد. طبقه‌ی با بزرگ‌ترین مقدار باور، به‌عنوان طبقه‌ی نمونه X تعیین می‌گردد. با توجه به اینکه برای تشخیص بیماری کبدی شامل دوطبقه‌ی ابتلا به بیماری و سالم است، مقدار M برابر ۲ است. در مدل پیشنهادی، پس از اعمال هر یک از طبقه‌بندی‌ها روی مجموعه داده‌ها، خروجی‌های دو طبقه‌بندی RF و MLP به‌عنوان شواهد با استفاده از نظریه‌ی ترکیب دمستر–شافر ترکیب شده و نتایج حاصل نیز با خروجی طبقه‌بندی SVM با استفاده از نظریه‌ی مذکور، ترکیب شد. روش ترکیبی با زبان برنامه‌نویسی پایتون پیاده‌سازی گردید.

به‌منظور بررسی کارایی مدل، همانند بسیاری از مطالعات دیگر (۴، ۶، ۸، ۱۰، ۲۲، ۲۳) از فن ارزیابی متقابل k تکه‌ی برابر (k-fold

cross validation) با مقدار k برابر ۱۰ استفاده شده است. در بررسی کارایی، معیارهای ارزیابی دقت (Accuracy)، حساسیت (Sensitivity) و ویژگی (Specificity) مورداستفاده قرار گرفت. معیار دقت، بیانگر دقت مدل در پیش‌بینی درست بیماری، معیار حساسیت، نسبت افراد درست بیمار تشخیص داده‌شده توسط مدل به کل افراد بیمار، و معیار ویژگی، نسبت افراد درست سالم تشخیص داده‌شده توسط مدل، به کل افراد سالم است. مقادیر این معیارها با استفاده از ماتریس تداخل (Confusion matrix) (جدول ۲) و از طریق روابط زیر به دست می‌آیند:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN+UP+UN} \quad (۱)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (۲)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (۳)$$

جدول (۲) : ماتریس تداخل

		نتیجه پیش‌بینی‌شده		
		بیمار کبدی	سالم	عدم قطعیت
نتیجه واقعی	بیمار کبدی	TP(True Positive)	FN(False Negative)	UP
	سالم	FP(False Positive)	TN(True Negative)	UN

TP به معنی تعداد نمونه‌های بیمار کبدی هستند که توسط مدل نیز بیمار تشخیص داده می‌شوند. FN: تعداد نمونه‌های بیمار کبدی هستند که توسط مدل سالم تشخیص داده می‌شوند. FP: تعداد نمونه‌های سالمی هستند که توسط مدل، بیمار کبدی تشخیص داده می‌شوند. TN: تعداد نمونه‌های سالمی هستند که توسط مدل نیز سالم تشخیص داده می‌شوند. ازآنجایی‌که در مدل پیشنهادی در صورت وجود عدم قطعیت، نمونه به هیچ کلاسی تعلق نمی‌یابد، دو حالت UP و UN نیز در ماتریس تداخل منظور گردیده است. UP: تعداد نمونه‌هایی که بیمار کبدی بوده‌اند و به کلاسی تعلق نگرفته‌اند و UN: تعداد نمونه‌هایی که سالم بوده‌اند و به کلاسی تعلق نگرفته‌اند.

برای بررسی معنادار بودن میزان بهبود دقت در مدل پیشنهادی نسبت به سایر روش‌های مقایسه شده ازنظر آماری، از آزمون آماری t زوجی[1] با سطح معنی‌داری ۰/۰۵ ($\alpha = 0.05$) استفاده شده است. این آزمون بین مقادیر دقت مدل پیشنهادی و هر یک از روش‌های دیگر به‌صورت جداگانه انجام شد.

یافته‌ها

جدول(۳)، معیارهای مربوط به کارایی شامل دقت، حساسیت و ویژگی به‌دست‌آمده برای هر یک از روش‌های طبقه‌بندی استفاده شده در ترکیب و همچنین مدل پیشنهادی را نشان می‌دهد. مقادیر پررنگ در جدول بیانگر بیشترین مقدار هستند.

