مدل ترکیبی برمبنای الگوریتم بهینه سازی شیرمورچه و k نزدیک ترین همسایه برای تشخیص بیماری کبد

شایان جوادزاده[۱]، هومن شایان فر[۲]، فرهاد سلیمانیان قره چپق[۲*]

۱) گروه مهندسی کامپیوتر، موسسه آموزش عالی کمال ارومیه، ارومیه، ایران

۲) گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران

تاریخ دریافت: ۱۳۹۸/۱۱/۱۲ تاریخ پذیرش: ۱۳۹۹/۶/۱۱

چکیده

مقدمه: از آن جایی که کلیه بیمارستان ها اعم از دولتی و خصوصی، هزینه های سنگینی را در بخش بیماری کبد تقبل می کنند، ارائه روشی به منظور پیش بینی بیماری کبد ضرورتی اجتناب ناپذیر است. در این مقاله، مدل ترکیبی بر مبنای الگوریتم بهینه سازی شیرمورچه و k نزدیک ترین همسایه به منظور تشخیص بیماری کبد ارائه می گردد.

مواد و روش ها: در این مطالعه توصیفی-تحلیلی یک مدل ترکیبی مبتنی بر الگوریتم های یادگیری ماشین برای طبقه بندی افراد به دو دسته سالم و مبتلا به بیماری کبد طراحی شده است. مدل پیشنهادی با استفاده از نرم افزار MATLAB شبیه سازی شده است. مجموعه داده مورد استفاده در این مقاله، مجموعه داده ILPD موجود در مخزن داده یادگیری ماشین دانشگاه ایروین کالیفرنیا است. این مجموعه داده شامل شامل ۵۸۳ رکورد مستقل شامل ۱۰ ویژگی برای بیماری کبد است.

یافته های پژوهش: داده های این مجموعه پس از پیش پردازش به صورت تصادفی به ۲۰ دسته از کل مجموعه داده تقسیم شدند که شامل داده های آموزش و آزمون متفاوت بودند. در هر دسته داده از ۹۰ درصد داده ها برای آموزش و ۱۰ درصد باقی مانده برای آزمایش استفاده شد. نتایج حاصله در بهترین حالت بر مبنای تمامی ویژگی ها بر اساس درصد صحت برابر با ۹۵/۲۳ درصد و بر مبنای معیارهای ویژگی و حساسیت درصد صحت به ترتیب برابر ۹۳/۹۵ درصد و ۹۴/۱۱ درصد می باشد. هم چنین درصد صحت مدل پیشنهادی با ۵ ویژگی برابر با ۹۸/۶۳ درصد می باشد.

بحث و نتیجه گیری: مدل پیشنهادی به منظور تشخیص و طبقه بندی بیماری کبد با دقت بالای ۹۰ درصد پیشنهاد گردید. نتایج حاصل از این مقاله می تواند برای مراکز درمانی و پزشکان مفید واقع شود.

واژه های کلیدی: تشخیص بیماری کبد، الگوریتم بهینه سازی شیرمورچه، الگوریتم k نزدیک ترین همسایه، طبقه بندی

* نویسنده مسئول: گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران

Email:bonab.farhad@gmail.com

[DOI: 10.29252/sjimu.28.5.76]

مقدمه

بیماری کبد جزء شایع ترین، خطرناک ترین و پرهزینه ترین بیمـــــاری ها در جهـــان و نیز کشور ما می باشد و زیان های جانی و مالی بسیاری را به جامعه وارد می سازد(۱). امروزه افراد زیادی به علت بیماری کبد در وضعیت ناخوشایندی به سر می برند. این بیماری با توجه به شیوع و سهمی که در مرگ و میر انسان ها دارد از اهمیت بالایی برخوردار است. دلیل اهمیت آن این است که برای درمان بیماری و یا ارائه روشی برای آن نیاز به تشخیص درست بیماری می باشد که اگر بیماری به هر دلیل درست تشخیص داده نشود می تواند منجر به مرگ بیمار شود(۲). رشد چشمگیر بیماری کبد و اثرات و عوارض آن و هزینه های بالایی که بر جامعه وارد می کند، باعث شده که جامعه پزشکی به دنبال برنامه هایی جهت بررسی بیشتر، پیشگیری، تشخیص زود هنگام و درمان موثر آن باشد. بنا بر این اگر سیستم دقیق و خـــبره ای موجود باشد که بتواند با توجه به ویژگی های بیمار پیش بینی درستی از بیماری کبد داشته باشد، کمک حائز اهمیتی به بیماران خواهد کرد(۲).

اهمیت تشخیص بیماری ها یک اصل مهم در علم پزشکی می باشد که بر مبنای آزمایشات گوناگون بر روی بیماران انجام می گیرد(۳). اگر تعداد ویژگی ها در تشخیص بیماری زیاد باشند ممکن است تشخیص بیماری حتی برای یک متخصص خبره پزشکی نیز به سختی ممکن پذیر باشد(۴). همین دلیل موجب شده است که در چند دهه اخیر ابزار تشخیص کامپیوتری با هدف کمک به پزشکان مورد استفاده قرار گیرد تا میزان اهمیت هر یک از ویژگی ها به منظور تشخیص دقیق تر بیماری مشخص شـــود(۵،۶). سپـــــس ویژگی های تاثیرگذار، شناسایی و تشخیص داده می شوند. مجموعه داده های پزشکی حاوی حجم زیادی از اطلاعات درباره بیماران و وضعیت پزشکی آن ها می باشند. استخراج دانش مفید و تصمیم سازی های علمی برای تشخیص و بررسی بیماری به کمک این مجموعه داده ها می تواند بسیار مفید واقع شود(۷). استفاده از دانش موجود در مجموعه داده های پزشکی موجب کاهش خطاهای احتمالی و هزینه های درمانی اضافی می شود.

به منظور انجام این مقاله، ابتدا مطالعه اجمالی در زمینه تحقیقات سایر محققان در حوزه تشخیص بیماری کبد انجام می گیرد. کلیه جوانب از لحاظ راهکار به کار رفته و میزان دقت روش های ارائه شده و کیفیت مطلوب آن ها در تشخـــیص بـــیماری کبد مورد بررسی قرار می گیرد. آبدار و همکاران(۸) از درخت تصمیم گیری C5.0 برای تشخیص بیماری کبد استفاده کرده اند. برای تشخیص بیماری کبد از ۲۰ قانون مختلف استفاده شده ست. هر قانون بر مبـــنای عملگرهای and و مقدار ویژگی ها ارزیابی شده است. ارزیابی برروی مجموعه داده ILPD با ۵۸۳ نمونه انجام شده است. نتایج نشان داده که درصد صحت درخت C5.0 برابر با ۹۳/۷۵ بوده است.

جلوداری و همکاران(۹) یک مدل ترکیبی بر مبنای بهینه سازی اجتماع ذرات و ماشین بردار پشتیبان برای تشخـــیص بیـــماری کبد پیشنهاد داده اند. الگوریتم بهینه سازی اجتماع ذرات برای تجزیه و تحلیل داده ها با تعداد زیادی ویژگی استفاده می شود. این الگوریتم برای بهینه سازی توابع، کاهش ابعاد استفاده می شود. در مدل ترکیبی از الگوریتم بهینه سازی اجتماع ذرات برای انتـــخاب ویژگی و از ماشین بردار پشتیبان برای طبقه بندی نمونه ها استفاده می شوند. نتایج بر روی ۵۸۳ نمونه نشان داده که درصد صحت مدل ترکیبی برابر با ۹۴/۴۲ درصد بوده است.

