Diagnosing Liver Disease using Firefly Algorithm based on Adaboost

Sheyda Ardam[1], Farhad Soleimanian Gharehchopogh[*2]

Abstract

Introduction: Liver disease is one of the most common and dangerous diseases the early detection of which can be very effective in preventing complications as well as controlling and treating the disease. The purpose of this study was to improve Adaboost algorithm using Firefly Algorithm for diagnosing liver disease.

Method: This is a descriptive-analytic study. The dataset consists of 583 independent records including 10 features of machine learning dataset in the University of California, Irvine. In this study, Adaboost and Firefly Algorithm were combined to increase the effectiveness of liver disease diagnosis. 80% of the data were used for training and 20% for testing.

Results: The results highlighted the superiority of the hybrid model of feature selection over the models without feature selection. Of course, the selection of important features affect the performance of the model. The accuracy of the hybrid model considering 5 and all features was 98.61% and 94.15%, respectively. Overall, the hybrid model proved more accurate compared with most of the other data mining models.

Conclusion: Hybrid model can be used to help physicians identify and classify healthy and unhealthy individuals; it can also be used in medical centers to enhance accuracy and speed, and reduce costs. It cannot be claimed that the hybrid model is the best model; however, it proved more accurarate.

Keywords: Liver Disease, Adaboost Algorithm, Firefly Algorithm, Classification

● **Received:** 05/Jan/2019 ● **Modified:** 10/March/2019 ● **Accepted:** 17/March/2019

DOI:

1 Msc Student, Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran (sheyda.bokani2913@gmail.com)

2 Asistant Professor, Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia,Iran (bonab.farhad@gmail.com)

تشخیص بیماری کبد با الگوریتم کرم شب‌تاب مبتنی بر الگوریتم آدابوست

شیدا آردم[۱]، فرهاد سلیمانیان قره چپق*[۲]

چکیده

مقدمه: بیماری کبدی یکی از بیماری‌های شایع و خطرناک می‌باشد و تشخیص بهموقع این بیماری می‌تواند در پیشگیری از عوارض، کنترل و درمان بیماری بسیار موثر باشد. هدف پژوهش حاضر بهبود الگوریتم آدابوست با الگوریتم کرم شب‌تاب برای تشخیص بیماری کبد می‌باشد.

روش ها: مطالعه حاضر، از نوع توصیفی-تحلیلی می‌باشد. مجموعه داده آن شامل ۵۸۳ رکورد مستقل شامل ۱۰ ویژگی موجود در مجموعه داده یادگیری ماشین دانشگاه کالیفرنیا، ایروین (University of California, Irvine (UCI)) می‌باشد. در این مقاله از ترکیب الگوریتم آدابوست و کرم شب‌تاب در راستای افزایش کارایی تشخیص بیماری کبد استفاده شده است. از ۸۰ درصد داده‌ها جهت آموزش و از ۲۰ درصد باقی‌مانده جهت آزمون استفاده شده است که این مبنا توسط ارزیابی‌های مختلف انتخاب شده است.

یافته ها: نتایج نشان داد که عملکرد مدل ترکیبی با انتخاب ویژگی در مقایسه با حالت بدون انتخاب ویژگی بهتر است. البته انتخاب ویژگی‌های مهم در عملکرد مدل ترکیبی موثر هستند. درصد صحت (accuracy) مدل ترکیبی با پنج ویژگی در بهترین حالت برابر با ۹۸/٦ درصد و درحالت کلی و با تمام ویژگی‌ها برابر با ۹٤/۱ درصد است. در مقایسه کلی، مدل ترکیبی در مقایسه با اغلب مدل‌های داده کاوی از درصد صحت بیشتری برخوردار است.

نتیجه‌گیری: با توجه به نتایج به دست آمده مطالعه حاضر، مدل ترکیبی در تشخیص و طبقه‌بندی افراد سالم و ناسالم می‌تواند نقش مؤثری در کمک به پزشکان داشته باشد و در مراکز پزشکی برای بالا بردن دقت، سرعت و کاهش هزینه‌ها می‌توان از این مدل استفاده نمود. نمی‌توان ادعا کرد که مدل ترکیبی در مقایسه با کل مدل‌ها بهتر است اما در مقایسه با بیشتر مدل‌ها دارای درصد صحت بیشتری است.

واژه‌های کلیدی: بیماری کبد، الگوریتم آدابوست، الگوریتم کرم شب‌تاب، طبقه‌بندی

• وصول مقاله: ۹۷/۱۰/۱۵ اصلاح نهایی: ۹۷/۱۲/۱۹ پذیرش نهایی: ۹۷/۱۲/۲٦

DOI:

۱. دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران (sheyda.bokani2913@gmail.com)

۲. استادیار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران (bonab.farhad@gmail.com)

مقدمه

سلامت هر جامعه مهمترین بحث در پیشرفت و ارتقاء آن جامعه به حساب می‌آید و همه فعالیت‌های انسانی بر این اصل استوار است. با توجه به گسترش تکنولوژی، همه جوامع سعی در بکارگیری فن‌آوری‌های جدید در همه حوزه‌های زندگی انسانی دارند و حوزه پزشکی نیز از این قاعده مستثنی نیست. در سال‌های اخیر پیشرفت‌های قابل توجهی در روش‌ها و ابزارهای هوشمند برای علم پزشکی صورت گرفته است و بکارگیری این ابزارهای هوشمند در تشخیص و کاهش اشتباهات پزشکی، خسارت جانی و مالی کمک شایانی به جامعه پزشکی کرده است. [۲,۱] مطالعه روش‌های تشخیص هوشمند نشان می‌دهد که آنها قادر هستند دقت تشخیص را بهبود داده و موارد مشکوک از دست رفته به دلیل خستگی یا بی‌تجربگی فرد خبره را کاهش دهند و در نتیجه اشتباهات تشخیصی ناشی از تنوع زیاد بیماری‌ها را نیز به حداقل برسانند. [۴,۳] از این‌رو، استفاده از سیستم‌های هوشمند در پزشکی و ارائه روش‌های جدید در این حوزه مورد توجه پژوهشگران قرار گرفته است و مطالعات و تحقیقات زیادی پیرامون آن انجام می‌گیرد.

از طرفی تشخیص درست بیماری‌ها که بر اساس آزمایشات مختلف بر روی بیمار بدست می‌آید در علم پزشکی اهمیت بالایی دارد. در تشخیص بیماری، پزشک با مشکلاتی همچون تنوع بیماری‌ها، رشد چشمگیر بیماری‌ها، اثرات و عوارض آنها مواجه است و این عوامل ممکن است پزشک را در تشخیص زودهنگام بیماری دچار سردرگمی کند. [۵] همچنین، تشخیص نادرست عوارض سنگین و جبران‌ناپذیری برای جامعه پزشکی و بیماران دارد. لذا، همین مشکلات باعث شده جامعه پزشکی در چند دهه اخیر به دنبال ابزارهایی هوشمند جهت تشخیص زود هنگام بیماری‌ها، بررسی بیشتر، پیشگیری و درمان موثر بیماری‌ها باشند و روش‌ها و ابزارهای تشخیص کامپیوتری را با هدف کمک به پزشک ارائه و مورد مطالعه قرار دهند. [۶] بکارگیری ابزارهای هوشمند به نحوی بی‌نظمی را از داده‌ها خارج می‌نمایند و همچنین، خطاهای احتمالی ناشی از خستگی و یا بی‌تجربگی فرد خبره (پزشک) را کاهش می‌دهند و اطلاعات پزشکی موردنیاز و یا جزئیات بیشتر را در اختیار پزشک قرار می‌دهند.

کبد بزرگترین اندام درونی و مهمترین عضو پس از قلب و مغز در بدن انسان است و بدون آن ادامه حیات غیرممکن است. [۷] بیماری‌های کبدی در شمار ده بیماری کشنده در جهان قرار داشته و در ایران حدود ده درصد کل مرگ‌ها و در اروپا پنجمین عامل مرگ و میر بعد از ناراحتی قلبی، سرطان، سکته و بیماری‌های تنفسی به شمار می‌رود. موثرترین راه برای کاهش مرگ و میر ناشی از بیماری‌های کبد، درمان این بیماری‌ها در مراحل اولیه آن است. [۸] درمان زود هنگام، مستلزم تشخیص زودهنگام براساس روش‌های تشخیصی دقیق و قابل اعتماد است. بر این اساس در عصر حاضر، یکی از مسائل مطرح در علم پزشکی که توجه بسیاری از محققان را به خود جلب کرده است، تشخیص بیماری کبد و کمک به درمان آنها است. به دلیل اهمیت بیماری کبد، مطالعه حاضر بر روی این عضو صورت گرفته است و روشی براساس ترکیب الگوریتم آدابوست با الگوریتم کرم شب‌تاب برای تشخیص بیماری کبد ارائه شده است.