جدول (۳) : کارایی مدل پیشنهادی و هر یک از طبقه‌بندی‌های استفاده شده در ترکیب (درصد)

روش	حساسیت	ویژگی	دقت
RF	۸۲/۴۵	۷۴/۵۵	۷۸/۹۳
MLP	۷۷/۱۶	۶۳/۱۷	۷۰/۹۳
SVM	۸۷/۵	۵۰/۷۲	۷۲/۱۳
مدل پیشنهادی	۹۳/۰۳	۸۹/۵۲	۹۱/۴۷

جدول(۴) مقایسه‌ی کارایی مدل پیشنهادی با چند روش جدید برای تشخیص بیماری کبدی در مجموعه داده‌ی ILPD را نشان می‌دهد. مقدار معیار ویژگی در سه تا از روش‌های مقایسه شده، توسط ارائه‌دهندگان این روش‌ها محاسبه نشده است که در این جدول نیز درج نشده‌اند.

[1] Paired t-test

جدول (۴) : مقایسه کارایی روش پیشنهادی با روش‌های جدید دیگر (درصد)

روش	حساسیت	ویژگی	دقت
Kumar و Thakur (۶)	۹۰/۰۳	۸۴/۷۹	۷۱/۸۷
Murugesan و همکاران (۲۴)	۸۴/۲۱	۶۳/۴۱	۷۰
Kumar و Thakur (۱۰)	۹۲/۷۵	۹۲/۷۵	۹۰/۶۵
Sreejith و همکاران (۲۲)	۸۳/۸۸	۸۰/۸۴	۸۲/۴۶
Sharma و Mehrotra (۸)	۸۳	-	۸۳/۹
Li و همکاران (۲۳)	۷۷/۴۸	-	۸۳/۰۶
فتحی و همکاران (۴)	۸۹/۲	-	۹۰/۹
مدل پیشنهادی	۹۳/۰۳	۸۹/۵۲	۹۱/۴۷

مقادیر value-p به‌دست‌آمده‌ی حاصل از آزمون t نیز در جدول (۵) مشاهده می‌شود.

جدول (۵) : مقادیر value-p آزمون test-t بین دقت مدل پیشنهادی و روش‌های مورد مقایسه

روش	value-p
RF	۰/۰۰۱۲
MLP	۰/۰۰۴۳
SVM	۰/۰۰۷۱
Kumar و Thakur (۶)	۰
Murugesan و همکاران (۲۴)	۰/۰۰۰۲
Kumar و Thakur (۱۰)	۰/۰۰۳۵
Sreejith و همکاران (۲۲)	۰/۰۰۵۴
Sharma و Mehrotra (۸)	۰
Li و همکاران (۲۳)	۰
فتحی و همکاران (۴)	۰/۰۰۵۹

بحث و نتیجه‌گیری

در سال‌های اخیر به‌کارگیری یادگیری ماشینی و روش‌های طبقه‌بندی به‌منظور پیش‌بینی و تشخیص بیماری‌های کبدی موردتوجه قرار گرفته است. Wu و همکاران با مقایسه و ارزیابی روش‌های مختلف طبقه‌بندی شامل جنگل تصادفی (RF)، بیزین ساده (NB)، شبکه عصبی مصنوعی (NNA) و رگرسیون لجستیک (LR) برای تشخیص بیماری کبد چرب، پیشنهاد کردند که مدل RF بهتر از سایر روش‌های طبقه‌بندی می‌تواند به پزشکان در تشخیص این بیماری کمک کند (۲۷). دقت مدل RF در این مطالعه برابر ۸۶/۴۸ درصد بوده است. مطالعه‌ی مشابهی نیز توسط Ghosh و همکاران انجام شده است (۲۱). در این مطالعه با مقایسه و ارزیابی روش‌های طبقه‌بندی رگرسیون لجستیک، جنگل تصادفی، XGBoost، ماشین بردار پشتیبان (SVM)، Adaboost، نزدیک‌ترین k همسایه و درخت تصمیم، روش RF به‌عنوان روشی که با دقت بهتری (دقت برابر ۸۳/۷ درصد) می‌تواند بیماری کبد را تشخیص دهد، پیشنهاد شده است. فتحی و همکاران روشی مبتنی بر ماشین بردار پشتیبان برای تشخیص بیماری کبد ارائه کرده‌اند (۴). آن‌ها ابتدا ویژگی‌های مهم و تأثیرگذار در بیماری را به روش انتخاب ویژگی معکوس (Backward selection) انتخاب کردند و سپس سه مدل طبقه‌بندی SVM خطی، SVM درجه دوم و SVM گاوسی برای تشخیص بیماران در دو مجموعه داده‌ی کبدی پیشنهاد