پورپناه و همکاران(۱۰) در تحقیقی که بر روی بیماری کبد انجام داده اند به این نتیجه دست یافته اند که استفـــاده از الگوریتم یادگیری تقویتی مبتنی بر عامل های چندگانه راهکار مناسبی برای تشخیص بیماری کبد می باشد. آن ها در تحقیقشان از عامل های یادگیرنده که فضای جستجو را به صورت موازی کاوش می کنند استفاده کرده اند. نتایج بر روی ۵۸۳ نمونه از مجموعه داده ILPD نشان داده که درصد صحت برابر با ۸۰/۸۲ درصد بوده است.

ونگ و همکاران(۱۱) برای تشخیص بیماری کبد از شبکه عصبی مصنوعی استفاده کرده اند. آموزش و تست داده ها به ترتیب برابر با ۸۰ و ۲۰ درصد بوده است. تعداد نرون های ورودی در شبکه عصبی مصنوعی برای تشخیص بیماری کبد برابر با ۱۰ نرون بوده و بر مبنای الگوریتم رای گیری، وزن نرون ها بهینه شده و در هر

[DOI: 10.29252/sjimu.28.5.76]

مرحله مقدار خطا کاهش یافته است. ارزیابی بر روی مجموعه داده کبد نشان داده که درصد صحت در بهترین حالت برابر با ۷۴/۴۸ درصد بوده است.

لیانگ و پنگ(۱۲) یک مدل ترکیبی بر مبنای الگوریتم ایمنی مصنوعی و ژنتیک برای تشخیص بیماری کبد پیشنهاد داده اند. در مدل ترکیبی شان از الگوریتم ژنتیک به منظور کشف راه حل های جدید استفاده شده است. الگوریتم ژنتیک بر مبنای عملگرهای ادغام و جهش سعی در بهبود دادن درصد صحت و کشف ویژگی هایی دارد که باعث افزایش درصد طبقه بندی می شوند. نتایج بر روی مجموعه داده کبد نشان داده که درصد صحت برابر با ۹۸/۱ درصد بوده است.

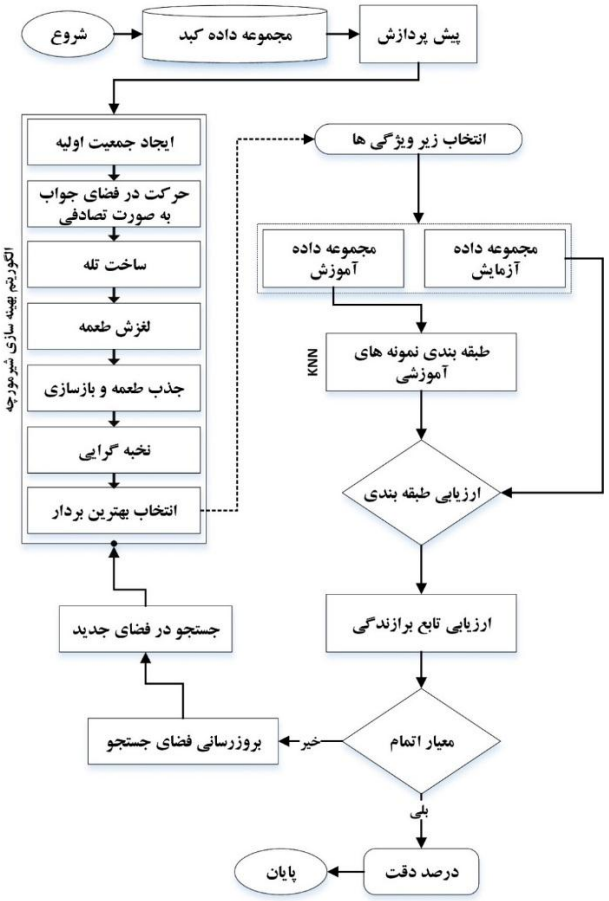
کومار و تانکور(۱۳) مدل ترکیبی بر مبنای فازی و k نزدیک ترین همسایه وزنی برای تشخیص بیماری کبد پیشنهاد داده اند. در مدل آن ها از الگوریتم فازی برای بهبود الگوریتم k نزدیک ترین همسایه وزنی استفاده شده و اوزان k نزدیک ترین همسایه وزنی به منظور دقت بیشتر بر مبنای فازی کشف شده اند. هدف وزن های بهینه این است که الگوریتم نمونه های مشابه به هم را پیدا کند و طبقه بندی افزایش یابد. نتایج نشان داده که درصد صحت برابر با ۸۴/۲۹ درصد بوده است.

هدف اصلی مقاله حاضر، ارائه یک سیستم تشخیص بیماری کبد برای مجموعه داده ILPD (۱۴) مبتنی بر ترکیب الگوریتم بهینه سازی شیرمورچه(۱۵) و طبقه بندی k نزدیک ترین همسایه(۱۶) است. در مدل پیشنهادی از الگوریتم بهینه سازی شیرمورچه برای انتخاب ویژگی(۱۷،۱۸) و از الگوریتم k نزدیک ترین همسایه برای طبقه بندی نمونه ها استفاده می شود. الگوریتم بهینه سازی شیرمورچه یکی از الگوریتم های فرا ابتکاری است که بر مبنای جمعیت اولیه و تکرار به راه حل بهینه دست می یابد. هم چنین الگوریتم k نزدیک ترین همسایه یکی از الگوریتم های داده کاوی است که برای طبقه بندی و تشخیص نمونه ها استفاده می شود(۱۹). در این مقاله به کمک مدل پیشنهادی، ویژگی های مهم بیماری کبد شناسایی می شوند.

مواد و روش ها

مطالعه حاضر، از نوع توصیفی-تحلیلی است. در این مطالعه بر مبنای ویژگی های ورودی به تشخیص وضعیت بیماران کبد از نظر سالم یا ناسالم بودن می پردازیم. مجموعه داده مورد استفاده در این مطالعه از مجموعه داده مربوط به بیماران مبتلا به کبد، موجود در مجموعه داده یادگیری ماشین دانشگاه ایروین، کالیفرنیا تامین شده است. مجموعه داده ها شامل ۵۸۳ رکورد کبد با ۱۰ ویژگی(و یک ویژگی متعلق به کلاس) می باشند. مجموعه داده کبد را می توان 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) می باشد. در شکل شماره ۱ فلوچارت مدل پیشنهادی نشان داده شده است.

[DOI: 10.29252/sjimu.28.5.76]



شکل شماره ۱. فلوچارت مدل پیشنهادی

مرحله اول: پیش پردازش

یک مرحله مـــهم در پیش پردازش، نرمال سازی داده ها است. هدف نرمال سازی این است که داده ها در یک محدوده مساوی تعریف شوند. در بسیاری از کاربردها، نایکسان بودن مقادیر ویژگی ها باعث بوجود آمدن بی کیفیتی جواب می شوند. برای مثال ویژگی سن بیمار دارای بازه متفاوتی است که مقادیر آن بر کارایی دیگر ویژگی ها تاثیر می گذارد و باعث می شود که تابع هزینه اثر نامطلوبی داشته باشد. برای رفع این مشکل می توان از نـــرمال ســـازی داده ها استفاده کرد. نرمال سازی داده ها با یک تبدیل خطی یا غیرخطی، داده ها را به بازه ای که معمولاً [۱,۱−] و [۰,۱] است، نگاشت می کند. اصولاً وارد کردن داده ها به صورت خام باعث کاهش سرعت الگوریتم طبقه بندی می شود. برای اجتناب از چنین شرایطی باید داده ها استاندارد شوند. استانداردسازی داده های اولیه در مدل پیشنهادی طبق معادله(۱) انجام گرفته است. در معادله(۱)، x_r ،x_n ،x_{max} و x_{min} به ترتیب نشان دهنده مقادیر واقعی، استاندارد شده، حداکثر و حداقل داده های تحت بررسی هستند(۲۰،۲۱).