در سال‌های اخیر استفاده از روش‌های هوش مصنوعی به منظور تشخیص بیماری‌ها، مورد توجه بسیاری از پژوهشگران قرار گرفته است. حیدری و تیموری، روشی برای پیش‌بینی نارسایی کبدی با استفاده از ترکیب شبکه عصبی مصنوعی و الگوریتم ژنتیک ارائه داده‌اند. دو حالت کلی در ترکیب روش‌ها که عبارتند از ترکیب حمایتی و ترکیب تشریک مساعی در این مقاله مطرح شده است که در روش پیشنهادی از ترکیب تشریک مساعی برای ترکیب دو الگوریتم ژنتیک و شبکه عصبی مصنوعی استفاده شده است. در این روش شبکه عصبی و الگوریتم ژنتیک به صورت همزمان اعمال می‌شوند و یک سیستم یکپارچه را می‌سازند و تعداد نرون‌های لایه پنهان شبکه عصبی و همچنین وزن بین گره‌ها از طریق الگوریتم ژنتیک بهبود یافته است. در مطالعه حاضر، از دو مجموعه داده بوپا (British United Provident (BUPA)) Association) شامل ۳۴۵ نمونه از بیماران کبدی و مجموعه

داده های بیمارستانی کبد ((Indian Liver Patient Dataset (ILPD)شامل ۵۸۳ نمونه است، استفاده شده است. برای پیاده‌سازی روش پیشنهادی از نرم‌افزار وکا (WEKA) استفاده شده است. درصد صحت دسته‌بندی این تکنیک ترکیبی روی مجموعه داده BUPA برابر با ۴۶ درصد و روی مجموعه داده ILPD برابر با ۶۰ درصد بدست آمده است و مشخص گردید این روش دسته‌بندی نسبت به شبکه عصبی سنتی عملکرد بهتری داشته است. [۹]

لین و چوانگ از روش‌های داده کاوی مانند شبکه عصبی مصنوعی و استدلال برمبنای منطق برای تشخیص بیماری کبد استفاده کرده‌اند. مجموعه داده‌های پزشکی شامل ۵۱۰ نمونه به یک مرکز پزشکی در تایوان در طی یک سال از ماه مارس ۲۰۰۵ تا فوریه ۲۰۰۶ می‌باشد. با کمک پزشکان و متخصصین پزشکی کبد، ۲۱۰ مورد سالم و ۳۰۰ مورد با شرایط کبدی (هپاتیت مزمن، هپاتیت الکل، سیروز کبدی، هپاتیت B و دیگر موارد) تشخیص داده شده‌اند. درصد صحت شبکه عصبی مصنوعی در مقایسه با مدل‌های دیگر بیشتر است. درصد صحت شبکه عصبی مصنوعی برابر با ۹۸ درصد است. [۱۰]

پژوهشگران یک سیستم تشخیص خودکار بیماری کبد مبتنی بر برنامه‌نویسی ژنتیک طراحی کرده‌اند. روش یادگیری سیستم مطابق با برنامه‌نویسی ژنتیک می‌باشد که بهترین راه حل را از اطلاعات مشخصه‌های محلی در مسائل پیچیده پیدا می‌کند. در این مقاله از روش‌های هوش مصنوعی مانند رگرسیون خطی چندگانه، ماشین بردار پشتیبان، جنگل تصادفی، و J48 برای تشخیص بیماری کبد استفاده شده است. این سیستم توسط مجموعه داده ILPD ارزیابی شده که درصد صحت آن برای برنامه‌نویسی ژنتیک برابر با ۸۴,۷ درصد است. [۱۱]

از دو الگوریتم درخت تصمیم یعنی الگوریتم‌های C5.0 و تشخیص متقابل اتوماتیک ((Chi-square Automatic Interaction Detector (CHAID) برای پیش‌بینی بیماری کبد استفاده شده است. [۱۲] از طریق درخت C5.0 و CHAID مجموعه داده ILPD مورد تجزیه و تحلیل قرار گرفته است. مجموعه داده بکار برده شده در مقاله حاضر، ۵۸۳ نمونه از مجموعه داده بیماری کبد از سایت UCI در نظر گرفته شده است. الگوریتم CHAID یک درخت تصمیم مبتنی بر شناسایی قوی‌ترین روابط متقابل بین متغیرهای مستقل و وابسته است و به این منظور از مقدار احتمال آماره کای دو مربوط به آزمون استقلال جداول توافقی استفاده می‌کند. درخت ایجاد شده در CHAID وسیع‌تر از درخت تصمیم C5.0 است. ضعف CHAID در عدم توانایی ایجاد بهینه‌ترین تقسیمات بر اساس متغیرهای موجود است. نتایج بدست آمده بیانگر این است که در هر دو الگوریتم، ویژگی‌های TB, SGPT, ALB, DB و A/G تاثیر قابل توجهی در پیش‌بینی بیماری کبد با توجه به قوانین تولید شده دارند. ویژگی‌های ذکر شده در مجموعه داده بیماری کبد از درصد صحت بیشتری بهره‌مند هستند. مقایسه کارایی این دو الگوریتم مشخص کرد که الگوریتم C5.0 تقویت شده عملکرد بهتری نسبت به CHAIDدارد و با بررسی عوامل موثر در فواصل کوچک صحت الگوریتم افزایش می‌یابد و در واقع الگوریتم C5.0 همه پارامترها را در نظر می‌گیرد که این نشان‌دهنده این است که این الگوریتم از پارامترها با احتیاط بیشتری استفاده می‌کند، در حالیکه الگوریتم CHAID تنها از پنج پارامتر فوق استفاده می‌کند. بنابراین الگوریتم درخت C5.0 با درصد صحت ۹۳/۷ نسبت به الگوریتم CHAID با درصد صحت ۶۵، صحت بالاتر و همچنین قوانین بیشتری دارد. [۱۲]

روشی با استفاده از درخت تصمیم برای تشخیص زودهنگام بیماری کبد ارائه شده است. در مقاله حاضر، مدلی برای پیش‌بینی بیماری کبدی بر اساس برخی معیارهای آزمایشگاهی افراد ارائه می‌شود. روش ارائه شده با استفاده از یک مجموعه داده معتبر مربوط به بیماران کبدی مقایسه و ارزیابی می‌شود. نتایج بدست آمده، کارایی مدل پیش‌بینی کننده بر اساس الگوریتم درخت تصمیم را نشان می‌دهد. [۱۳]

سموات و صف آرا، یک سیستم ترکیبی هوشمند با هدف تشخیص بیماری کبد با استفاده از ویژگی‌های استخراج شده از آزمایش خون و الگوریتم داده کاوی، ارائه داده‌اند. برای کاهش ویژگی‌ها از الگوریتم بهینه‌سازی ازدحام ذرات باینری

به عنوان راهبرد جستجو و از شبکه عصبی به عنوان تابع ارزیابی در روش پیشنهادی این مقاله استفاده شده است و در نهایت تکنیک طبقه‌بندی جنگل تصادفی برای افزایش دقت تشخیص سیستم به کار برده شده است. از مجموعه داده ILPD که شامل ۱۶۷بیمار غیر کبد و۴۱۶ بیمار کبد با ۱۰ ویژگی برای تست و ارزیابی روش پیشنهادی استفاده شده است که در نهایت دقت روش پیشنهادی برابر با ۷۶/۲ درصد محاسبه گردیده است. [۱۴] همچنین کریستوفر و همکاران یک رویکرد برمبنای الگوریتم بهینه‌سازی ازدحام باد برای کاوش دانش پزشکی ارائه داده‌اند که در آن از معیار برازندگی مقدار جایز ((JVAL) Jabez Value) برای ارزیابی دسته‌بندی‌های مبتنی بر قوانین استفاده شده است. الگوریتم بهینه‌سازی ازدحام باد برای به‌دست آوردن جایگشت‌ها و ترکیبات مختلف قوانین استفاده شده که به موجب آن، قوانین بهینه مطلوب برای پیش‌بینی داده آزمون استفاده شده است. در این مقاله از شش مجموعه داده پزشکی معین برای تست استفاده شده است. نتایج معیار دقت برای الگوریتم بهینه‌سازی اجتماع ذرات و الگوریتم بهینه‌سازی ازدحام باد به ترتیب ۰/۵۹۲۱ و ۰/۵۸۴۶ بدست آمده است. [۱۵]

هدف اصلی پژوهش حاضر، ارائه یک سیستم تشخیص بیماری کبد برای مجموعه داده ILPD [۱۶] مبتنی بر الگوریتم کرم شب‌تاب [۱۷] و طبقه‌بند آدابوست [۱۸] است. در مدل ترکیبی از الگوریتم کرم شب‌تاب برای انتخاب ویژگی و از الگوریتم آدابوست برای طبقه‌بندی نمونه‌ها استفاده می‌شود. الگوریتم کرم شب‌تاب یکی از الگوریتم‌های فرا ابتکاری است که برمبنای جمعیت اولیه و تکرار به راه حل بهینه دست می‌یابد. همچنین الگوریتم آدابوست یکی از الگوریتم‌های داده کاوی است که برای طبقه‌بندی و تشخیص نمونه‌ها استفاده می‌شود.