کردند. ارزیابی آن‌ها نشان داده است که مدل SVM گاوسی در هر یک از مجموعه داده‌ها دقتی به ترتیب برابر ۹۰/۹ درصد و ۹۲/۲ درصد داشته و در مقایسه با دو مدل طبقه‌بندی دیگر و همچنین سایر روش‌های مقایسه شده، از عملکرد بهتری برخوردار بوده است.

Sreejith و همکاران یک سیستم پشتیبان تصمیم بالینی ارائه کرده‌اند که از روش طبقه‌بندی RF برای تشخیص بیماری کبد استفاده می‌کند (۲۲). با توجه به اینکه اغلب داده‌های پزشکی ازجمله کبد با مسئله‌ی عدم توازن داده‌ها مواجه هستند (۱، ۶)، این سیستم ابتدا از فن SMOTE برای متوازن کردن داده‌ها استفاده می‌کند. همچنین از یک روش توسعه‌یافته‌ی مبتنی بر wrapper (۱) به نام بهینه‌سازی چند نظمی آشفته (CMVO) برای انتخاب ویژگی‌های مهم استفاده می‌کند. نتایج نشان داده‌اند که دقت این سیستم در تشخیص بیماری کبد برابر ۸۲/۴۶ درصد بوده است. Kumar و Thakur روشی مبتنی بر طبقه‌بندی نزدیک‌ترین k همسایه‌ی فازی برای تشخیص بیماری کبدی ارائه کرده‌اند (۶). این روش داده‌ها را با وزن‌گذاری متوازن می‌کند. به‌طوری‌که وزن زیادی را به همسایه‌های متعلق به کلاس اقلیت و وزن نسبتاً کمی را به همسایه‌های متعلق به کلاس اکثریت انتساب می‌کند. دقت این روش در تشخیص بیماری کبد برابر ۸۷/۷۱ درصد بوده است.

در پژوهشی دیگر، Kumar و Thakur روشی ترکیبی مبتنی بر الگوریتم طبقه‌بندی Adaboost که به‌صورت مجزا روی هر یک از طبقه‌بندی‌های SVM، LR، NB وRF آموزش داده می‌شود (۱۰)، برای تشخیص بیماری‌های کبدی ارائه کرده‌اند. در این روش با فازی سازی داده‌ها، عدم قطعیت در داده‌ها مدیریت‌شده و استفاده از الگوریتم Adaboost نیز باعث متوازن‌سازی داده‌ها گردیده است. نتایج ارزیابی آن‌ها نشان داده است که الگوریتم Adaboost با استفاده از طبقه‌بندی LR با دقت بهتری توانسته است بیماری کبد را تشخیص دهد (دقت برابر ۹۰/۶۵ درصد). به‌طور مشابه، Li و همکاران (۲۳) نیز مدلی ترکیبی مبتنی بر الگوریتم Adaboost ارائه کرده‌اند. در این مدل از فن درخت تصمیم CART به‌عنوان یک طبقه‌بندی ضعیف برای آموزش و افزایش قدرت الگوریتم Adaboost استفاده شده است. دقت این روش در تشخیص بیماری کبد برابر ۸۳/۰۶ درصد بوده است. Murugesan و همکاران نیز یک روش طبقه‌بندی ترکیبی برای تشخیص بیماری‌های کبدی پیشنهاد کرده‌اند که ابتدا ویژگی‌های مهم داده‌ها را از طریق یک روش wrapper که از سه الگوریتم الهام گرفته از طبیعت به همراه طبقه‌بندی SVM استفاده می‌کند، استخراج کرده و سپس ویژگی‌های استخراج‌شده توسط هر الگوریتم برای آموزش سه شبکه‌ی عصبی به‌صورت مجزا به کار می‌روند (۲۴). کلاس تعیین‌شده توسط هر یک از این سه شبکه‌ی عصبی به یک شبکه‌ی عصبی دیگر واردشده و طبقه‌بندی نهایی توسط آن انجام می‌شود. دقت این روش در تشخیص بیمار کبد برابر ۷۰ درصد بوده است. Sharma و Mehrotra یک روش ترکیبی دومرحله‌ای برای تشخیص بیماری‌های کبدی پیشنهاد کرده‌اند (۸). در مرحله‌ی اول طبقه‌بندی نمونه‌ها به دسته‌های بیمار و سالم توسط یک طبقه‌بندی معمولی انجام شده و درصورتی‌که یک نمونه‌ی بیمار به‌اشتباه توسط طبقه‌بندی سالم تشخیص داده شود، این نمونه وارد مرحله‌ی دوم شده و سالم و یا بیمار بودن آن توسط روش استنتاج مبتنی بر مورد (CBR: Case Based Reasoning) مشخص می‌شود. ارائه‌دهندگان این روش با انتخاب هر یک از طبقه‌بندی‌های ANN، FR، LR، SVM و NB برای مرحله‌ی اول، نشان دادند که دقت ترکیب NB و CBR برابر ۸۳ درصد بوده که در مقایسه با ترکیب سایر طبقه‌بندی‌ها با CBR بیشتر است.