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

مرحله دوم: انتخاب ویژگی

مرحـــله دوم انتخاب ویژگی مبتنی بر الگوریتم بهینه سازی شیـــرمورچه باینری می باشد. الگوریتم بهینه سازی شیرمـــورچه از عـــکس العمـــل بین شیرمورچه ها و مورچه های در دام افتـــاده اســـتفاده می کند. الگوریتم بهینه سازی شیرمورچه باینری شامل مراحل زیر است:

**مرحله اول(راه رفتن تصادفی مورچه ها):* ابتدا مورچه ها به عنوان حرکت در فضای جستجو در نظر گرفته می شوند، سپس شیرمورچه مجاز به شکار آن ها می شود. از آن جایی که مورچه ها به صورت تصادفی

[DOI: 10.29252/sjimu.28.5.76]

در محل حرکت می کنند، حرکت مورچه ها طبق معادله(۲) تعریف می شود.

$$X(t) = [0, cumsum(2r \times (t_1) - 1)\ cumsum(2r \times (t_2) - 1)\ \dots\ cumsum(2r \times (t_n) - 1)] \quad (۲)$$

در معادله(۲) برای تولید هر کدام از اعداد تصادفی که می توانند مثبت، منفی و صفر باشند از تابعی به نام cumsum استفاده می شود. پارامتر n نشان دهنده بیشترین تکرار، t بیانگر مرحله راه رفتن تصادفی است(t امین تکرار)، t(r) تابع تصادفی است که طبق معادله(۳) تعریف می شود.

$$\mathrm{r(t)} = \begin{cases} 1 & if\ rand > 0.5 \\ 0 & if\ rand \leq 0.5 \end{cases} \quad (۳)$$

در معادله(۳) t بیانگر مرحله قدم زدن تصادفی و rand عدد تصادفی در بازه [−۱,۱] می باشد. موقعیت مورچه ها در یک ماتریس طبق معادله(۴) ذخیره و در طول بهینه سازی مورد استفاده قرار می گیرد.

$$\mathrm{M}_{Ant} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & \cdots & A_{1,d} \\ A_{2,1} & A_{2,2} & \vdots & \vdots & A_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ A_{n,1} & A_{n,1} & & & A_{n,d} \end{bmatrix} \quad (۴)$$

در معادله(۴) $A_{i,j}$ مشخص کننده مقدار متغیر j–ام از مورچه i–ام است. پارامتر n تعداد مورچه ها و d تعداد متغیرها است. برای ارزیابی هر مورچه، یک تابع برازندگی طبق معادله(۵) در طول بهینه سازی تعریف می شود. بعد از مشخص شدن موقعیت هر مورچه، مقدار تابع هدف آن محاسبه می شود؛ آن گاه این مقادیر به صورت صعودی مرتب می شوند و در ماتریس $\mathrm{M_{OA}}$ ذخیره می شوند. در هر مرحله از بهینه سازی، مورچه ها با قدم زدن تصادفی موقعیت خود را به‌روز می کنند.

$$\mathrm{M}_{OA} = \begin{bmatrix} f(\{A_{1,1}, A_{1,2}, \cdots, A_{1,d}\}) \\ f(\{A_{2,1}, A_{2,2}, \cdots, A_{2,d}\}) \\ \vdots \\ \vdots \\ f(\{A_{n,1}, A_{n,2}, \cdots, A_{n,d}\}) \end{bmatrix} \quad (۵)$$

**مرحله دوم(ساختن تله توسط شیرمورچه و گرفتار شدن مورچه در آن):* با توجه به مطالب بیان شده قدم زدن مورچه تحت تاثیر تله های شیرمورچه قرار می گیرد. تله هایی که در نقاط بهینه هستند از برازش بهتری برخوردار هستند و لذا شانس بیشتری برای رفتن طعمه به این نقاط وجود دارد.

**مرحله سوم(لغزش طعمه به سمت شیرمورچه):* برای مدل کردن توانایی شکار شیرمورچه ها، از ساختار چرخ گردان استفاده می شود. الگوریتم بهینه سازی شیرمورچه نیاز به یک عملگر چرخ گردان برای تعیین شیرمورچه ها بر پایه تابع برازندگی آن ها در طول بهینه سازی دارد. پس فرآیند چرخ گردان معادل تله شیرمورچه جهت شکار مورچه است به این صورت که، هر چه تله بزرگ تر باشد به شیرمورچه امکان شکار مورچه های بیشتری می دهد.

**مرحله چهارم(جذب طعمه و بازسازی تله):* مرحله نهایی شکار زمانی است که طعمه به پایین گودال رسیده و در دهان شیرمورچه قرار می گیرد. پس از این مرحله شیرمورچه طعمه را به داخل گودال می کشد و می خورد. در استفاده از این فرآیند فرض می شود که شکار هنگامی انجام می گیرد که مورچه داخل ماسه فرو رفته باشد. سپس می بایست موقعیت شیرمورچه نسبت به موقعیتی که مورچه را شکار کرده است، برای افزایش شانس شکار جدید به روزرسانی شود.

**مرحله پنجم(نخبه گرایی):* نخبه گرایی یک ویژگی مهم از الگوریتم های تکاملی است که اجازه می دهد تا بهترین راه حل به دست آمده در هر مرحله از فرآیند بهینه سازی حفظ شود. در این مطالعه بهترین شیرمورچه به دست آمده در هر تکرار ذخیره شده است و به عنوان یک نخبه در نظر گرفته می شود. از آن جایی که نخبه ها مناسب ترین پاسخ های هستند باید قادر به تاثیرگذاری بر همه مورچه ها باشند. بنا بر این فرض می شود هر مورچه با ساختار چرخ گردان به یک شیرمورچه نزدیک می شود.

الگوریتم های فرا ابتکاری با یک جواب اولیه شروع به حل مسئله می نمایند و سپس به جستجوی جواب های مناسب تر می پردازند، تا جایی که دیگر قادر به یافتن جواب های بهتر نباشد، در این مرحله الگوریتم متوقف شده و نتیجه حاصل شده به عنوان جواب بهینه ذخیره می شود.

[DOI: 10.29252/sjimu.28.5.76]

با توجه به این که هر مسئله از تعداد زیادی ویژگی تشکیل شده است لذا باید ویژگی های تاثیرگذار در مسئله برای تولید جواب های بهینه انتخاب شوند. در زمینه انتخاب ویژگی، تعداد ویژگی در مسئله یک پارامتر مهم و تعیین کننده بوده و جواب های کسب شده بر اساس تعداد ویــژگی های موجود در شبکه رتبه بندی می شوند. از این رو برای ایجاد جواب اولیه، برداری شامل اعداد تصادفی مثبت بین صفر و یک تولید می شود. تعداد عناصر این بردار برابر با تعداد ویژگی های موجود در مجموعه داده بیماری خواهد بود و داخل عناصر بردار با اعداد تصادفی پر می شود. به عنوان مثال برای یک بردار با ویژگی های مختلف، برداری با n خانه به صورت شکل شماره ۲ ایجاد می کنیم که عناصر این بردار با اعداد تصادفی مثبت و کوچک تر از یک پر شده است. شیرمــــورچه ها موقعیت خود را بر مبنای تعداد شکار مورچه ها به روزرسانی می کنند. در مناطقی که تعداد به دام افتادن مورچه ها بالا باشد تابع هدف شیرمورچه ها به سمت آن مناطق تمایل خواهد داشت و موقعیت شیرمورچه ها برمبنای کشف همسایه های نقاط برازنده محاسبه می شود و موقعیتی به عنوان موقعیت بعدی انتخاب می شود که فاصله کمتری با موقعیت جاری دارد. مقدار موقعیت جدید بر مبنای مجموع مقادیر بردار محاسبه می شود و برداری که مجموع مقادیر آن بیشتر باشد به عنوان بردار برازنده انتخاب می شود.