روش‌ها

مطالعه حاضر، از نوع توصیفی-تحلیلی است که بر اساس ویژگی‌های ورودی به تشخیص وضعیت بیماران کبد از نظر سالم یا ناسالم بودن می‌پردازد. داده‌های مورد استفاده در این مقاله از مجموعه داده مربوط به بیماران مبتلا به کبد، موجود در مجموعه داده یادگیری ماشین دانشگاه ایروین، کالیفرنیا تأمین شده است. [۱۶] مجموعه داده‌ها شامل ۵۸۳ رکورد کبد با ۱۰ ویژگی (و یک ویژگی متعلق به کلاس) می‌باشند. مجموعه داده کبد را می‌توان یک ماتریس ۱۰×۵۸۳ تلقی کرد. از مجموعه داده کبد، ۴۱۶ نمونه (۷۲ درصد)، پرونده بیماران کبدی (کلاس یک) و ۱۶۷ نمونه (۲۸ درصد) پرونده افراد سالم (کلاس دو) می‌باشد. یکی از ویژگی‌های عمومی داده‌های پزشکی، نامتوازن بودن آنها است. [۳۳] برمبنای تقسیم‌بندی نمونه‌ها (سالم و غیرسالم)، مجموعه داده بکار گرفته شده در این مطالعه از این مساله مستثنی نیست. [۳۴] لذا، در مقاله حاضر، داده‌ها به حالت استاندارد تبدیل می‌شوند.

در الگوریتم‌های طبقه‌بندی استاندارد، تقسیم کلاس‌ها باید در حالت متوازن انجام شود و این نوع از الگوریتم‌ها در مقابله با مجموعه داده‌های نامتوازن کارایی مناسبی را از خود ارایه نمی‌دهند؛ چرا که الگوریتم‌های معمول طبقه‌بندی به سمت نمونه‌های آموزشی کلاس بزرگ‌تر متمایل می‌شوند که این موضوع باعث افزایش خطا در تشخیص نمونه‌های اقلیت می‌شود.[۳۵] این مسئله یکی از چالش‌های مهم برای طبقه‌بندی داده‌های نامتوازن محسوب می‌شود. از روش‌های متنوعی برای حل مساله عدم توازن در علم یادگیری ماشین استفاده می‌شود. [۳۶] یکی از این روش‌ها، روش‌های بازبینی در سطح الگوریتم است که با تغییر در الگوریتم طبقه‌بندی به نوعی مساله عدم توازن مرتفع می‌شود. [۳۷] از دیگر روش‌های حل نامتوازن بودن داده‌ها روش‌های مبتنی بر ترکیب طبقه‌بندها است. هدف اصلی روش ترکیب طبقه‌بندها تلاش برای بهبود عملکرد طبقه‌بندی داده‌ها از طریق ترکیب چندین طبقه‌بند است. به طوری که ترکیب چند طبقه‌بند عملکرد بهتری نسبت به یکی از همان طبقه‌بندها خواهد داشت. روش سوم؛ روش‌های سطح داده است. در این نوع از روش‌ها توزیع کلاس نامتوازن با نمونه‌گیری مجدد در فضای داده‌ها متوازن می‌شود. [۳۸] و نهایتاً روش‌های حساس به هزینه، دسته دیگری از روش‌های ارایه شده برای حل عدم توازن در داده‌ها محسوب می‌شود. این دسته از روش‌ها به نوعی از ترکیب روش‌های تغییر در الگوریتم طبقه‌بند و روش‌های سطح داده حاصل می‌شوند. [۳۹] در این

بین، روش‌های سطح داده با دو رویکرد زیرنمونه‌برداری و بیش‌نمونه‌برداری از روش‌های موثر در متوازن نمودن داده‌ها محسوب می‌شوند. [۴۰]

در مطالعه حاضر به دلیل نامتوازن بودن نمونه‌های بیماری کبد از روش‌های سطح داده شامل زیرنمونه‌برداری تصادفی کلاس اکثریت، بیش‌نمونه‌برداری تصادفی کلاس اقلیت و بیش‌نمونه‌برداری مصنوعی کلاس اقلیت برای افزایش دقت طبقه‌بندی آدابوست در تشخیص کلاس‌های سالم از غیرسالم استفاده می‌شود. در مجموعه داده اولیه، داده‌های وابسته به کلاس سالم یعنی کلاس اقلیت برابر با ۱۶۷ رکورد و داده‌های متعلق به کلاس غیرسالم یعنی کلاس اکثریت برابر با ۴۱۶ رکورد می‌باشد که بر این اساس مجموعه داده این مطالعه دارای نرخ عدم توازن ۲/۵ درصد است. در ابتدا از کلاس اکثریت یک نمونه تصادفی به اندازه داده‌های کلاس اقلیت انتخاب شد تا هر دو کلاس به یک اندازه مشاهده شوند. [۴۱] در مرحله دوم از کلاس اقلیت به صورت تصادفی نمونه‌هایی انتخاب و به همین مجموعه اضافه شد تا اندازه داده‌های کلاس اقلیت با اندازه داده‌های کلاس اکثریت برابر شود. در مرحله سوم با روش بیش‌نمونه‌برداری مصنوعی، کلاس اقلیت بر روی مجموعه داده اصلی ایجاد شد.

در شکل(۱)، فلوچارت مدل ترکیبی نشان داده شده است. در این مقاله درصد نمونه‌های آموزشی برابر با ۸۰ درصد می‌باشد. به‌منظور جلوگیری از تأثیر نامطلوب بازه تغییرات ورودی‌های مختلف بر روی الگوریتم طبقه‌بندی، رکوردها بین ۰ و ۱ نرمالیزه می‌شوند. محدوده ویژگی‌های مجموعه داده با هم متفاوت است. در بسیاری از الگوریتم‌های طبقه‌بندی، این تفاوت در محدوده ویژگی‌ها، باعث می‌شود که ویژگی با محدوده‌ی بزرگ‌تر تأثیر منفی روی درصد صحت داشته باشد. لذا، برای انجام فرآیند مدل سازی و یادگیری طبقه‌بندی باید نمونه‌ها نرمالیزه شوند. بنابراین، ضروری است که داده‌ها نرمالسازی شوند تا تأثیری که هر ویژگی روی نتایج دارد، به حالت استاندارد تبدیل شود. در مرحله دوم با استفاده از روش انتخاب ویژگی مبتنی بر الگوریتم کرم شب‌تاب، داده‌های زائد(ویژگی‌های کم اولویت) حذف می‌شود و در مرحله سوم با استفاده از الگوریتم آدابوست عملیات طبقه‌بندی انجام می‌گیرد.