در این مطالعه، مدلی باهدف بهبود دقت در تشخیص افراد مبتلا به بیماری‌های کبدی ارائه شد که مبتنی بر ترکیب نتایج سه طبقه‌بندی FR، MLP و SVM با استفاده از نظریه‌ی ترکیب دمپستر-شافر است. با توجه به نتایج جدول(۳)، بررسی مقادیر حساسیت، ویژگی و دقت هر یک از سه طبقه‌بندی استفاده شده در ترکیب نشان می‌دهد که روش RF نسبت به دو روش دیگر از ویژگی و دقت بالاتری برخوردار است. بااین‌وجود، حساسیت RF نسبت به روش SVM کمتر است. از طرفی مقایسه‌ی مدل پیشنهادی با هر یک از این طبقه‌بندی‌های استفاده شده در ترکیب نشان می‌دهد که مدل پیشنهادی حساسیت، ویژگی و دقت بالاتری دارد و افزایش کارایی آن نسبت به سه طبقه‌بندی FR، MLP و SVM قابل‌توجه است. برتری مدل پیشنهادی و اختلاف چشمگیر مقادیر حساسیت، ویژگی و دقت آن با طبقه‌بندی‌های شرکت‌کننده در ساخت مدل، درستی این ادعا که ترکیب طبقه‌بندی‌ها می‌تواند بر محدودیت‌های هر یک از آن‌ها غلبه کرده و باعث بهبود دقت شود را نشان می‌دهد.

با بررسی مقادیر به‌دست‌آمده در جدول (۴) مشاهده می‌شود که حساسیت و دقت مدل پیشنهادی نسبت به سایر روش‌های مقایسه شده بیشتر است. این مدل با دقت ۹۱/۴۷ درصد و حساسیت ۹۳/۰۳ درصد، دقیق‌تر از سایر روش‌های مورد مقایسه توانسته است بیماران کبدی را شناسایی کند. همچنین مقدار ویژگی در مدل پیشنهادی نیز در بین چهار روشی که مقادیر ویژگی آن‌ها در جدول (۴) مشخص است، از سه روش با اختلاف قابل‌توجهی بیشتر است. مقدار ویژگی در مدل پیشنهادی نسبت به روش Kumar و Thakur (۱۰) حدود ۳/۲۳ درصد کمتر است. بااین‌وجود، مقادیر حساسیت و دقت مدل پیشنهادی از روش Kumar و Thakur (۱۰) بیشتر است. اگرچه این اختلاف کمتر از یک درصد است، ولی حتی یک بهبود جزئی در مقادیر این معیارها در کاربردهای حیاتی از قبیل پزشکی

بسیار قابل‌توجه است (۱۵). همچنین ازآنجایی‌که مقادیر value-p به‌دست‌آمده‌ی حاصل از آزمون t در جدول (۵)، کمتر از ۰/۰۵ هستند، می‌توان نتیجه گرفت که میزان بهبود مدل پیشنهادی در مقایسه با طبقه بندهای استفاده شده در ترکیب و همچنین سایر روش‌های مورد مقایسه، ازنظر آماری معنادار است.