۰/۶	۰/۲	۰/۳	۰/۷	۰/۲	۰/۱	۰/۹	۰/۸

شکل شماره ۲. نمایش راه حل اولیه

برای انتخاب ویژگی ها از بین عناصر بردار با اعداد تصادفی، عناصر با مقدار بزرگ تر به عنوان ویژگی انتخاب می شوند. مراحل ایجاد جواب اولیه برای انتخاب ویژگی با ۱۰ ویژگی به ترتیب نشان داده شده است. بنا بر این اگر مقدار یک ویژگی صفر شود، آن گاه این ویژگی با بیماری کبد رابطه معنادار ندارد. هم چنین مقدار یک در ویژگی ها نشان دهنده وجود ارتباط معنادار با بیماری کبد است. در شکل شماره ۳، مقادیر یک به عنوان ویژگی های انتخاب شده و مقادیر صفر به عنوان ویژگی انتخاب نشده در نظر گرفته می شوند.

۰/۶	۰/۲	۰/۳	۰/۷	۰/۲۱	۰/۳۲	۰/۶۲	۰/۹
۱	۰	۰	۱	۰	۰	۱	۱

شکل شماره ۳. نمایش ویژگی انتخاب شده

تابع برازندگی برای انتخاب ویژگی معادله(۶) تعریف می شود. در معادله(۶)، |n| تعداد کل ویژگی ها و |S| تعداد ویژگی های انتخاب شده است. پارامتر accuracy درصد صحت و مقدار پارامترهای δ و ρ ثابت هستند و مقدار آن ها به ترتیب برابر با ۹۹ و ۱ می باشد.

$$Fitness = \delta.Accuracy + \rho.\frac{|n| - |S|}{|n|} \quad (۶)$$

مرحله سوم: طبقه بندی

طبقه بندی یک روش مبتنی بر یادگیری است که در آن برای شناخت الگوهای ناشناختــه از قانون های IF-THEN استفاده می شود. هدف اصــلی کاربرد طبقه بنـــدی کمک به علم پزشکی در یافتن علائم بیـــماری ها و تلاش برای یافتن بهترین درمان به منظور کاهـــش هزینه های پرداختی و پیشگیری از مرگ است. یکی از روش های طبـــقه بندی، الگوریتم k نزدیک ترین همســـایه که به صورت گسترده در حوزه پزشـــکی و خصـــوصاً طبقه بندی داده های پزشکی جهت تشخیص بیـــماری استفاده می شود. در این الگوریتم با اســـتفاده از معـــادله(۷) فاصـــله اقلیدســـی بین نمونه جـــدید و هـــمه نمونه های مـــوجود در مجمـــوعه نمـــونه های یادگـــیری محـــاسبه انـــجام می شود.

[DOI: 10.29252/sjimu.28.5.76]

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \; ; \; x_1 = (x_{11}, x_{12}, \dots, x_{1n}) \, , x_2 = (x_{21}, x_{22}, \dots, x_{2n}) \quad (۷)$$

در این الگوریتم، k تا از نمونه های یادگیری که کمتـــرین فاصله را با نمونه جدید دارند، به عنوان همسایه های آن نمونه انتخاب می شوند. در بین این همسایه ها برچسب کلاسی که در اکثریت باشد به عنوان برچسب دسته این نمونه جدید پیش بینی می شود.

مرحله چهارم: معیارهای ارزیابی

مرحله چهارم، معیارهای ارزیابی می باشد. در مطالعه حاضر، جهت ارزیابی دقت طبقه بندی از معیارهای شاخص های ویژگی، دقت، بازخوانی یا حـــساسیت، F-Measure و صحت استفاده شده است که صحت معیار اصلی است(۲۲،۲۳). حال آن که معیار بازخوانی (recall) و معیار دقت(precision) و F-Measure به طور مجزا عملکرد طبقه بندی کننده را برای کلاس های مختلف نشان می دهند. هر چه میزان بازخوانی بالاتر باشد بـــیانگر این است که قابلیت شناسایی درست کلاس ها بیشتر است.

$$\textbf{Specificity} = \frac{\textbf{TN}}{\textbf{TN} + \textbf{FP}} * \textbf{100} \quad (۸)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100 \quad (۹)$$

$$\text{Sensitivity or Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100 \quad (۱۰)$$

$$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۱۱)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (۱۲)$$

یافته‌های پژوهش

مجموعه داده مورد استفاده در این مطالعه، مجموعه داده استانداردی به نام ILPD (۱۴) است که از شمال شرق آندراپرادش هندوستان جمع آوری 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 می باشد که از این میان ۴۱۶ نمونه(۷۲ درصد) پرونده بیماران کبدی (کلاس یک) و ۱۶۷ نمونه(۲۸ درصد) پرونده افراد سالم (کلاس دو) می باشد. از این ۵۸۳ نمونه تحت بررسی، ۴۴۱ نفر مرد و ۱۴۲ نفر زن هستند. این مجموعه داده شامل ده ویژگی و یک فیلد هدف می باشد. فیلد هدف یک برچسب کلاس است که مجموعه داده را به دو گروه (بیمار و سالم) تقسیم کرده است. ویژگی های این مجموعه داده در جدول شماره ۱ نشان داده شده است که در ادامه هر ویژگی نیز به صورت مختصر معرفی گردیده است.

جدول شماره ۱. ویژگی های مجموعه داده کبد(ILPD)

شماره ویژگی	نام ویژگی	نوع ویژگی	توضیحات	محدوده
۱	Age	عددی	سن	۴ تا ۹۰
۲	Gender	اسمی	جنس	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
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

به منظور تولید مجموعه آموزشی و مجموعه آزمایشی از روش 10-Fold Cross Validation استفاده شده است. در این روش مجموعه داده ها به صورت تصادفی به ۱۰ بخش مساوی تقسیم می شود که

[DOI: 10.29252/sjimu.28.5.76]

در هر تکرار، یک بخش به عنوان آزمایش و نه بخش دیگر به عنوان آموزش انتخاب شدند. یعنی در هر اجرا ۹۰ درصد داده ها به عنوان داده های آموزشی و ۱۰ درصد باقی مانده به عنوان آزمایش انتخاب شدند. در انتها میانگین ۱۰ بار تکرار الگوریتم به عنوان نتیجه نهایی انتخاب شد. برای اجرای مدل پیشنهادی باید در ابتدا مقداردهی اولیه انجام گیرد. تعداد تکرار و جمعیت اولیه در مدل ترکیبی به ترتیب برابر با ۳۰۰ و ۵۰ هستند. در جدول شماره ۲، نتایج مدل ترکیبی بر مبنای تکرارهای مختلف نشان داده شده است. درصد صحت در مدل ترکیبی برای ۱۰۰ و ۳۰۰ بار تکرار بر روی مجموعه داده های آموزشی به ترتیب برابر با ۹۰/۱۹ درصد و ۹۳/۶۱ درصد است. هم چنین درصد صحت در مدل پیشنهادی برای ۱۰۰ و ۳۰۰ بار تکرار بر روی مجموعه داده های آزمایشی به ترتیب برابر با ۹۱/۸۲ درصد و ۹۵/۱۹ درصد است. نتایج جدول شماره ۲ بر مبنای مجموعه داده های آموزشی و آزمایشی در ۳۰۰ تکرار، نتایج بهتری داشته است و به عنوان نتیجه نهایی ثبت شده است.

جدول شماره ۲. نتایج مدل پیشنهادی بر مبنای تعداد تکرار

مجموعه داده	تعداد تکرار	ویژگی	دقت	بازخوانی یا حساسیت	F-Measure	صحت
مجموعه داده آموزشی	۱۰۰	۸۹/۲۲	۸۹/۳۶	۸۹/۴۸	۸۹/۴۱	۹۰/۱۹
	۱۵۰	۸۹/۱۲	۸۹/۵۶	۸۹/۷۲	۸۹/۶۴	۹۱/۰۳
	۲۰۰	۹۰/۳۳	۹۰/۱۱	۹۰/۶۸	۹۰/۳۹	۹۱/۸۲
	۲۵۰	۹۱/۲۵	۹۱/۲۳	۹۱/۷۶	۹۱/۴۹	۹۳/۱۷
	۳۰۰	۹۲/۱۹	۹۱/۷۲	۹۲/۳۵	۹۲/۰۳	۹۳/۶۱
مجموعه داده آزمایشی	۱۰۰	۹۰/۴۵	۹۰/۰۸	۹۰/۵۱	۹۰/۲۹	۹۱/۸۲
	۱۵۰	۹۱/۰۹	۹۱/۱۵	۹۱/۳۴	۹۱/۲۴	۹۲/۱۵
	۲۰۰	۹۲/۴۶	۹۲/۸۷	۹۲/۹۶	۹۲/۹۱	۹۳/۶۸
	۲۵۰	۹۳/۹۵	۹۳/۳۸	۹۴/۰۶	۹۳/۷۲	۹۴/۰۷
	۳۰۰	۹۴/۳۷	۹۴/۲۵	۹۴/۶۹	۹۴/۴۷	۹۵/۱۹

در جدول شماره ۳، نتایج مدل پیشنهادی با ۳۰۰ بار تکرار و بر مبنای تعداد ویژگی های انتخاب شده نشان داده شده است. هم چنین جمعیت اولیه در الگوریتم بهینه سازی شیرمورچه برابر با ۵۰ می باشد. جدول شماره ۳ نشان می‌دهد که اگر تعداد ویژگی های منتخب کم باشند درصد صحت بیشتر است و هم چنین، نوع ویژگی ها در درصد صحت موثر هستند. برای مثال برای پنج ویژگی درصد صحت در بهترین حالت برابر با ۹۸/۶۳ درصد است.

[DOI: 10.29252/sjimu.28.5.76]

جدول شماره ۳. نتایج مدل پیشنهادی برمبنای انتخاب ویژگی

تعداد متغیر	ویژگی ها	درصد صحت	ویژگی	دقت	بازخوانی یا حساسیت	F-Measure
۵	TB, Alkphos, SGPT, TP, ALB	۹۸/۲۱	۹۷/۵۶	۹۷/۶۱	۹۷/۹۳	۹۷/۹۹
۵	DB, SGPT, SGOT, TP, A/G	۹۸/۶۳	۹۸/۳۱	۹۸/۲۴	۹۸/۴۵	۹۸/۳۴
۵	Age, DB, Alkphos, TP, ALB	۹۸/۵۲	۹۸/۰۲	۹۸/۱۹	۹۸/۱۳	۹۸/۱۶
۵	TB, DB, Alkphos, SGPT, TP	۹۸/۳۸	۹۸/۲۹	۹۸/۴۱	۹۸/۵۶	۹۸/۴۸
۶	Gender, TB, DB, SGPT, TP, ALB	۹۷/۸۱	۹۷/۱۹	۹۷/۳۲	۹۷/۶۷	۹۷/۴۹
۶	Age, DB, Alkphos, SGOT, TP, A/G	۹۷/۲۸	۹۶/۸۹	۹۷/۵۴	۹۷/۱۵	۹۷/۳۴
۷	Gender, TB, DB, SGPT, SGOT, TP	۹۷/۰۵	۹۶/۴۶	۹۶/۱۹	۹۶/۷۹	۹۶/۴۹
۷	DB, Alkphos, SGPT, SGOT, TP, ALB, A/G	۹۷/۲۴	۹۶/۳۷	۹۶/۴۹	۹۶/۶۲	۹۶/۵۵
۸	Age, TB, DB, Alkphos, SGOT, TP, ALB, A/G	۹۶/۸۴	۹۶/۱۱	۹۶/۴۳	۹۶/۵۱	۹۶/۴۷
۸	Gender, TB, DB, SGPT, SGOT, TP, ALB, A/G	۹۶/۲۹	۹۶/۰۴	۹۶/۰۹	۹۶/۱۷	۹۶/۱۳
۸	Age, Gender, TB, Alkphos, SGPT, SGOT, ALB, A/G	۹۶/۵۷	۹۶/۲۲	۹۶/۱۶	۹۶/۳۲	۹۶/۲۴
۹	Gender, TB, DB, Alkphos, SGPT, SGOT, TP, ALB, A/G	۹۶/۱۱	۹۵/۶۳	۹۶/۴۳	۹۶/۷۴	۹۶/۵۸
۹	Age, Gender, TB, DB, Alkphos, SGPT, SGOT, ALB, A/G	۹۵/۹۴	۹۵/۲۵	۹۵/۲۸	۹۵/۴۹	۹۵/۳۸
۹	Age, Gender, TB, DB, SGPT, SGOT, TP, ALB, A/G	۹۵/۶۱	۹۴/۵۳	۹۴/۳۳	۹۴/۳۸	۳۵,۹۴
۱۰	Age, Gender, TB, DB, Alkphos, SGPT, SGOT, TP, ALB, A/G	۹۵/۲۳	۹۳/۹۵	۹۴/۰۵	۹۴/۱۱	۹۴/۰۸

در جدول شماره ۳، نتیجه درصد صحت بیماری کبد بر مبنای ویژگی های مختلف نشان داده شده است. نتایج حاکی از آن است که در حالت اول برای شش ویژگی (Gender, TB, DB, SGPT, TP, ALB) درصد صحت برابر با ۹۷/۸۱ درصد است و در حالت دوم برای شش ویژگی(Age, DB, Alkphos, SGOT, TP, A/G) درصد صحت برابر با ۹۷/۲۸ درصد است. نتایج برای هشت ویژگی(Age, TB, DB, Alkphos, SGOT, TP, ALB, A/G) در حالت اول برابر با ۹۶/۸۴ درصد است. در حالت دوم برای هشت ویژگی (Gender, TB, DB, SGPT, SGOT, TP, ALB, A/G) برابر با ۹۶/۲۹ درصد است. در حالت سوم برای هشت ویژگی(Age, Gender, TB, Alkphos, SGPT, SGOT, ALB, A/G) برابر با ۹۶/۵۷ درصد است. از مقایسه نتایج به دست آمده از ویژگی ها می توان فهمید که اگر تعداد ویژگی ها کمتر باشند آن گاه درصد صحت بیشتر است و هم چنین نوع ویژگی ها هم در افزایش و کاهش درصد صحت موثر هستند.

جدول شماره ۴ مقایسه و ارزیابی مدل پیشنهادی با مدل های دیگر را نشان می دهد. درصد صحت درخت تصمیم گیری C5.0 (۸) با همه ویژگی ها برابر با ۹۳/۷۵ درصد بوده است. درصد صحت مدل پیشنهادی با همه ویژگی ها برابر با ۹۵/۲۳ است. درصد صحت جنگل تصادفی(۹) با نه ویژگی(ALB, A/g, Alkphos, TP, TB, DB, SGOT, Age, SGPT)، ماشین بردار پشتیبان(۹)، شبکه بیزین(۹) و شبکه عصبی مصنوعی چندلایه(۹) به ترتیب برابر با ۸۶/۲۶، ۷۵/۱۰، ۶۶/۰۹ و ۷۸/۱۱ درصد بوده است. درصد صحت ترکیب بهینه سازی اجتماع ذرات و ماشین بردار پشتیبان(۹) با هفت ویژگی(ALB, Alkphos, DB, TB, SGOT, Age, SGPT) برابر با ۹۴/۴۲ بوده است.