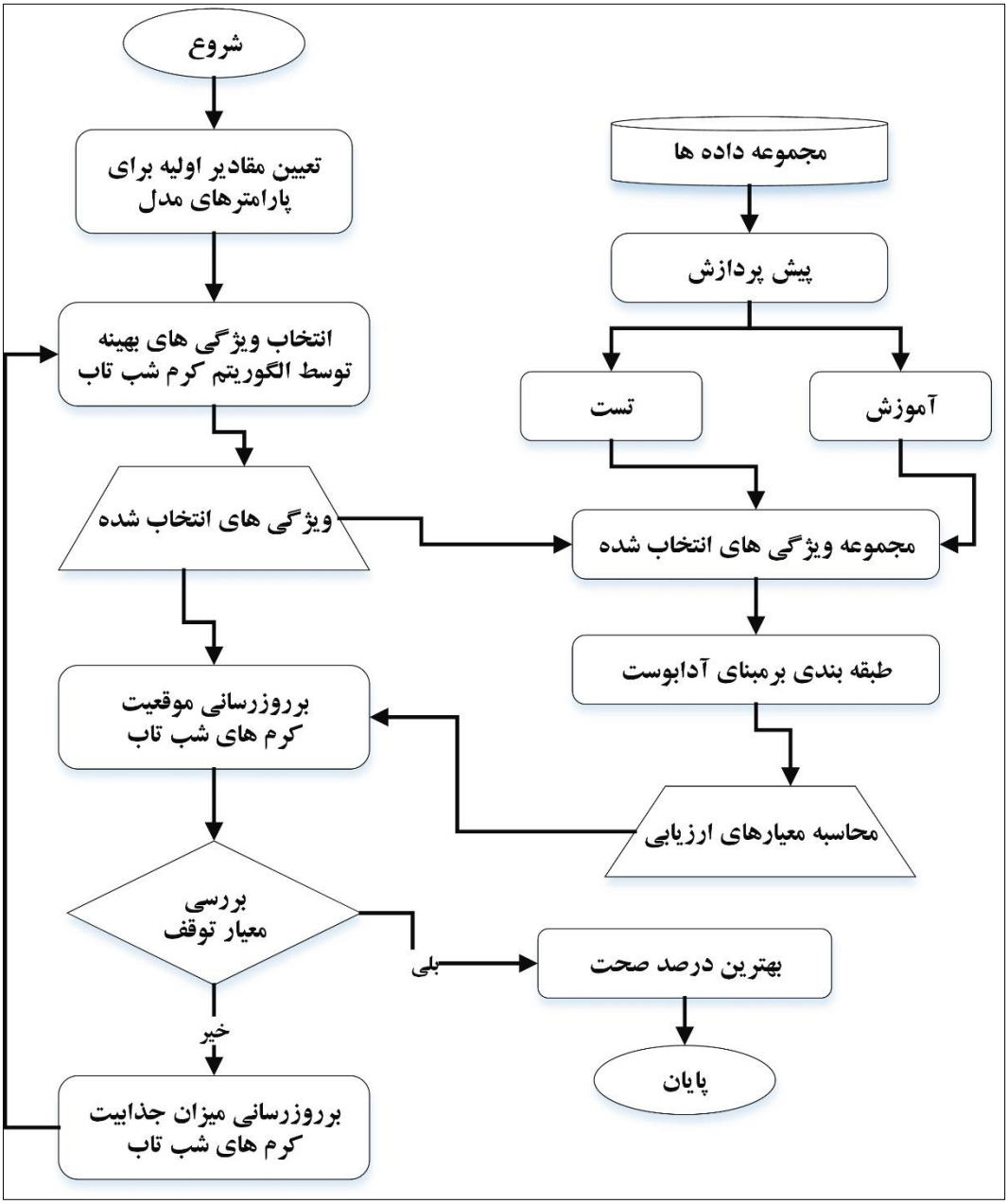
مرحله اول پیش‌پردازش می‌باشد که مهمترین بخش آن نرمال‌سازی داده‌ها است. نرمال‌سازی داده‌ها با یک تبدیل خطی یا غیرخطی، داده‌ها را به بازه‌ای که معمولاً [۱,۱-] و [۰,۱] است، نگاشت می‌کند. اصولاً وارد کردن داده‌ها به‌صورت خام باعث کاهش سرعت الگوریتم طبقه‌بندی می‌شود. [۴۲] برای اجتناب از چنین شرایطی باید داده‌ها استاندارد شوند. استانداردسازی داده‌های اولیه در مدل ترکیبی طبق معادله(۱) انجام گرفته است. [۲۰,۱۹]

$$x_n = \frac{(x_r - x_{\min})}{(x_{\max} - x_{\min})} \quad (۱)$$

در معادله(۱)، X_r، X_n، X_{max} و X_{min} به ترتیب نشان‌دهنده مقادیر واقعی، استاندارد شده، حداکثر و حداقل داده‌های تحت بررسی هستند. هدف از استانداردسازی مقدار ویژگی‌ها این است که همه ویژگی‌ها به منظور افزایش درصد صحت در یک محدوده مشخص باشند، که این محدوده در بازه صفر و ۱ است. پس از استانداردسازی داده‌ها، کل داده‌ها به دو بخش آموزش و آزمون تفکیک شدند. البته داده‌ها از حالت غیرمتوازن به حالت متوازن تبدیل شدند و سپس به دو بخش آموزش و آزمون تبدیل شدند.

مرحله دوم انتخاب ویژگی مبتنی بر کرم شب‌تاب می‌باشد. در این مرحله ویژگی‌هایی از مجموعه خصیصه‌ها که برای پیشگویی خروجی موثرتر هستند، انتخاب می‌شود. علاوه بر آن، مفهوم موجود در ویژگی‌ها بعد از انتخاب ویژگی حفظ می‌شود. در مقاله حاضر از الگوریتم کرم شب‌تاب[۴۳] برای انتخاب ویژگی استفاده شده است. کرم‌های شب‌تاب به منظور جذب جفت و شکار، نورهایی تولید می‌کنند که الگوی نوری هر کدام با دیگری متفاوت است. میزان این نور رابطه مستقیم با جذابیت کرم شب‌تاب دارد. با در نظر گرفتن میزان نور کرم به

عنوان تابع هدف، می‌توان رفتار کرم‌های شب‌تاب را به صورت یک الگوریتم بهینه‌ساز مدل نمود.



شکل ۱- فلوچارت مدل ترکیبی

الگوریتم کرم شب‌تاب به‌صورت تصادفی یک جمعیت اولیه از کرم‌های شب‌تاب به تعداد F عدد $p=\{x_1,x_2,\ldots x_F\}$ تولید می‌کند. برای حل مسئله به‌صورت t بعدی، کرم iام $x_i=\{x_{i1},x_{i2},\ldots,x_{it}\}$ و $(1<i<F)$ قرار می‌گیرد. بعد از تولید جمعیت کرم‌های شب‌تاب، مقدار شایستگی هر کرم شب‌تاب(x_i) برمبنای فاصله نزدیکی (فاصله کمتر) به کرم‌های همسایه محاسبه می‌شود، سپس کرم‌ها براساس شایستگی (کرم‌هایی که فاصله کمتری دارند) که در این مسئله به‌صورت یک روند نزولی است مرتب می‌شوند. در الگوریتم کرم شب‌تاب، هر کرم شب‌تاب یک راه حل ممکن مسئله در یک

فرم برداری است که طول هر بردار برابر با ۱۰ می‌باشد. به دلیل اینکه تعداد ویژگی‌ها در مجموعه داده‌های بیماری کبد برابر با ۱۰ ویژگی است. پس از ایجاد جمعیت اولیه به صورت تصادفی و تعیین ارزیابی شایستگی متناظر با هر کرم شب‌تاب با استفاده از تابع ارزیابی، میزان درخشندگی کرم شب‌تاب i طبق معادله(۲) محاسبه می‌شود.

$$B_i = \frac{1}{Sum_{x(i)} / VL_{(i)}}$$

(۲)

که در معادله(۲)، $sum_{x(i)}$ مجموع بردار iام و $VL_{(i)}$ طول بردار iام است و B_i میزان درخشندگی کرم شب‌تاب i را بیان می‌کند. منظور از میزان درخشندگی این است که اگر مقدار Bi بیشتر باشد آن کرم می‌تواند کرم‌های بیشتری را به اطراف خود جذب کند و لذا نقاط بهینه را در فضای مسئله کشف کرده است. مقدار Bi زمانی بیشتر خواهد بود که مجموع بردار iام کمتر باشد، به عبارتی فاصله بین کرم‌ها باید کمتر باشد. اگر مجموع ویژگی‌های یک بردار در مقایسه با بردارهای دیگر بیشتر باشد آن بردار به عنوان یک بردار درخشنده (برازنده) انتخاب می‌شود. در این الگوریتم برای بهبود سرعت همگرایی الگوریتم، در هر تکرار، تعدادی از کرم‌های شب‌تاب که بیشترین میزان درخشندگی را بدست آورده‌اند، انتخاب شده و از بین کرم‌های باقی‌مانده، هر کدام به سمت نزدیک‌ترین کرم شب‌تاب درخشنده حرکت می‌کنند. فاصله بین دو کرم طبق معادله(۳) محاسبه می‌شود. [۱۷]

$$r_{i,j} = \left\| x_i - x_j \right\|$$

(۳)

میزان جذابیت کرم‌های شب‌تاب (β) نسبی بوده و به فاصله بین دو کرم ($r_{i,j}$) و ضریب جذب نور (γ) بستگی دارد که از رابطه(۴) قابل محاسبه است. [۱۷] جذابیت هر کرم شب‌تاب برای کرم شب‌تاب دیگر رابطه مستقیم با درخشندگی دارد. این بدین معنی است که کرم شب‌تابی که جذابیت کمتری دارد به سمت کرم شب‌تاب جذاب‌تر حرکت خواهد کرد. اگر حول یک کرم شب‌تاب، کرم شب‌تابی با جذابیت بیشتر حاضر نباشد، کرم شب‌تاب به طور تصادفی حرکت خواهد نمود.

$$\beta_{(r)} = \beta_0 e^{-\gamma r_{i,j}}$$

(۴)

در معادله(۴) β_0 بیانگر ماکزیمم جذابیت کرم درخشان‌تر در $r=0$ بوده و مقداری در بازه $[0,1]$ دارد. پارامتر γ بیانگر ضریب جذب می‌باشد و مقداری در بازه $[0,\infty)$ دارد که این پارامتر مربوط به الگوریتم کرم شب‌تاب است و برای جذب کرم‌ها و کمینه کردن فاصله بین کرم‌ها استفاده می‌شود. اگر $\beta_0=0$ باشد هر یک از کرم‌های شب‌تاب بدون همکاری با سایر کرم‌های شب تاب به تنهایی فضای مساله را جستجو می‌کنند و جستجو به صورت تصادفی انجام می‌گیرد. همچنین اگر مقدار پارامتر $\gamma=\infty$ باشد منجر به انجام یک جستجوی تصادفی در فضای مساله می‌گردد. موقعیت کرم iام پس از حرکت به سمت کرم jام که درخشان‌تر است طبق معادله(۵) محاسبه می‌شود. [۱۷]

$$x_i = x_i + \beta(x_j - x_i) + \alpha(rand - 0.5)$$

(۵)