با توجه به نتایج به‌دست‌آمده، مدل پیشنهادی می‌تواند به‌عنوان یک ابزار مفید به پزشکان در تشخیص به‌موقع بیماری‌های کبدی کمک کرده و ضمن کاهش خطاها و آزمایش‌های غیرضروری، موجب استفاده‌ی بهینه از منابع و کاهش هزینه‌های تشخیص و درمان بیماری گردد. اگرچه استفاده از این مدل ترکیبی باعث افزایش زمان اجرایی و پیچیدگی محاسباتی می‌گردد، ولی در کاربردهایی مانند پزشکی که دقت و اعتماد از اهمیت و اولویت بیشتری برخوردار است (۲۸)، به‌کارگیری مدل پیشنهادی مفید خواهد بود.

برای پژوهش‌های آینده، به‌کارگیری مدل پیشنهادی بر روی داده‌های مربوط به بیماری‌های کبدی در مراکز درمانی داخل کشور و بررسی نتایج موردتوجه است. علاوه بر این، توسعه و تعمیم مدل پیشنهادی برای تشخیص و پیش‌بینی سایر بیماری‌ها از قبیل آلزایمر، دیابت، تیروئید و سرطان نیز پیشنهاد می‌شود. همچنین استفاده از سایر فن‌های طبقه‌بندی و یادگیری ماشینی در مدل ترکیبی پیشنهادی و بررسی کارایی آن مفید خواهد بود.

References:

1. Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Inform Med Unlocked 2019; 17:100255.
2. Decharatanachart P, Chaiteerakij R, Tiyarattanachai T, Treeprasertsuk S. Application of artificial intelligence in chronic liver diseases: a systematic review and meta-analysis. BMC Gastroenterol 2021;21(1):1-16.
3. Devikanniga D, Ramu A, Haldorai A. Efficient diagnosis of liver disease using support vector machine optimized with crows search algorithm. EAI Endorsed Trans Energy Web 2020;7(29):1–10.
4. Fathi M, Nemati M, Mohammadi SM, Abbasi-Kesbi R. A machine learning approach based on SVM for classification of liver diseases. Biomed Eng - Appl Basis Commun 2020;32(03):1-9.
5. Harafani H, Suryani I, Ispandi, Lutfiyana N. Neural network parameters optimization with genetic algorithm to improve liver disease estimation. J Phys Conf Ser 2020;1641(1).
6. Kumar P, Thakur RS. Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach. Multimed. Tools Appl 2021;80(11):16515–35.
7. Kuzhippallil MA, Joseph C, Kannan A. Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients. 2020 6th Int Conf Adv Comput Commun Syst (ICACCS) 2020;778–82.
8. Sharma S, Mehrotra D. Two-Stage CBR Based Healthcare Model to Diagnose Liver Disease. Int J Comput Digit Syst 2021;10:1–8.
9. Spann A, Yasodhara A, Kang J, Watt K, Wang B, Goldenberg A, et al. Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review. Hepatology 2020;71(3):1093–105.
10. Kumar P, Thakur RS. An approach using fuzzy sets and boosting techniques to predict liver disease. Comput Mater Contin 2021;68(3):3513–29.
11. Tanwar N, Rahman KF. Machine learning in liver disease diagnosis: Current progress and future opportunities. IOP Conf Ser Mater Sci Eng 2021;1022(1):1–18.
12. Mabrouk AG, Hamdy A, Abdelaal HM, Elkattan AG, Elshourbagy MM, Alansary HAY. Automatic Classification Algorithm for Diffused Liver Diseases Based on Ultrasound Images. IEEE Access 2021; 9:5760–8.
13. Tahmasbi H, Jalali M, Shakeri H. An Expert System for Heart Disease Diagnosis Based on Evidence Combination in Data Mining. J Health Biomed Inf 2017;3(4):251–8. (Persian)
14. Khan RA, Luo Y, Wu FX. Machine learning based

liver disease diagnosis: A systematic review. Neurocomputing 2022;468:492-509.