[DOI: 10.29252/sjimu.28.5.76]

جدول شماره ۴. مقایسه مدل پیشنهادی با مدل های دیگر

مراجع	مدل‌ها	ویژگی‌ها	درصد صحت
(۸)	درخت تصمیم گیری C5.0	همه ویژگی ها	۹۳/۷۵
(۹)	جنگل تصادفی	ALB, A/g, Alkphos, TP, TB, DB, SGOT, Age, SGPT	۸۶/۲۶
	ماشین بردار پشتیبان	همه ویژگی ها	۷۵/۱۰
	شبکه بیزین	همه ویژگی ها	۶۶/۰۹
	شبکه عصبی مصنوعی چندلایه	همه ویژگی ها	۷۸/۱۱
	ترکیب بهینه سازی اجتماع ذرات و ماشین بردار پشتیبان	ALB, Alkphos, DB, TB, SGOT, Age, SGPT	۹۴/۴۲
(۱۰)	یادگیری تقویتی مبتنی بر عامل های چندگانه	همه ویژگی ها	۸۰/۸۲
(۱۱)	شبکه عصبی مصنوعی	همه ویژگی ها	۷۴/۴۸
(۱۲)	ترکیب الگوریتم ایمنی مصنوعی و ژنتیک	همه ویژگی ها	۹۸/۱
(۱۳)	ترکیب فازی و k نزدیک ترین همسایه وزنی	همه ویژگی ها	۸۴/۲۹
(۲۴)	شبکه عصبی مصنوعی مبتنی بر درخت C5.0	همه ویژگی ها	۹۴/۱۲
(۲۵)	بگینگ	همه ویژگی ها	۷۱/۸۷
	آدابوست	همه ویژگی ها	۶۶/۵۵
	رای‌گیری اکثریت	همه ویژگی ها	۷۱/۵۳
(۲۶)	الگوریتم C-Means	همه ویژگی ها	۷۱/۳۵
(۲۷)	رگرسیون لجستیک	همه ویژگی ها	۷۲/۵۰
	درخت تصمیم گیری J48	همه ویژگی ها	۶۸/۷۸
	جنگل تصادفی	همه ویژگی ها	۷۱/۵۳
–	مدل پیشنهادی	همه ویژگی ها	۹۵/۲۳

درصد صحت یادگیری تقویتی مبتنی بر عامل های چندگانه(۱۰)، شبکه عصبی مصنوعی(۱۱)، ترکیب الگوریتم ایمنی مصنوعی و ژنتیک(۱۲)، ترکیب فازی و k نزدیک ترین همسایه وزنی(۱۳) به ترتیب برابر با ۸۰/۸۲ درصد، ۷۴/۴۸ درصد، ۹۸/۱ درصد، ۸۴/۲۹ درصد بوده است. درصد صحت شبکه عصبی مصنوعی مبتنی بر درخت C5.0 و رای گیری اکثریت به ترتیب برابر با ۹۴/۱۲ درصد و ۷۱/۵۳ درصد بوده است.

طبق جدول شماره ۴ درصد صحت در مدل پیشنهادی برابر ۹۵/۲۳ و در مدل ترکیبی الگوریتم ایمنی مصنوعی و ژنتیک(۱۲) برابر با ۹۸/۱ بوده است. ترکیب الگوریتم ایمنی مصنوعی و ژنتیک قادر بوده است ابهاماتی از قبیل داده های اضافی یا پیش پردازش که اغلب در تجزیه و تحلیل آزمایش های عملکردی کبد می باشند را حـــذف کرده است. روش یادگیری آن به گونه ای بوده است که الگوریتم ژنتیـــک، بهترین راه حل ها را کشف کرده است و سپس الگوریتم ایمنی مصنوعی عملیات بهینه سازی و تشخیص را انجام داده است.

قابل ذکر است که مقایسه مدل پیشنهادی با رگرسیون لجستیک(۲۷)، درخت تصمیم گیری J48 و جنگل تصادفی انجام شده است. رگرسیون لجستیک، جزو مدل های خطی بهبود یافته است که می تواند برای بیان رابطه چندین متغیر مستقل(X) با یک متغیر وابسته دو یا چند حالتی(Y) مورد استفاده قرار گیرد. طبق نتایج به دست آمده، درصد صحت در مدل رگرسیون لجستیک ۷۲/۵۰ درصد بوده است. مدل پیشنهادی در مقایسه با رگرسیون لجستیک درصد صحت بیشتری دارد و در حدود ۲۲/۷۳ درصد اختلاف دقت دارد. هم چنین درصد صحت درخت تصمیم گیری J48 و جنگل تصادفی به ترتیب برابر با ۶۸/۷۸ درصد و ۷۱/۵۳ درصد بوده است.

در جدول شماره ۵ مقایسه مدل پیشنهادی با الگوریتم های دیگر بر مبنای تعداد k (تعداد k در الگوریتم k نزدیک ترین همسایه) و تعداد ویژگی نشان داده شده است. اگر تعداد k کمتر باشد آن گاه درصد صحت بیشـــتر خواهد بود. به دلیل این که کشف همسایه های نزدیک با تشابه بیشتر در اولویت می باشند و لذا دقت طبقه بندی با کشف ویژگی های مشابه که دارای فاصله کمتری هستند بالاتر خواهد بود. طبق نتایج به دست آمده می توان ادعا کرد که مدل پیشنهادی در مقایسه با الگوریتم های دیگر، درصد صحت بیشتری دارد. هم چنین الگوریتم بهینه سازی گرگ خاکستری

[DOI: 10.29252/sjimu.28.5.76]

(۲۸) در مقایسه با الگوریتم بهینه سازی اجتماع ذرات (۲۹)، الگوریتم کلونی زنبور مصنوعی(۳۰) و الگوریتم بهینه سازی نهنگ(۳۱) از درصد صحت بیشتری بهره مند است. اگر تعداد k برابر با ۳ باشد آن گاه درصد صحت مدل پیشنهادی برابر با ۹۵/۲۳ درصد می باشد و اگر تعداد k برابر با ۵ باشد آن گاه درصد صحت مدل پیشنهادی برابر با ۹۴/۶۳ درصد می باشد. اگر تعداد k برابر با ۴ باشد آن گاه درصد صحت در مدل پیشنهادی و الگوریتم بهینه سازی گرگ خاکستری به ترتیب برابر با ۹۵/۱۱ درصد و ۹۴/۹۱ درصد می باشد.

جدول شماره ۵. مقایسه مدل پیشنهادی با الگوریتم های دیگر بر مبنای تعداد k

تعداد k	الگوریتم ها	تعداد ویژگی/درصد صحت			
		۵	۷	۸	۱۰
۳	الگوریتم بهینه سازی اجتماع ذرات	۹۷/۹۲	۹۶/۷۵	۹۵/۱۴	۹۴/۷۸
	الگوریتم کلونی زنبور مصنوعی	۹۷/۸۶	۹۶/۳۵	۹۵/۹۳	۹۴/۳۲
	الگوریتم بهینه سازی گرگ خاکستری	۹۸/۲۵	۹۶/۸۲	۹۶/۲۹	۹۵/۰۵
	الگوریتم بهینه سازی نهنگ	۹۷/۹۴	۹۶/۱۷	۹۶/۰۲	۹۴/۹۶
	مدل پیشنهادی	۹۸/۶۳	۹۷/۲۴	۹۶/۸۴	۹۵/۲۳
۴	الگوریتم بهینه سازی اجتماع ذرات	۹۷/۱۲	۹۵/۵۲	۹۵/۱۴	۹۳/۶۲
	الگوریتم کلونی زنبور مصنوعی	۹۷/۲۵	۹۵/۱۶	۹۵/۹۳	۹۳/۵۹
	الگوریتم بهینه سازی گرگ خاکستری	۹۷/۷۱	۹۶/۷۴	۹۶/۲۹	۹۴/۹۱
	الگوریتم بهینه سازی نهنگ	۹۷/۵۶	۹۶/۲۹	۹۶/۰۲	۹۴/۵۲
	مدل پیشنهادی	۹۸/۳۴	۹۷/۰۲	۹۶/۱۵	۹۵/۱۱
۵	الگوریتم بهینه سازی اجتماع ذرات	۹۶/۸۲	۹۴/۵۵	۹۳/۹۱	۹۲/۷۶
	الگوریتم کلونی زنبور مصنوعی	۹۶/۵۷	۹۴/۱۲	۹۳/۳۸	۹۲/۶۸
	الگوریتم بهینه سازی گرگ خاکستری	۹۷/۱۵	۹۵/۲۱	۹۴/۸۲	۹۳/۴۸
	الگوریتم بهینه سازی نهنگ	۹۶/۹۴	۹۴/۶۱	۹۴/۲۲	۹۳/۳۲
	مدل پیشنهادی	۹۸/۰۹	۹۶/۸۵	۹۵/۷۹	۹۴/۶۳

بحث و نتیجه گیری

مقایسه ها حاکی از آن است که در سه مدل طبقه بندی ماشین بردار پشتیبان، شبکه عصبی مصنوعی و شبکه بیزین، بیشترین دقت مربوط به ماشین بردار پشتیبان بود. دقت تشخیص در مدل درخت تصمیم گیری C5.0 نیز از کارایی مناسبی بهره مند بود. ویژگی هایی از قبیل TB, DB, Alkphos, SGPT, TP جزو ویژگی های اثرگذار بر تشخیص بیماری کبد شناخته شدند. الگوریتم های هوشمند می تواند راهنمای پزشکان در تشخیص بیماری کبد باشند و دقت نتایج حاصل از آن ها نیز کاملاً به واقعیت نزدیک است.

مطالعات قابل توجهی در مورد مسئله تشخیص بیماری کبد بر روی مجموعه داده ILPD انجام شده گرفته است؛ این کارهای تحقیقاتی، دید با ارزشی را در خصوص ماهیت این مسئله به ارمغان آورده اند. در مطالعه آبدار و همکاران(۸) که یک پژوهش گسترده در زمینه تشخیص بیماری کبد می باشد، ویژگی های بیماری کبد به صورت جداگانه ارزیابی شده اند و برای انجام این کار از درختان مختلف تصمیم گیری از قبیل C5.0 و CHAID استفاده کرده اند. آن ها از دو روش آنتروپی و نرخ اطلاعات برای این منظور بهره برده اند و در نهایت به درصد صحت ۹۳/۷۵ درصدی برای تشخیص بیماری کبد دست پیدا کرده اند.

جلوداری و همکاران(۹) با بهره گیری از الگوریتم بهینه سازی اجتماع ذرات توانسته اند که مهم ترین ویژگی ها را انتخاب کنند و عملیات طبقه بندی توسط الگوریتم ماشین بردار پشتیبان بر مبنای نمونه های مشابه و آسان برای جداکننده ممکن‌پ ذیر شده است. الگوریتم بهینه سازی اجتماع ذرات از روش های محلی و سراسری برای کشف نقاط بهینه و مهم استفاده کرده و ذرات درجهت بهترین ویژگی ها حرکت کرده اند. نتایج ترکیب بهینه سازی اجتماع ذرات با ماشین بردار پشتیبان برابر با ۹۴/۴۲ درصد گزارش شده است.

[DOI: 10.29252/sjimu.28.5.76]

در این مقاله به منظور تشخیص بیماری کبد، مدل ترکیبی بر مبنای الگوریتم بهینه سازی شیرمورچه و k نزدیک ترین همسایه ارائه گردید. به دلیل این که در مدل پیشنهادی، ویژگی های مهم توسط الگوریتم بهینه سازی شیرمورچه انتخاب شدند، دقت طبقه بندی و تشخیص بسیار بالا بود؛ هم چنین اندازه بردار ویژگی نیز در حالت کاهشی از درصد صحت بالاتری برخوردار بود که این منجر به کاهش فضای ذخیره سازی و افزایش دقت می گردد.

پیشنهاد می شود که به منظور انتخاب ویژگی از روش های دیگر انتخاب ویژگی همانند الگوریتم گرگ خاکستری، الگوریتم جستجوی هارمونی و الگوریتم بهینه سازی امواج آب استفاده شود. هم چنین می توان از روش های دیگر استخراج ویژگی نظیر روش تحلیل تفکیک کننده خطی در ترکیب با روش های مختلف انتخاب ویژگی استفاده کرد و اثر اعمال این روش ها را بر روی داده های استخراج شده از بیماران کبد با الگوریتم های مختلف داده کاوی مقایسه نمود. بعد از انتخاب ویژگی ها برای طبقه بندی نمونه ها می توان از روش رگرسیون لجستیک استفاده کرد.

الگوریتم بهینه سازی شیرمورچه که به اختصار ALO نامیده می شود، بر اساس رفتار شیرمورچه ها در طبیعت، برای اولین بار در سال ۲۰۱۵ توسط میرجلیلی پیشنهاد شده است. این الگوریتم بر روی ۱۹ تابع پیچیده ریاضی آزمایش شده است و با الگوریتم های مختلف مقایسه شده است. کدنویسی الگوریتم بهینه سازی شیرمورچه برای توابع ریاضی بهینه سازی در نرم افزار متلب در لینک (https://www.mathworks.com/matlabcentral /fileexchange/49920-ant-lion-optimizer-alo) و مقاله کامل الگوریتم بهینه سازی شیرمورچه در لینک با شماره DOI (doi.org/10.1016/j.advengsoft.2015.01.010) موجود است. پژوهشگران می توانند برای مطالعه بیشتر در مورد این الگوریتم به لینک های ذکر شده مراجعه کنند و از این الگوریتم در تحقیق های مختلف استفاده نمایند.

References

1.Calvopina DA, Noble C, Weis A, Hartel GF, Ramm GA. Supersonic shear wave elastography and APRI for the detection and staging of liver disease in pediatric cystic fibrosis. J Cyst Fibros2019; 2:1-6. doi.10.1016/j.jcf.2019.06.017

2.Lewindon PJ, Puertolas-Lopez MV, Ramm LE, Noble C, Ramm GA. Accuracy of transient elastography data combined with APRI in detection and staging of liver disease in pediatric patients with cystic fibrosis. Clin Gastroenterol Hep 2019; 17: 2561-9. doi.10.1016/j.cgh.2019.03.015.

3.Vanderlocht J, Cruys MVD, Stals F, Bakker L, Damoiseaux J. Multiplex autoantibody detection for autoimmune liver diseases and autoimmune gastritis. J Immunolo Meth 2017; 448: 21-5. doi.10.1016/j.jim.2017.05.003.

4.Gharehchopogh FS, Mousavi SK. [A decision support system for diagnosis of diabetes and hepatitis. based on the combination of particle swarm optimization and firefly algorithm]. J Health Bio Inform 2019; 6: 32-45. (Persian)

5.Gharehchopogh FS, Shayanfar H, Gholizadeh H. A comprehensive survey on symbiotic organisms search algorithms. Art Int Rev 2019; 1-48. doi.10.1007/s10462-019-09733-4

6.Gharehchopogh FS, Gholizadeh H. A comprehensive survey: whale optimization algorithm and its applications. Swarm Evol Comput2019; 48: 1-24. doi. 10.1016/j.swevo.2019.03.004

7.Gharehchopogh FS, FarokhZad MR. [Determining fuzzy logic parameters by using genetic algorithm for the diagnosis of liver disease]. J Health a Bio Info2018; 5: 384-39. (Persian)

8.Abdar M, Zomorodimoghadam M, Das R, Ting IH. Performance analysis of classification algorithms on early detection of liver disease. Exp Syst Appl2017; 67: 239-51. doi.10.1016/j.eswa.2016.08.065

9.Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer aided decision-making for predicting liver disease using PSO based optimized SVM with feature

[DOI: 10.29252/sjimu.28.5.76]

selection. Info Med Unlocked 2019; 17: 1-17. doi.org/10.1016/j.imu.2019.100255
10.Pourpanah F, Tan CJ, Lim CP, Mohamad J. A Q-learning based multi agent system for data classification. Appl Soft Comput 2017; 52: 519-31. doi. 10.1016/j.asoc.2016.10.016
11.Weng CH, Cheng T, Han RP. Disease prediction with different types of neural network classifiers. Tel Info 2016; 33: 277-92. doi.10.1016/j.tele.2015.08.006
12.Liang C, Peng L. An automated diagnosis system of liver disease using artificial immune and genetic algorithms. J Med Syst 2013; 2:1-10. doi.10.1007/s10916-013-9932-9
13.Kumar P, Thakur RS. Diagnosis of liver disorder using fuzzy adaptive and neighbor weighted k-nn method for lft imbalanced data. Int Con Struc Syst 2019; 3:1-5. doi.10.1109/ICSSS.2019.8882861
14.Rajeswari P, Reena GS. Analysis of liver disorder using data mining algorithm. Global J Comput Sci Technol2010; 2:71-6.
15.Mirjalili S. The ant lion optimizer. Adv Eng Soft2015; 83: 80-98. doi. 10.1016/j.advengsoft.2015.01.010
16.Martin B. Instance Based Learning: Nearest Neighbour with Generalisation. Uni Waikato Dept Comput Sci Newzealand1995; 95:1-76.
17.Mahmoudi M, Gharehchopogh FS. an improvement of shuffled frog leaping algorithm with a decision tree for feature selection in text document classification. CSI J Compu Sci Eng 2018; 16: 60-72.
18.Orooji A, Langarizadeh M. Evaluation of the effect of feature selection and different kernel functions on svm performance for breast cancer diagnosis. J Health Biomed Info2018; 5:244-51. doi.jhbmi.ir/article-1-284-en.html
19.Allahverdipour A, Gharehchopogh FS. An improved K-nearest neighbor with crow search algorithm for feature selection in text documents classification. J Adv Compu Res 2018; 9: 37-48. doi.jacr.iausari.ac.ir/article_655529.html
20.Jain S, Shukla S, Wadhvani R. Dynamic selection of normalization techniques using data complexity measures. Exp Syst Appl 2018; 106: 252-62. doi. 10.1016/j.eswa.2018.04.008
21.Han J, Kamber M. Data mining concepts and techniques. 2 th ed. Morgan Kuafmann Publication. 2006; P.133-9. doi.10.1016/C2009-0-61819-5
22.Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Info Med Unlocked 2018; 10:100-07. doi.10.1016/j.imu.2017.12.006
23.Edla DR, Cheruku R. Diabete finder: a bat optimized classification system for type 2 diabetes. Proce Comput Sci 2017; 115: 235-42. doi.10.1016/j.procs.2017.09.130
24.Abdar M, Yen NY, Hung JCS. Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Bio Eng2018; 38: 953-65. doi.10.1007/s40846-017-0360-z
25.Bashir S, Qamar U, Khan FH. Intelli health a medical decision support application using a novel weighted multi-layer classifier ensemble framework. J Biomed Info 2015; 59: 185-200. doi.10.1016/j.jbi.2015.12.001
26.Zhou J, Lai Z, Gao C, Miao D, Yue X. Rough possibilistic C-means clustering based on multigranulation approximation regions and shadowed sets. Knowl Base Syst 2018; 160: 144-66. doi.10.1016/j.knosys.2018.07.007
27.Singh J, Bagga S, Kaur R. Software based prediction of liver disease with feature selection and classification techniques. Proce Comput Sci2020; 167: 1970-80. doi. 10.1016/j.procs.2020.03.226
28.Mirjalili S, Mirjalili SM, Lewis A. Grey wolf optimizer. Adv Eng Soft 2014; 69: 46-61. doi.10.1016/j.advengsoft.2013.12.007
29.Kennedy J, Eberhart RC. Particle swarm optimization. Proce Int Con Neur Net 1995;3: 1942-8. doi.10.1109/ICNN.1995.488968
30.Karaboga D. An idea based on honeybee swarm for numerical optimization technical report tr06 Erciyes University engineering faculty. Com Eng Dep2005; 4:81-6. doi. 015d/f4d97ed1f541752842c49d12e429a78 5460b.pdf
31.Mirjalili S, Lewis A. The whale optimization algorithm. Adv Eng Soft 2016; 95: 51-67. doi.10.1016/j.advengsoft.2016.01.008

A Hybrid Model based on Ant Lion Optimization Algorithm

[DOI: 10.29252/sjimu.28.5.76]

and K-Nearest Neighbors Algorithm to Diagnose Liver Disease

*Javadzadeh S[1], Shayanfar H[2], Soleimanianqarachapaq F[2]**

(Received: February 1, 2020 Accepted: September 1, 2020)

Abstract

Introduction: Given that a huge amount of cost is imposed on public and private hospitals from the department of liver diseases, it is necessary to provide a method to predict liver diseases. This study aimed to propose a hybrid model based on the Ant Lion Optimization algorithm and K-Nearest Neighbors algorithm to diagnose liver diseases.

Materials & Methods: This descriptive-analytic study proposed a hybrid model based on machine learning algorithms to classify individuals into two categories, including healthy and unhealthy (those with liver diseases). The proposed model has been simulated using MATLAB software. The datasets used in this study were obtained from the Indian Liver Patient Dataset available in the Machine Learning Repository at the University of Irvine, California. This dataset contains 583 independent records, including 10 features for liver diseases.

Findings: After pre-processing, the dataset was randomly divided into 20 categories of the entire dataset, which included different training and test data. In each category of the dataset, 90% and 10% of the data were used for training and test, respectively. Regarding all features, the results obtained the most accurate mode at 95.23%. Moreover, according to the criteria of specificity and sensitivity accuracy, the corresponding values were 93.95% and 94.11%, respectively. Furthermore, the accuracy of the proposed model along with five features was estimated at 98.63%.

Discussions & Conclusions: This model was proposed to diagnose and classify liver diseases along with an accuracy rate of higher than 90%. Healthcare centers and physicians can utilize the results of this study.

Keywords: Ant lion optimization (ALO) algorithm, Classification, Diagnosis of liver disease, K-nearest neighbors (KNN) algorithm

1. Dept of Computer Engineering, Kamal Institute of Higher Education, Urmia, Iran
2. Dept of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
**Corresponding author Email:bonab.farhad@gmail.com*

Scientific Journal of Ilam University of Medical Sciences

[DOI: 10.29252/sjimu.28.5.76]