که در معادله(۵)، x_i موقعیت کرم شبتاب کم نورتر، x_j موقعیت کرم شب‌تاب درخشان‌تر و α عددی تصادفی است که نشان‌دهنده میزان جهش هر کرم شب‌تاب می‌باشد. همچنین در بخش سوم در معادله(۵)، حرکت تصادفی در فرآیند جذب بیان شده است که باعث جستجوی جامع‌تر فضای تصمیم مسأله توسط الگوریتم می‌شود. در مدل ترکیبی باید الگوریتم کرم شب‌تاب از حالت پیوسته به گسسته تبدیل شود. به دلیل اینکه مقدار مجموعه داده‌ها گسسته هستند و در بازه ۰ و ۱ قرار دارند.

$$g(x) = \frac{1}{1 + e^{-x}}$$

(۶)

$$x_i = \begin{cases} 0, & if\ g(x) < 0.5 \\ 1, & otherwise \end{cases}$$

(۷)

برای تبدیل اعداد پیوسته به باینری از تابع سیگموئید طبق معادله(۶)، استفاده شد. خروجی تابع سیگموئید در یک

محدوده عددی خاصی (عموماً بین صفر و یک) قرار داده شد. در این تابع جواب ۰ یا ۱ نخواهد بود بلکه مجموعه اعدادی بین صفر و یک است. لذا در الگوریتم کرم شب‌تاب برای تبدیل حالت پیوسته به گسسته موقعیت هر کرم شب‌تاب x_i طبقه معادله(۷)، تعریف می‌شود.

در مدل ترکیبی با استفاده از الگوریتم کرم شب‌تاب زیرمجموعه‌ای از ویژگی‌ها که به بیشترین درصد صحت منجر می‌شوند، انتخاب می‌گردد. تابع برازندگی برای انتخاب ویژگی از هر بردار طبق معادله (۸) تعریف می‌شود. در معادله (۸)، $|n|$ تعداد کل ویژگی‌ها و $|S|$ تعداد ویژگی‌های انتخاب شده است. معیار صحت و پارامترهای α و β دو پارامتر مربوط به اهمیت کیفیت طبقه‌بندی و طول زیر مجموعه انتخاب شده می‌باشند، مقدار α در بازه صفر و یک و $\beta=(1-\alpha)$ می‌باشد. [۴۴]

$$Fitness = \alpha.Accuracy + \beta.\frac{|n| - |S|}{|n|} \quad (۸)$$

مرحله سوم، طبقه‌بندی می‌باشد. در الگوریتم آدابوست [۱۸] در مرحله آموزش، برمبنای نرخ صحت به هر دسته ضعیف (نمونه‌های تست)، وزنی اختصاص داده می‌شود و به هر نمونه آموزشی نیز وزنی اختصاص داده می‌شود که نشان‌دهنده درستی فرآیند طبقه‌بندی است. در طول توالی اضافه شدن یک دسته‌بند ضعیف، اگر هر یک از نمونه‌های آموزش به‌درستی طبقه‌بندی شوند، وزنش کاهش می‌یابد. در غیر این صورت، وزن آن افزایش خواهد یافت؛ بنابراین دسته‌بند در تکرار بعد می‌تواند بر روی نمونه‌های اشتباه طبقه‌بندی شده تمرکز کند. مراحل الگوریتم آدابوست برای طبقه‌بندی نمونه‌ها بشرح زیر است:

۱) وارد کردن داده‌های آموزشی $D = \{(x_1, y_1), \dots, (x_m, y_m)\}$ که x_m بردار ویژگی‌ها، y_i کلاس داده‌ها و m تعداد داده‌های آموزشی است.

۲) مقداردهی اولیه بردار وزن $w_i(x_i) = 1/m \leftarrow$

۳) پیدا کردن T عدد دسته‌بند ضعیف $h_t(t = 1,2, \dots, T)$ و مقداردهی اولیه (t=0)

۳-۱) تعیین $P_t(x_i) = \frac{w_t(x_i)}{\sum_{i=1}^{m} w_t(x_i)} \leftarrow :P_t(x_i)$

۳-۲) محاسبه نرخ خطای دسته‌بند $\varepsilon_t = \sum P_t(x_i)[h_t(x_i) \neq y_i] \leftarrow :h_t$

اگر $\varepsilon_t > 0.5$ بود T=t-1 و از چرخه خارج می‌شود.

۳-۳) محاسبه وزن $\alpha_t = log\left(\frac{\varepsilon_t}{1-\varepsilon_t}\right) \leftarrow :h_t$

۳-۴) به‌روزرسانی وزن تمام داده‌های آموزش از i تا m: $w_{t+1}(x_i) = w_t(x_i) * \begin{Bmatrix} e^{-\alpha_t} \\ e^{\alpha_t} \end{Bmatrix} \leftarrow$

که $e^{-\alpha_t}$ برای داده‌هایی که صحیح طبقه‌بندی شده‌اند. و e^{α_t} برای داده‌هایی که اشتباه طبقه‌بندی شده‌اند.

۳-۵) اگر $\varepsilon_t > 0.001$ و t>T باشد آنگاه برو به مرحله ۳-۱

٤) خروجی یک دسته‌بند قوی است اگر از ترکیب دسته‌بند کنندگان ضعیف وزن‌دار تشکیل شود. $H(x) = \arg max_{1 \le j \le J} \sum_{t=1}^{T} \alpha_t [h_t(x) = y]$

در مطالعه حاضر به منظور طبقه‌بندی نمونه‌های سالم و ناسالم از روش AdaBoost.M1 استفاده شد. آدابوست تعدادی طبقه‌بند می‌سازد وخروجی پیش‌بینی شده توسط طبقه‌بندها را از طریق نمونه‌هایی با حداکثر وزن با یکدیگر ترکیب می‌نماید. طبقه‌بندها از طریق آموزش یک طبقه‌بند ضعیف، بر روی توزیعی از داده‌های آموزش که در هر مرحله بروز می‌شود بدست می‌آیند. عمل توزیع بروز شده بدین صورت است که نمونه‌هایی که در مراحل قبل به اشتباه طبقه‌بندی شده‌اند با

احتمال بیشتری برای آموزش طبقه‌بند بعدی مورد استفاده قرار می‌گیرند. از این رو، داده‌های آموزشی طبقه‌بندی شده، به سمت نمونه‌هایی که به سختی طبقه‌بندی می‌شوند حرکت داده می‌شوند.

الگوریتم آدابوست یکی از روش‌های طبقه‌بندی است که برمبنای تکرار و خطا سعی می‌کند که نمونه‌های مشابه را دریک دسته قرار دهد. آموزش برروی هر ویژگی انجام می‌گیرد و لذا برای حداقل کردن خطا از معادله(۹) استفاده می‌شود. در معادله(۹)، w وزن ویژگی‌ها و x و y مقدار ویژگی‌ها است.

$$\varepsilon_j = \sum_{i=1}^{n} w_i \left| h_j(x_i) - y_i \right| \quad (۹)$$

در الگوریتم آدابوست برای کاهش دادن خطا، مراحل آموزش در چندین مرحله تکرار می‌شود و در هر مرحله نمونه‌های مشابه به هم در یک دسته قرار می‌گیرند. مدل آدابوست برمبنای نمونه‌های آموزشی، مدل طبقه‌بندی را ایجاد می‌کند و سپس نمونه‌های تست برمبنای مدل سازی می‌توانند طبقه‌بندی شوند و در نهایت دقت تشخیص مدل ترکیبی ارزیابی می‌گردد.

مرحله چهارم، معیارهای ارزیابی می‌باشد. در مطالعه حاضر، جهت ارزیابی دقت طبقه‌بندی از معیارهای درصد بازخوانی، دقت، F-Measure و صحت استفاده شده است [۲۲,۲۱] که صحت معیار اصلی است. حال آنکه معیار بازخوانی (recall) و معیار دقت (precision) و F-Measure به طور مجزا عملکرد طبقه‌بندی کننده را برای کلاس‌های مختلف نشان می‌دهند. هرچه میزان بازخوانی بالاتر باشد بیانگر این است که قابلیت شناسایی درست کلاس‌ها بیشتر است.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۱۰)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (۱۱)$$

$$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} * \text{Recall}} \quad (۱۲)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱۳)$$

یافته ها

مجموعه داده مورد استفاده در این مطالعه، مجموعه داده استانداردی به نام ILPD [۱۶] است که از شمال شرق آندراپرادش هندوستان جمع‌آوری شده است و در قالب یک فایل اکسل در مخزن داده دانشگاه کالیفرنیای ایروین ثبت گردیده است. مجموعه داده ILPD یک مجموعه داده نامتوازن شامل ۵۸۳ نمونه می‌باشد که از این میان ۴۱۶ نمونه (۷۲ درصد)، پرونده بیماران کبدی (کلاس یک) و ۱۶۷ نمونه (۲۸ درصد) پرونده افراد سالم (کلاس دو) می‌باشد. از این ۵۸۳ نمونه تحت بررسی، ۴۴۱ نفر مرد و ۱۴۲ نفر زن هستند. این مجموعه داده شامل ده ویژگی و یک فیلد هدف می‌باشد. فیلد هدف یک برچسب کلاس است که مجموعه داده را به دو گروه (بیمار و سالم) تقسیم کرده است. ویژگی‌های این مجموعه داده در جدول (۱) نشان داده شده است که در ادامه هر ویژگی نیز به صورت مختصر معرفی گردیده است.

جدول ۱: ویژگی‌های مجموعه داده ILPD

شماره ویژگی	نام ویژگی	نوع ویژگی	توضیحات	محدوده
۱	Age	عددی	سن	۴ تا ۹۰
۲	Gender	اسمی	جنسیت	مرد-زن
۳	TB (Total Bilirubin)	عددی	بیلی‌روبین کلی	۰/۴ تا ۷۵
۴	DB (Direct Bilirubin)	عددی	بیلی‌روبین مستقیم	۰/۱ تا ۱۹/۷
۵	Alkphos Alkaline Phosphotase	عددی	آلکالین فسفاتاز	۶۳ تا ۲۱۱۰

جدول ۱: (ادامه)

شماره ویژگی	نام ویژگی	نوع ویژگی	توضیحات	محدوده
۶	SGPT Alamine Aminotransferase	عددی	آلانین آمینوترانسفراز	۱۰ تا ۲۰۰۰
۷	SGOT Aspartate Aminotransferase	عددی	آسپارتات آمینوترانسفرا	۱۰ تا ۴۹۲۹
۸	TP (Total Protein)	عددی	پروتئین کلی	۲/۷ تا ۹/۶
۹	ALB (Albumin)	عددی	آلبومین	۰/۹ تا ۵/۵
۱۰	A/G Ratio (Albumin and Globulin Ratio)	عددی	نسبت آلبومین به گلوبولین	۰/۳ تا ۲/۸
۱۱	Selector field	اسمی	فیلد انتخاب کننده ناسالم(۱)، سالم(۲)	۱ یا ۲

برای اجرای مدل ترکیبی باید در ابتدا مقداردهی اولیه انجام گیرد. تعداد تکرار و جمعیت اولیه در مدل ترکیبی به ترتیب برابر با ۵۰۰ و ۱۰۰ هستند. در جدول(۲)، نتایج مدل ترکیبی برمبنای تکرارهای مختلف نشان داده شده است. درصد صحت در مدل ترکیبی برای ۱۰۰ و ۵۰۰ بار تکرار برروی مجموعه داده‌های آموزشی بترتیب برابر با ۹۰/۷ درصد و ۹۴/۱ درصد است. همچنین درصد صحت در مدل ترکیبی برای ۱۰۰ و ۵۰۰ بار تکرار برروی مجموعه داده‌های تست بترتیب برابر با ۹۱/۱ درصد و ۹۴/۳ درصد است. نتایج مجموعه داده‌های تست بیانگر این است که مدل ترکیبی توانسته است که مجموعه داده‌های آموزش را با دقت بالایی طبقه‌بندی نماید و لذا در طبقه‌بندی مجموعه داده‌های تست، مشکل تشخیص نوع کلاس برای نمونه‌های جدید، کمتر می‌باشد. اگر تعداد تکرار افزایش یابد مدل ترکیبی بهتر می‌تواند نقاط مشابه را کشف کند و همچنین بروزرسانی موقعیت در الگوریتم کرم شب‌تاب برای پیدا کردن فاصله نزدیک در بهترین حالت انجام می‌گیرد. در مدل ترکیبی به دلیل اینکه حجم مجموعه داده‌ها بزرگ نیست، مشکل بیش‌آموزش(over-training) رخ نخواهد داد. هنگامی که آموزش بیش از حد صورت گیرد مشکل بیش‌آموزش رخ می‌دهد. در صورتی که تعداد تکرارهای مدل ترکیبی بیشتر از ۵۰۰ بار باشد مدل به سمت پدیده بیش‌آموزش پیش می‌رود و بهبودی در جواب نهایی حاصل نخواهد شد. در نتیجه بهترین تکرار برای مدل ترکیبی برابر با ۵۰۰ تکرار است.

جدول۲: نتایج مدل ترکیبی برمبنای تعداد تکرار

مجموعه داده	تعداد تکرار	دقت	بازخوانی	F-Measure	صحت
مجموعه داده آموزش	۱۰۰	۸۸/۸۵	۸۹/۷۳	۸۹/۲۹	۹۰/۶۸
	۲۰۰	۸۹/۲۳	۹۰/۱۲	۸۹/۶۷	۹۱/۴۵
	۳۰۰	۹۱/۰۶	۹۲/۱۷	۹۱/۶۱	۹۲/۸۵
	۴۰۰	۹۲/۱۳	۹۲/۵۱	۹۲/۳۲	۹۳/۸۵
	۵۰۰	۹۳/۲۷	۹۳/۸۴	۹۳/۵۵	۹۴/۱۵
مجموعه داده تست	۱۰۰	۸۹/۱۱	۸۹/۴۷	۸۹/۲۹	۹۱/۱۳
	۲۰۰	۸۹/۴۲	۸۹/۷۵	۸۹/۵۸	۹۱/۸۹
	۳۰۰	۹۱/۲۳	۹۱/۷۸	۹۱/۵۰	۹۳/۰۵
	۴۰۰	۹۲/۳۶	۹۲/۴۴	۹۲/۴۰	۹۴/۰۸
	۵۰۰	۹۳/۵۷	۹۳/۹۲	۹۳/۷۴	۹۴/۳۸

در جدول(۳)، نتایج مدل ترکیبی با ۵۰۰ بار تکرار و برمبنای تعداد ویژگی‌های انتخاب شده نشان داده شده است. همچنین جمعیت اولیه در الگوریتم کرم شب‌تاب برابر با ۱۰۰ می‌باشد. تعداد جمعیت کم باعث انتخاب ویژگی‌های نامرتبط می‌گردد که می‌تواند باعث افزایش محاسبات و کاهش دقت پیش‌بینی گردد. از طرف دیگر اطلاعات کافی به‌منظور آموزش سیستم هوشمند فراهم نمی‌گردد؛ بنابراین یک مقدار بهینه می‌تواند باعث افزایش دقت پیش‌بینی گردد. لذا با ۳۰ بار اجرای برنامه و تغییر جمعیت اولیه به این نتیجه رسیدیم که تعداد جمعیت اولیه اگر برابر با ۱۰۰ باشد دقت طبقه‌بندی بالا است. جدول(۳) نشان می‌دهد که اگر تعداد ویژگی‌های منتخب کم باشند درصد صحت بیشتر است و همچنین، نوع ویژگی‌ها در درصد صحت موثر هستند. برای مثال برای پنج ویژگی درصد صحت برابر با ۹۷/۲ و ۹۸/۶ درصد است. ویژگی‌هایی که در جدول(۳) نشان داده شده‌اند، ویژگی‌هایی هستند که توسط الگوریتم کرم شب‌تاب انتخاب شده‌اند و طبقه‌بندی برمبنای آنها انجام شده است.

جدول ۳: نتایج مدل ترکیبی برمبنای انتخاب ویژگی

تعداد ویژگی	ویژگی‌ها	درصد صحت
۵	Age, Gender, DB, SGPT, SGOT	۹۷/۲
۶	Age, Gender, DB, Alkphos, SGOT, ALB	۹۶/۸
۵	Age, Gender, TB, TP, ALB	۹۸/۶
۶	TB, DB, Alkphos, SGPT, TP, A/G	۹۶/۹
۷	DB, Alkphos, SGPT, SGOT, TP, ALB, A/G	۹۶/۲
۸	Age, Gender, TB, DB, Alkphos, SGOT, TP, A/G	۹۵/۵
۷	TB, DB, SGPT, SGOT, TP, ALB, A/G	۹۶/۳
۷	DB, Alkphos, SGPT, SGOT, TP, ALB, A/G	۹۳/۶
۸	Age, Gender, TB, DB, Alkphos, SGPT, SGOT, TP	۹۴/۰
۸	Age, Gender, TB, SGPT, SGOT, TP, ALB, A/G	۹۴/۳
۹	Age, Gender, TB, Alkphos, SGPT, SGOT, TP, ALB, A/G	۹۵/۲
۹	Age, Gender, TB, DB, Alkphos, SGPT, SGOT, TP, A/G	۹۵/۱
۱۰	Age, Gender, TB, DB, Alkphos, SGPT, SGOT, TP, ALB, A/G	۹۴/۱

در جدول(۳)، نتیجه درصد صحت بیماری کبد برمبنای ویژگی‌های مختلف نشان داده شده است. نتایج حاکی از آن است که در حالت اول برای شش ویژگی (Age, Gender, DB, Alkphos, SGOT, ALB) درصد صحت برابر با ۹۶/۹ درصد است و در حالت دوم برای شش ویژگی (TB, DB, Alkphos, SGPT, TP, A/G) درصد صحت برابر با ۹۶/۹ درصد است.

در جدول(۴)، مقایسه مدل ترکیبی با مدل‌های دیگر نشان داده شده است. درصد صحت درخت نیوی بیز [۲۶] برروی ۵۸۳ نمونه برابر با ۶۷ درصد است که مدل ترکیبی با همه ویژگی‌ها در مقایسه با آن ۲۷/۱ درصد بهبود داشته است و با شش ویژگی در حدود ۲۹/۹ درصد در مقایسه با درخت نیوی بیز بهبود داشته است. درخت نیوی بیز ترکیب الگوریتم نیوی بیز با درخت تصمیم‌گیری است. هدف الگوریتم درخت نیوی بیز این است که الگوریتم نیوی بیز بهبود داده شود. الگوریتم نیوی بیز در مقایسه با درخت تصمیم‌گیری برای مقابله با مجموعه داده بزرگ طراحی نشده است. علاوه بر آن، درصد صحت ترکیب درخت C5.0 و نزدیکترین همسایه [۲۸] با چهار ویژگی (DB, Age, Alkphos and SGOT) برابر با ۷۱ درصد است که مدل ترکیبی با همه ویژگی‌ها در مقایسه با آن ۲۳/۱

درصد بهبود داشته است. همچنین، درصد صحت درخت C5.0 [۲۹] برابر با ۸۷/۹ درصد است که درصد صحت مدل ترکیبی با همه ویژگی‌ها برابر با ۹۴/۱ است و در مقایسه با آن در حدود ۶/۲ درصد بهبود داشته است.

نتایج مقایسه برای همه مدل‌ها برمبنای یک مجموعه داده انجام شده است. در جدول(۴) نشان داده شده که درصد صحت ماشین بردار پشتیبان-شبکه عصبی مصنوعی چندلایه [۳۰] با شش ویژگی (SGPT, SGOT, TB, DB, Alkphos and Age) برابر با ۹۸/۸ درصد است و در مقایسه با مدل ترکیبی بیشتر است. همچنین، درصد صحت درخت تصمیم‌گیری C5.0 برمبنای بوستینگ [۱۲] برابر با ۹۳/۷ درصد است که درصد صحت مدل ترکیبی با همه ویژگی‌ها برابر با ۹۴/۱ است. درصد صحت ترکیب شبکه عصبی مصنوعی با الگوریتم رای‌گیری اکثریت [۳۲] برابر با ۷۱/۵ درصد است که مدل ترکیبی با همه در مقایسه با آن ۲۲/۶ درصد بهبود داشته است. مدل ترکیبی در مقایسه با اغلب مدل‌ها درصد صحت بیشتری دارد و به عنوان یک طبقه‌بند کارا برای تشخیص بیماری کبد مناسب است.

جدول ٤: مقایسه مدل ترکیبی با مدل‌های دیگر

مراجع	مدل‌ها	ویژگی‌ها	درصد صحت
[۲۳]	شبکه عصبی مصنوعی چندلایه	Age, Sex, SGOT, SGPT and ALP	۹۸
[۲۴]	K نزدیک‌ترین همسایه	Alkphos, SGPT and SGOT	۷۲/۲
[۲۵]	ماشین بردار پشتیبان	TB, DB,TP, ALB and A/G	۹۹/۷
[۲۶]	درخت نیوی بیز	TB, DB, Alkphos, SGOT, SGPT and ALB	۶۷
[۲۷]	درخت تصمیم‌گیری	همه ویژگی‌ها	۶۹/۴
[۲۸]	ترکیب درخت C5.0 و k نزدیکترین همسایه	DB, Age, Alkphos and SGOT	۷۱
[۲۹]	درخت تصمیم‌گیری C5.0	همه ویژگی‌ها	۸۷/۹
[۳۰]	ماشین بردار پشتیبان- شبکه عصبی مصنوعی چندلایه	SGPT, SGOT, TB, DB, Alkphos and Age	۹۸/۸
[۳۱]	شبکه عصبی مصنوعی چندلایه	همه ویژگی‌ها	۷۹/۳
[۱۲]	درخت تصمیم‌گیری C5.0 برمبنای بوستینگ	همه ویژگی‌ها	۹۳/۷
[۳۲]	ترکیب شبکه عصبی مصنوعی با الگوریتم رای‌گیری اکثریت	همه ویژگی‌ها	۷۱/۵
-	مدل پیشنهادی	همه ویژگی‌ها	۹۴/۱

بحث

در این مقاله از الگوریتم کرم شب‌تاب جهت انتخاب ویژگی نمونه‌ها استفاده گردید. ابتدا در مرحله پیش‌پردازش، داده‌های مربوطه نرمال سازی شدند و در مرحله بعد انتخاب ویژگی بین داده‌ها با کمک الگوریتم کرم شب‌تاب و در نهایت طبقه‌بندی با استفاده از الگوریتم آدابوست انجام گرفت. ارزیابی مدل ترکیبی برمبنای تعداد تکرار و تعداد ویژگی‌ها انجام شد. نتایج نشان داد که درصد صحت در مدل ترکیبی با تعداد تکرار رابطه مستقیم دارد و با افزایش تعداد تکرارها، درصد صحت افزایش یافته است.

بر اساس نتایج مطالعات موجود در مدل‌های درخت تصمیم‌گیری (**٦٩/٤**) [۲۷]، درخت تصمیم‌گیری C5.0 (**۸۷/۹**) [۲۹]، شبکه عصبی مصنوعی چندلایه(**۷۹/۳**) [۳۱] درخت تصمیم‌گیری C5.0 برمبنای بوستینگ (**۹۳/۷**) [۱۲] و ترکیب شبکه عصبی مصنوعی با الگوریتم رای‌گیری اکثریت (**۷۱/۵**) [۳۲] می‌توان نتیجه گرفت که مدل ترکیبی (**۹٤/۱**)، درصد صحت بیشتری دارد و این نشان‌دهنده دقت زیاد و کارایی خوب مدل ترکیبی است. نتایج شبکه عصبی مصنوعی

چندلایه [۲۳] با پنج ویژگی برابر با **۹۸** درصد، k نزدیکترین همسایه [۲۴] با سه ویژگی برابر با **۷۲/۲** درصد، ماشین بردار پشتیبان [۲۵] با پنج ویژگی برابر با **۹۹/۷** درصد، درخت نیوی بیز [۲۶] با شش ویژگی برابر با ۶۷ درصد، ترکیب درخت C5.0 و k نزدیکترین همسایه [۲۸] با چهار ویژگی برابر با **۷۱** درصد، و ماشین بردار پشتیبان- شبکه عصبی مصنوعی چندلایه [۳۰] با شش ویژگی برابر با **۹۸/۸** درصد و مدل ترکیبی با پنج ویژگی برابر با **۹۸/٦** درصد می‌باشد.

از محدودیت‌های مطالعه حاضر می‌توان به عدم بکارگیری مدل ترکیبی برای مجموعه داده‌های مختلف بیماری کبد و همچنین عدم ثبت نادرست برخی صفات مانند سن بیمار، مقدار بیلی‌روبین، مقدار آلبومین و احتمال خطا در ثبت داده‌ها اشاره نمود.

مدل ترکیبی پیشنهاد شده در این مطالعه، یک سیستم تشخیص بیماری کبد می‌باشد. امید است که مدل ترکیبی پیشنهاد شده در مراحل بعدی، جهت استفاده در تشخیص بیماری کبد در بیمارستان‌های مختلف ایران که نظام آنها مبتنی بر تشخیص بیماری در مراحل ابتدایی می‌باشد کاربرد داشته باشد. از آنجاییکه نتایج این مطالعه وابسته به مجموعه داده ILPD می‌باشد، پیشنهاد می‌شود برای بررسی بیشتر در این زمینه، در مطالعه‌های بعدی از مجموعه داده‌های مراکز بیمارستانی دیگر و یا الگوریتم‌های دیگر استفاده گردد. همچنین ویژگی‌های سن، بیلی‌روبین و آلبومین می‌توانند به عنوان ویژگی‌های تاثیرگذار بر بیماری کبد مورد بررسی قرار گیرند و بهتر است در مطالعه‌های بعدی، محدوده این ویژگی‌ها در نظر گرفته شوند. همچنین برای کارهای آتی پیشنهاد می‌شود که از طبقه‌بندی فازی برای طبقه‌بندی بیماران استفاده شود. روش طبقه‌بندی فازی به دلیل نسبت دادن میزان درجه تعلق نمونه‌ها در هر دسته از دقت قابل قبولی در تشخیص بیماران برخوردار خواهد بود.

ملاحظات اخلاقی

رعایت دستورالعمل‌های اخلاقی: مطالعه حاضر از نوع زیست پزشکی نمی‌باشد.

حمایت مالی: مطالعه حاضر از جایی حمایت مالی نشده است.

تضاد منافع: نویسندگان اظهار داشتند که تضاد منافعی وجود ندارد.

تشکر و قدردانی: مقاله حاضر مستخرج از پایان‌نامه با عنوان **"بهبود عملکرد دسته‌بند آدابوست برای تشخیص بیماری کبد"**، در مقطع کارشناسی ارشد، رشته مهندسی کامپیوتر، گرایش نرم‌افزار، در سال ۱۳۹۸، در دانشگاه آزاد اسلامی واحد ارومیه است.

References

1. Feizabadi M, Vaziri E, Haseli D. Analysis of the Factors Influencing Citations in Systematic Reviews of Medical Research in Iran. JHA. 2017; 20 (68): 86-98.
2. Jahani J, Rezaeenoor M, Mahdavi M, Hadavandi E. Prediction of diabetes by Neural Network. JHA. 2017; 20 (67):24-35.
3. Rezaii Farokh Zad M, Soleimanian Gharehchopogh F. Determining Fuzzy Logic Parameters by using Genetic Algorithm for the Diagnosis of Liver Disease. Journal of Health and Biomedical Informatics. 2018; 5 (3):384-397.
4. Jin XY, Jin QL, Yang X. A Disease Detection Method of Liver Based on Improved Back Propagation Neural Network. 8th International Symposium on Computational Intelligence and Design (ISCID). 2015; 2: 111-113.
5. Kumar SS, Devapal D. Survey on recent CAD system for liver disease diagnosis, International Conference on Control. Instrumentation, Communication and Computational Technologies (ICCICCT). 2014; 763-766.
6. Sebastian A, Varghese SM. Fuzzy logic for Child-Pugh classification of patients with cirrhosis of liver. International Conference on Information Science (ICIS); 2016; 168-171.
7. Lee CC, Chen SH, Chiang YC. Automatic Liver Diseases Diagnosis for CT Images Using Kernel-Based Classifiers, World Automation Congress. 2006; 1-5.
8. Ribeiro RT, Marinho RT, Sanches JM. Classification and Staging of Chronic Liver Disease from Multimodal Data. IEEE Transactions on Biomedical Engineering. 2013; 60(5):1336-1344.
9. Heydari M, and Teymouri M. [Prediction of Hepatic Failure Using Artificial Neural Network and Genetic Algorithm]. National Computer Engineering Conference and Sustainable Development with a Focus on Computer Networks, Modeling and Systems Security, Mashhad, Khavaran Higher Education Institution. 2014. (In Persian)
10. Lin R, and Chuang C. A Hybrid Diagnosis Model for Determining the Type of the Liver Disease, Computers in Biology and Medicine; 2010; 40: 665-670.
11. Pahareeya J, Vohra R, Makhijani J, and Patsariya S. Liver Patient Classification using Intelligence Techniques, International Journal Of Advanced Research in Computer Science and Software Engineering. 2014; 295-299.
12. Abdar M, Zomorodi-Moghadam M, Das R, Ting IH. Performance analysis of classification algorithms on early detection of liver disease, Expert Systems with Applications. 2017; 67: 239-251.
13. Mazaheri P, Norouzi A, Karimi A, and Kazemi M. [Using Decision Tree Algorithm for Early Detection of Hepatic Disease]. Second National Conference on Technology, Energy, and Data with the

Approach of Electrical and Computer Engineering. Kermanshah. IEEE Association. Kurdistan Student Branch. 2016. (In Persian).

14. Samavat M, and Safara F. [A Comprehensive Intelligent System for Diagnosis of Liver Disease]. 2nd International Knowledge Based Research Conference in Computer Engineering and Information Technology. Tehran, Majlisi University. 2017. (In Persian)
15. Christopher J, Nehemiah HK and Kannan A. A Swarm Optimization Approach for Clinical Knowledge Mining. Computer Methods and Programs in Biomedicine; 2015. 1-43.
16. https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)
17. Yang XS. Nature-Inspired Meta-heuristic Algorithms, Luniver Press. 2008.
18. Freund Y, and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 1997; 5(1): 119-139.
19. Jain S, Shukla S, Wadhvani R. Dynamic selection of normalization techniques using data complexity measures. Expert Systems with Applications. 2018; 106: 252-262.
20. Han J, & Kamber M. Data mining: Concepts and techniques, Morgan Kuafmann Publish, 2006.
21. Wu H, Yang S, Huang Z, He J, Wang X, Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 2018; 10:100-107.
22. Edla DR, Cheruku R, Diabetes-Finder: A Bat Optimized Classification System for Type-2 Diabetes. Procedia Computer Science. 2017; 115: 235-242.
23. Ramana BV, Babu MSP, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems. 2011; 3: 101-114.
24. Ramana BV, Babu MSP, Venkateswarlu NB. A critical comparative study of liver patients from USA and India: An exploratory analysis. International Journal of Computer Science Issues. 2012; 9: 506-516.
25. Tiwari AK, Sharma LK, & Krishna GP. Comparative Study of Artificial Neural Network based Classification for Liver Patient. Journal of Information Engineering and Applications. 2013; 3: 2225-0506.
26. Alfisahrin SDNN, & Mantoro T. Data Mining Techniques for Optimization of Liver Disease Classification. In 2013 International Conference on Advanced Computer Science Applications and Technologies. 2013; 379-384.
27. Jin H, Kim S, & Kim J. Decision factors on effective liver patient data prediction. International Journal of Bio-Science and Bio-Technology. 2014; 6:167-178.
28. Montazeri M, Montazeri M, Beygzadeh A, Zahedi MJ. Identifying efficient features in diagnose of liver disease by decision tree models. HealthMED. 2014; 8: 1115-1124.
29. Abdar M. A Survey and Compare the Performance of IBM SPSS Modeler and Rapid Miner Software for Predicting Liver disease by Using Various Data Mining Algorithms. Cumhuriyet Science Journal. 2015; 36:3230-3241.

30. Nagaraj K, Sridhar A. NeuroSVM: A Graphical User Interface for Identification of Liver Patients. ArXiv preprint arXiv: 1502.05534; 2015.
31. Weng CH, Huang TCK, Han RP. Disease prediction with different types of neural network classifiers. Telematics and Informatics. 2016; 33: 277-292.
32. Bashir S, Qamar U, Khan FH, Naseem L. HMV: A medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science. 2016; 13: 10-25.
33. Raghuwanshi B.S, Shukla S. Class imbalance learning using UnderBagging based kernelized extreme learning machine, Neurocomputing. 2019; 329:172-187.
34. Chawla NV. Data Mining for Imbalanced Datasets: An Overview. Data mining know discov handbook. 2005.
35. Sun Y, Wong AKC, Kamel MS. Classification of Imbalancd Data: A Review. Int J Patt Recogn Artif Intell. 2009; 4:687-719.
36. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging- Boosting- and HybridBased Approaches. IEEE Trans on Syst Man Cyber Part C AppRevi.2012; 4:463-84.
37. Barandela R, Sanchez JS, Garcia V, Rangel E. Strategies for learning in class imbalance problems. Patt Recogn. 2003; 3:849-51.
38. Napierała K, Stefanowski j, Wilk S. Learning from Imbalanced data in presence of noisy and borderline examples. In: Szczuka M, Kryszkiewicz M, Ramanna S, Jensen R, Hu Q, editors. RSCTC, LNAI 6086. Proceedingof 7th International Conference; 2010June 28-30; Warsaw, Poland. 2010; 158-167.
39. Zhang S, Liu L, Zhu X, Zhang C. A strategy for attributes selection in costsensitive decision trees induction. Proceeding of IEEE 8th International Conference onComputer and Information Technology Workshops. 2008; 8(11): 8-13.
40. Li DC, Liu CW, Hu SC. A learning method for the class imbalance problem with medical data sets. J Comput Bio Medi. 2010; 5: 509-518.
41. Rahman MM, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. Int J Machine Learning and Computer. 2013; 2: 224-8.
42. Cao XH, Stojkovic I, and Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinformatics. 2016; 17(1): 2-10.
43. Selvakumar B, Muneeswaran K. Firefly algorithm based feature selection for network intrusion detection. Computers & Security. 2019; 81:148-155.
44. Emary E, Zawbaa H.M, & Hassanien, AE. Binary ant lion approaches for feature selection. Neurocomputing. 2016; 213: 54-65.