15. Tang C, Ji J, Tang Y, Gao S, Tang Z, Todo Y. A novel machine learning technique for computer-aided diagnosis. Eng Appl Artif Intell 2020;92.
16. Kaur A, Kumar A. Prediction of Liver Disorders Using Simple Logistic Technique of Machine Learning. In Applications of Machine Intelligence in Engineering 2022 ;81-92.
17. Tahmasbi H, Amoozgar M, Adine H. Replacement of missing values and its effect on the classification accuracy in medical data mining. J Health Biomed Inf 2015;2(1):24–32. (Persian)
18. Wang YC, Cheng CH. A multiple combined method for rebalancing medical data with class imbalances. Comput Biol Med 2021;134:104527.
19. Zhao K, Li L, Chen Z, Sun R, Yuan G, Li J. A survey: Optimization and applications of evidence fusion algorithm based on Dempster-Shafer theory. Appl Soft Comput 2022:109075.
20. UCI Machine Learning Repository: Data Sets. (cited 2022 May 25). Available from: https://archive.ics.uci.edu/ml/datasets.php
21. Ghosh M, Mohsin Sarker Raihan M, Raihan M, Akter L, Kumar Bairagi A, S. Alshamrani S, et al. A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease. Intell Autom Soft Comput 2021;30(3):917–28.
22. Sreejith S, Khanna Nehemiah H, Kannan A. Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. Comput Biol Med 2020;126:103991.
23. Li X, Chen X, Yuan Z. Applicable model of liver disease detection based on the improved CART-AdaBoost algorithm. In: IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE 2021; 1177–81.
24. Murugesan S, Bhuvaneswaran RS, Khanna Nehemiah H, Keerthana Sankari S, Nancy Jane Y. Feature Selection and Classification of Clinical Datasets Using Bioinspired Algorithms and Super Learner. Comput Math Methods Med 2021:6662420.
25. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17.
26. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (cited 2022 May 29). Available from: https://www.cs.waikato.ac.nz/ml/weka/
27. Wu C-C, Yeh W-C, Hsu W-D, Islam MM, Nguyen PAA, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed 2019;170:23–9.
28. Abdar M, Yen NY, Hung JCS. Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees. J Med Biol Eng 2018;38(6):953–65.

Original Article

A MODEL FOR DIAGNOSING LIVER DISEASES USING "MACHINE LEARNING" TECHNIQUES

Hamidreza Tahmasbi[*1], Reza Besharati[2], Mohammad Alishahi[3]*

Received: 31 July, 2022; Accepted: 19 April, 2023

Abstract

Background & Aims: Early diagnosis of liver diseases has a significant effect on the prevention of its complications as well as control and treatment of the disease. "Machine learning" is one of the branches of artificial intelligence that has many applications in the field of medical diagnosis. This study aimed to provide a model with high accuracy and reliability for diagnosing liver diseases using machine learning methods that can help physicians in the early diagnosis and control of liver diseases.

Materials & Methods: This applied-developmental study used the dataset of 583 liver patients. In order to more accurately diagnose of the people with liver diseases, the results of the three classifiers including: Random Forest, Support Vector Machine, and Artificial Neural Network were combined using Dempster-Shafer theory. Weka data mining tool and Python programming language were used to implement the model. The k-fold cross-validation method was applied to evaluate efficiency of the model.

Results: The results showed that accuracy, sensitivity, and specificity in the proposed model were 91.47%, 89.52%, and 93.03%, respectively, which had a better performance than similar studies.

***Conclusion*:** The proposed model in the studied statistical population has a better performance in diagnosing liver diseases, and can help physicians in early diagnosis of the disease and appropriate treatment of it in the early stages of it and thus prevent development of the disease.

Keywords: Classification, Diagnosis, Liver Diseases, Machine Learning

Address: Islamic Azad University, Kashmar, Iran
Tel: +989151046117
***Email*:** htahma@gmail.com

SOURCE: STUD MED SCI 2023: 33(11): 822 ISSN: 2717-008X

[1] Assistant professor, Department of Computer Engineering, Kashmar Branch, Islamic Azad University, Kashmar, Iran (Corresopnding Author)
[2] Department of Nursing, Kashmar Branch, Islamic Azad University, Kashmar, Iran
[3] Assistant professor, Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran