آنالیز داده های مربوط به بیماران هپاتیت با استفاده از الگوریتم جلبک مصنوعی باینری مبتنی بر K نزدیکترین همسایه

عاطفه بیگلری صالح[۱]، فرهاد سلیمانیان قره چپق[*۱]

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

تاریخ دریافت: ۱۳۹۷/۱۰/۱۰ تاریخ پذیرش: ۱۳۹۸/۹/۳۰

چکیده

مقدمه: از مشکلات اصلی در علم پزشکی، تشخیص و پیش بینی به موقع بیماری ها می باشد. استفاده از سیستم های تصمیم یار به منظور کشف دانش نهفته در مجموعه اطلاعات بیماری و در سوابق مربوط به بیماران یکی از راهکارهایی است که در زمینه تشخیص و پیشگیری از بیماری بسیار موثر می باشد. هدف اصلی از این مقاله، طراحی یک سیستم تصمیم یار پزشکی است که بتواند بیماری هپاتیت را تشخیص دهد.

مواد و روش ها: این مطالعه از نوع توصیفی-تحلیلی می باشد. مجموعه داده آن شامل ۱۵۵ رکورد با ۱۹ ویژگی موجود در پایگاه داده یادگیری ماشین UCI می باشد. در این مقاله، از الگوریتم جلبک مصنوعی باینری برای انتخاب ویژگی و از k نزدیک ترین همسایه برای کلاس بندی هپاتیت به دو کلاس سالم و ناسالم استفاده شده است. از ۸۰ درصد داده ها جهت آموزش و از ۲۰ درصد باقی مانده جهت آزمون استفاده شده است. هم چنین جهت ارزیابی مدل از شاخص های دقت، بازخوانی، F-Measure و صحت استفاده شده است.

یافته های پژوهش: بررسی اولیه نشان داد که درصد صحت مدل پیشنهادی برابر با ۹۶/۴۵ درصد می باشد. بعد از انتخاب ویژگی با الگوریتم جلبک مصنوعی درصد صحت در بهترین حالت به ۹۸/۳۶ درصد رسید. در مدل پیشنهادی در حالت ۳۰۰ بار تکرار، مقدار معیارهای دقت، بازخـــوانی، F-Measure، و نرخ خطا به ترتیب برابر با ۹۶/۲۳ درصد، ۹۶/۷۴ درصد، ۹۶/۴۸ درصد، ۳/۵۵ درصد می باشند.

بحث و نتیجه گیری: هپاتیت یکی از شایع ترین بیماری ها در بین زنان و مردان می باشد. تشخیص به موقع بیماری ضمن کاهش هزینه ها، شانس درمان موفقیت آمیز بیمار را افزایش می دهد. در این مطالعه ضمن تشخیص بیماری به کمک روش ترکیبی، توانستیم با استفاده از انتخاب ویژگی به دقت بالایی در تشخیص بیماری دست یابیم.

واژه های کلیدی: سیستم تصمیم یار پزشکی، تشخیص بیماری هپاتیت، الگوریتم جلبک مصنوعی باینری، k نزدیک ترین همسایه، انتخاب ویژگی

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

Email:bonab.farhad@gmail.com

Copyright © 2019 Journal of Ilam University of Medical Science. This is an open-access article distributed under the terms of the Creative Commons Attribution international 4.0 International License (https://creativecommons.org/licenses/by-nc/4.0/) which permits copy and redistribute the material, in any medium or format, provided the original work is properly cited.

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

مقدمه

در پزشکی مدرن، اغلب تشخیص نوع بیماری پیچیده می باشد و دلایل منـــطقی و اصولی برای تصمیم گیری های دقیق وجود ندارد. به همین جهت تصمیم گیری های پزشکان معمولاً دلخـــواه و متغیر می باشد. از طرفی حجم داده های پزشکی به قدری زیاد شده است که پردازش اطلاعات برای پزشکان به منظور درمان در وضعیت های اضطراری مشکل است(۱). یکی از مشکلات دیگر، خطاهای تشخیصی توسط پزشکان است. سیستم های تصمیم یار پزشکی می توانند در این زمینه کمک کننده و یاری رسان به پزشکان باشند. با توجه به مسائل فوق الذکر به آسانی قابل درک است که پزشکان با سیل عظیمی از داده ها در پروسه تشخیص پزشکی مواجه هستند که نیاز به پیدایش سیستم های تصمیم یار پزشکی را آشکار خواهد نمود.

سیستم تصمیم یار پزشکی یک نرم افزار کامپیوتری تصمیم یار تعاملی است که به منظور یاری رساندن به پزشکان و متخصصین دیگر در حوزه سلامت برای وظایف تصمیم گیری، از جمله تشخیص بیماری از داده های بیمار، طراحی می شود(۲). این سیستم ها با استفاده از اطـــلاعات و دانش پزشکی به تشخیص عارضه های گوناگون و تجویز توصیه های پزشکی برای بیماران اقدام می نمایند. سیستم های تصمیم یار برای جایگزینی پزشکان طراحی نشده اند و تنهـــا جهت یاری رسانی به متخصصان علوم پزشکی در امر تشخیص بیمـــاری ها بر اساس یکسری از قواعد تجربی ارائه شده اند(۳). برخی دلایل استفاده از سیســـتم های تصمیم یار در پزشکی به شرح زیر است:

-کیفیت مراقبت پزشکی را افزایش می دهند و میزان سلامتی را ارتقا می بخشند.

-از خطا و رخدادهای مغایر که ممکن است در اثر بی دقتی پزشکان یا کمبود دانش یا موارد دیگر به وجود آید، جلوگیری کرده و به انتخاب تصمیم درست کمک می کند.

-بازدهی را افزایش و هزینه ها را کاهش می دهند و سطح رضایت بیماران را ارتقاء می دهند.

حوزه پزشکی و سلامت از بخش های مهم در جوامع است. تشخیص بیماری از میان حجـــم انبـــوه داده های مرتبط با سوابق بیماری و پرونده های پزشکی افراد با استفاده از سیستم های تصمیم یار می تواند منجر به شناسایی قوانین تشخیص، تسکین بیماری ها شود(۴). در حوزه پزشکی، جـــمع آوری داده هـــــا در مورد بیماری های مختلف از اهمیت زیادی برخوردار است. حجم داده های جمع آوری شده بسیار بالا است و برای این که بتوان از بین این حجم انبود داده ها الگوها و نتایج مورد نظر را به دست آورد، باید از سیستم های تصمیم یار استفاده شود.

بیماری هپاتیت از شایع ترین بیماری ها در میان بیماری های مزمن می باشد. استفاده از تکنیک های یادگیری ماشین برای ایجاد مدل های طبقه بند، جهت تشخیص افراد در معرض خطر برای کاهش عوارض ناشی از بیماری بسیار کمک کننده می باشند(۵). بیماری هپاتیت بیش از هر بیماری دیگر در کشورهای توسعه یافته موجب مرگ و ناتوانی شده و هزینه های اقتصادی تحمیل می نماید(۶). با توجه به افزایش سریع بیماری هپاتیت در سراسر جهان احتمالاً این بیماری تا سال ۲۰۲۰ به شایع ترین علت مرگ تبدیل خواهد شد.

در سال های اخیر استفاده از روش های هوش مصنوعی به منظور تشخیص بیماری ها، مورد توجه بسیاری از محققین قرار گرفته است. در بین روش های مختلف، الگوریتم های فرا ابتکاری به دلیل ماهیت ساختاری خود از محبوبیت ویژه ای برخوردارند(۹). در مدلی ترکیبی از داده کاوی برای انتخاب ویژگی های بیماری هپاتیت، بر مبنای الگوریتم جستجوی فاخته باینری و حداقل مربعات ماشین پشتیبان پیشنهاد شده است. انتخاب ویژگی ها با الگوریتم جستجوی فاخته باینری انجام شده و ماشین بردار پشتیبان حداقل مربعات برای پیش بینی استفاده شده است. درصد صحت حاصل از نتایج پیاده سازی مدل ترکیبی، با استفاده از مجموعه داده های هپاتیت از مجموعه داده UCI؛ برابر با ۹۹/۱۸ درصد می باشد. ابراهیمی و همکاران(۱۰) از درخت تصمیم گیری، شبکه عصبی مصنوعی و ماشین بردار پشتیبان برای تشخیص بیماری هپاتیت استفاده کرده اند. معیار ارزیابی روش های طبقه بندی نرخ دقت هر روش بوده و برای تست هر روش از نرم افزار clementine و پایگاه داده هپاتیت واقع در مخزن داده دانشگاه کالیفرنیا

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

استفاده شده است. نتایج به دست آمده نشان داده است که شبکه عصبی مصنوعی دقت بالاتری نسبت به سایر الگوریتم ها دارد. دقت تشخیص شبکه عصبی مصنوعی برابر با ۸۹/۷۴ درصد است.

خدارحمی و روحانی مقاله ای تحت عنوان «تشخیص بیماری هپاتیت با استفاده از ترکیب الگوریتم k-means و الگوریتم بهینه سازی علف های هرز» پیشنهاد داده اند(۱۱). نتایج استفاده از خوشه بندی K-Means و الگوریتم بهینه سازی علف های هرز حاکی از دقت ۱۰۰ درصد در خوشه بندی داده های آزمایشی و ۹۸/۹۰ درصد در خوشه بندی داده های آموزشی است. بعد از نرمال سازی به کمک الگوریتم تحلیل تفکیکی فیشر تعداد ۶ ویژگی از بین ۱۹ ویژگی انتخاب شده و برای خوشه بندی وارد الگوریتم k-means شده اند و مراحل کشف مراکز خوشه ها با استفاده از الگوریتم بهینه سازی علف های هرز، انجام شده است. محققین برای تشخیص بیماری هپاتیت از ترکیب مجموعه راف و یادگیری ماشین، استفاده کرده اند(۱۲). مدل ترکیبی شامل دو مرحله است. در مرحله اول، ویژگی های زائد از طریق مجموعه راف حذف می شوند و مرحله دوم، فرآیند دسته بندی با ویژگی های باقی مانده به وسیله یادگیری ماشین انجام می شود. هدف این مقاله، انتخاب ویژگی های مناسب، کاهش ابعاد و پیچیدگی مسئله و بالا بردن دقت دسته بندی است. نتایج تجربی بر روی ۱۵۵ نمونه افراد نشان می دهد که روش ترکیبی نسبت به دسته بندی بدون کاهش ویژگی، دقت دسته بندی بهتری می دهد. مدل ترکیبی با کاهش ۷۸/۹۵ درصد ویژگی ها، بالاترین دقت دسته بندی یعنی مقدار ۱۰۰ درصد را به دست آورده است.

یک سیستم هوشمند مبتنی بر تکنیک های یادگیری ماشین جهت تشخیص بیماری هپاتیت پیشنهاد شده است(۱۳). الگوریتم پیشنهادی شامل سه مرحله اساسی می باشد: کاهش ابعاد، طبقه‌ب ندی و هم جوشی طبقه بندی کننده ها با یکدیگر. مجموعه داده ها از پایگاه داده UCI گرفته شده است. در ابتدا تمام داده ها نرمال شده اند. سپس با استفاده از آنالیز اجزای اساسی تعداد ویژگی ها به ۱۰ ویژگی کاهش پیدا کرده است. در مرحله بعد از سه طبقه بندی کننده جهت مدل سازی داده ها استفاده شده است. الگوریتم پیشنهادی توانسته به درصد صحت ۹۶/۳۲ دست یابد.

مدلی بر مبنای سیستم استنتاج فازی برای تشخیص بیماری هپاتیت پیشنهاد شده است(۱۴). برای انتخاب ویژگی از درخت تصمیم گیری و برای طبقه بندی ویژگی ها از سیستم استنتاج فازی استفاده شده است. نتایج بر روی مجموعه داده هپاتیت نشان می دهد که دقت تشخیص مدل برابر با ۹۶/۰۳ درصد است(۱۵). در مدل ترکیبی بر مبنای الگوریتم کلونی زنبور مصنوعی و ماشین بردار پشتیبان برای تشخیص بیماری هپاتیت پیشنهاد شده است. در مدل ترکیبی از الگوریتم کلونی زنبور مصنوعی برای انتخاب ویژگی و از ماشین بردار پشتیبان برای طبقه بندی استفاده شده است. نتایج آزمایش ها نشان داده است که دقت تشخیص برابر با ۷۹/۲۹ درصد است.

هدف اصلی این مطالعه، ارائه یک سیستم تصمیم یار پزشکی مبتنی بر الگوریتم جلبک مصنوعی باینری(۷) و k نزدیک ترین همسایه(۸) برای تشخیص بیماری است. در مدل پیشنهادی از الگوریتم جلبک مصنوعی باینری برای انتخاب ویژگی و از k نزدیک ترین همسایه برای طبقه بندی نمونه ها استفاده شده است. الگوریتم جلبک مصنوعی باینری 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 شده است.

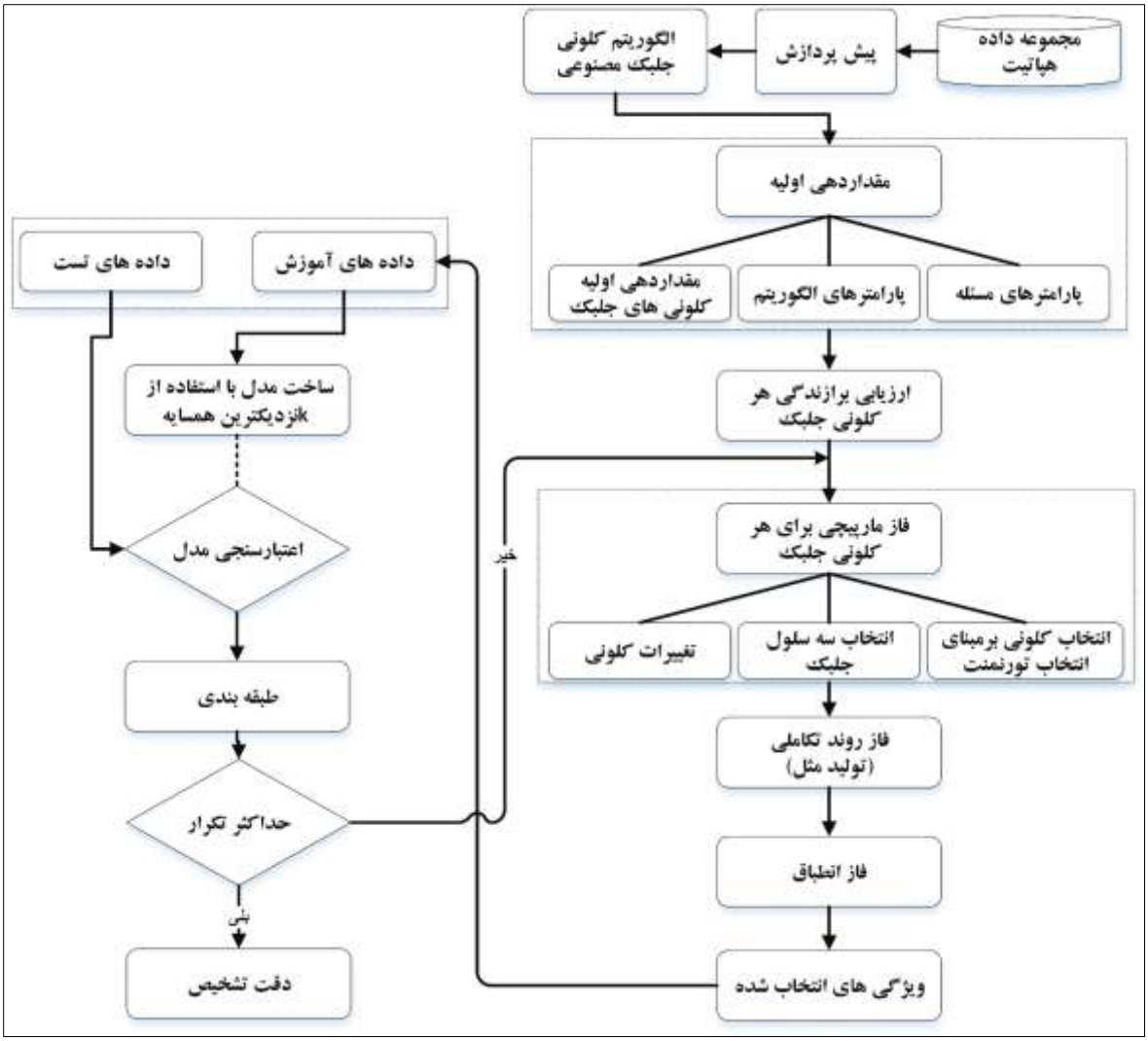
موارد و روش ها

این مطالعه از نوع توصیفی-تحلیلی است که بر اساس ویژگی های ورودی به تشخیص وضعیت بیماران هپاتیت از نظر سالم یا ناسالم بودن می پردازد. داده های مورد استفاده در این مقاله از مجموعه داده مربوط به بیماران مبتلا به هپاتیت، موجود در مجموعه داده یادگیری ماشین دانشگاه ایروین، کالیفرنیا تامین شده است(۱۶). مجموعه داده ها شامل ۱۵۵ رکورد هپاتیت با ۱۹ ویژگی(و یک ویژگی متعلق به کلاس) می باشند. مجموعه داده هپاتیت را می توان یک ماتریس ۱۵۵×۱۹

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

تلقی کرد. از مجموعه داده هپاتیت، ۲۰ درصد یعنی ۳۲ نفر ناسالم و ۸۰ درصد یعنی ۱۲۳ نفر سالم می باشند. در شکل شماره ۱، فلوچارت مدل پیشنهادی نشان داده شده است.



شکل شماره ۱. فلوچارت مدل پیشنهادی

بخش اول: پیش پردازش

اگر مقادیر ویژگی های مجموعه داده در دامنه متفاوتی قرار داشته باشند، احتمال بروز خطا در یافته ها افزایش می یابد. به قرار دادن داده های یک جامعه آماری در دامنه مشابه، نرمال سازی گفته می شود. در مدل پیشنهادی نحوه نرمال سازی به روش Max/Min و در بازه [۰،۱] اســـت(۱۷،۱۸). نرمال سازی مقدار ویژگی ها طبق معادله (۱) انجام شده است. در معادله (۱)، پارامتر v مقدار واقعی ویژگی ها است و min(v) و max(v) مقدار حداقل و حداکثر ویژگی می باشند. و x شامل مقـــدار عـــددی نرمال ســـازی ویژگی ها است.

$$x_i = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)} \qquad (۱)$$

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

در یک مجموعه داده، احتمال مقادیر گمشده برای رکوردها وجود دارد. داده های موجود در یک مجموعه داده، زمانی که وارد الگوریتم می شوند باید کامل و بدون مقادیر گمشده یا داده های ناقص باشند. هم چنین مواردی که مقادیر احتمالاً غلط به ویژگی های یک رکورد تخصیص یافته باشد، باید تصحیح و در صورت عدم تصحیح از مجموعه داده حذف گردد. متاسفانه در مجموعه داده هپاتیت، مقادیر گمشده وجود دارند. در این مقاله برای مقادیر گمشده از حداکثر مقدار ممکن استفاده شده است. در روش حداکثر مقدار ممکن در بین مقادیر قابل پذیرش برای یک ویژگی خاص، حداکثر مقدار آن برای جانشینی انتخاب می شود(۱۹).

بخش دوم: انتخاب ویژگی

بعد از مرحله خواندن مجموعه داده های هپاتیت و عملیات پیش پردازش برروی داده‌ها، عملیات تشکیل کلونی های جلبک و سلول ها انجام می گیرد. در الگوریتم کلونی جلبک مصنوعی، N تعداد کلونی های جلبک و D تعداد متغیرهای تصمیم یا ابعاد مسئله بهینه سازی می باشد. لذا کلونی های جلبک توسط یک ماتریس N*D شبیه سازی می شوند. هر سطر مربوط به یک راه حل ممکن از مسئله بهینه سازی است. در مدل پیشنهادی N تعداد رکوردهای مجموعه داده و D تعداد ویژگی ها می باشد. جمعیت کلونی های جلبک که شامل تعداد زیادی سلول است طبق معادله (۲) تعریف می شود. در مدل پیشنهادی، روش کار برای مجموعه داده هپاتیت بدین صورت است که الگوریتم جلبک مصنوعی باینری از ۵۰ کلونی جلبک و هر کلونی جلبک نیز از ۱۹ ویژگی(۱۹ بعُد) تشکیل شده است. هر کلونی جلبک توسط این ۱۹ ویژگی تعریف می شود.

$$\boldsymbol{Population\ of\ algae\ colony} = \begin{bmatrix} \boldsymbol{x_{11}} & \boldsymbol{x_{12}} & \cdots & \boldsymbol{x_{1d}} \\ \boldsymbol{x_{21}} & \boldsymbol{x_{22}} & \cdots & \boldsymbol{x_{2d}} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{x_{n1}} & \boldsymbol{x_{n2}} & \cdots & \boldsymbol{x_{nd}} \end{bmatrix} \qquad (۲)$$

در مجموعه $x_i = (x_{i1}, x_{i2}, \dots, x_{id}), i = 1,2, \dots, n$ هر xi نشان دهنده یک راه حل ممکن در فضای راه حل ها است. هر کلونی جلبک شامل یک گروه از سلول های جلبک است که به عنوان عناصر یک راه حل در نظر گرفته می شوند. تمام سلول های جلبک در یک کلونی جلبک به عنوان یک واحد کلی در نظر گرفته می شوند که به سمت یک مکان مناسب با منابع فراوان حرکت می کنند. اگر کلونی به یک موقعیت ایده آل برسد، راه حل بهینه ای به دست می آید. ارزیابی هر کلونی جلبک بر مبنای تابع هدف طبق معادله (۳) محاسبه می شود.

$$\mathrm{fit_i} = 1 - \frac{\mathrm{Obj_i} - \mathrm{worst(Obj)}}{\mathrm{best(Obj)} - \mathrm{worst(Obj)}} \qquad (۳)$$

در معادله (۳)، fit_i برازندگی کلونی جلبک i ام است. پارامتر Obj_i مقدار تابع هدف برای کلونی جلبک i ام است. هر کلونی جلبک بر مبنای معیار فاصله محاسبه می شود. پارامترهای worst و best بدترین و بهترین مقدار کلونی های جلبک هستند. در الگوریتم جلبک مصنوعی، سه بخش کلیدی وجود دارد: فاز مارپیچ، روند تکاملی و انطباق. کلونی های جلبک تلاش می کنند تا از طریق حرکت، تکامل و انطباق به موقعیت مطلوب حرکت کنند. فاز مارپیچی توسط انتخاب شدن سه سلول تصادفی متمایز و تغییر موقعیت آن ها شبیه سازی می شود و هدف این است که عملیات جهش بر روی کلونی ها انجام گیرد. در روند تکاملی برای به دست آوردن راه حل های بهینه، یک سلول جلبک از بزرگ ترین کلونی جلبک جایگزین یک سلول جلبک از کوچک ترین کلونی جلبک می شود. انطباق فرآیندی است که در آن کلونی جلبک ضعیف تلاش می کند به

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

سمت بزرگ ترین کلونی حرکت کند و خود را با محیط سازگار کند.

در مدل پیشنهادی باید الگوریتم جلبک مصنوعی از حالت پیوسته به گسسته تبدیل شود. به دلیل این که مقدار مجموعه داده ها گسسته هستند و در بازه ۰ و ۱ قرار دارند. لذا هر سلول جلبک xij طبقه معادله (۵)، تعریف می شود. جمعیت کلونی های جلبک با ۱۹ ویژگی کدگذاری می شود. برای تبدیل اعداد به باینری از تابع سیگموئید طبق معادله (۴)، استفاده می شود(۷). خروجی تابع سیگموئید در یک محدوده عددی خاصی(عموماً بین صفر و یک) قرار می گیرد. در این تابع جواب ۰ یا ۱ نخواهد بود بلکه مجموعه اعدادی بین صفر و یک است.

$$\boldsymbol{g(x)=\frac{1}{1+e^{-x}}} \qquad (۴)$$

$$x_{ij}=\begin{cases}0, & if\ g(x)<0.5\\ 1, & otherwise\end{cases} \qquad (۵)$$

در مدل پیشنهادی با استفاده از الگوریتم کلونی جلبک مصنوعی زیرمجموعه ای از ویژگی ها که به بهینه ترین مقدار منجر می شوند، انتخاب می گردد. تابع برازندگی برای انتخاب ویژگی از هر کلونی جلبک طبق 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.

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

مجموعه داده هایی که دارای ابعاد زیادی هستند علی رغم فرصت هایی که به وجود می آورند، چالش های محاسباتی زیادی را ایجاد می کنند. یکی از مشکلات داده های با ابعاد زیاد این است که در بیشتر مواقع تمام ویژگی های داده ها برای یافتن دانشی که در داده ها نهفته است مهم و حیاتی نیستند. ایده اصلی در انتخاب ویژگی، حذف زیرمجموعه ای از ویژگی های ورودی است که اطلاعات کمی دارند. از این رو انتخاب ویژگی برای کاهش فضای ویژگی و افزایش کارایی طبقه بندی به کار می رود.انتخاب ویژگی علاوه بر بالابردن دقت و کارایی طبقه بندی، قابلیت درک نتایج حاصل از آن را بالا می برد.

بخش سوم: طبقه‌بندی

برای طبقه بندی، ابتدا لازم است مجموعه داده ها به دو بخش آموزش(۸۰ درصد نمونه ها) و تست(۲۰ درصد نمونه ها) تقسیم شوند. داده های بخش آموزش مدل ارزیابی را تولید می کنند و داده های بخش تست با کمک تعدادی رکورد، مدل تولید شده را تست و برچسب مربوط به رکوردهای مذکور را تعیین و کلاس آن ها را مشخص می نمایند. در الگوریتم k نزدیک ترین همسایه فرض بر این است که نمونه ها به نزدیک ترین کلاس ها اختصاص داده شوند. معمولی ترین معیار برای تشخیص فاصله، استفاده از فاصله اقلیدسی است(۸). فاصله اقلیدسی بین دو نمونه از داده های $x=<x_1,x_2,\dots x_n>$ و $y=<y1,y2,\dots,yn>$ طبق معادله(۷) تعریف می‌شود.

$$d(x,y)=\sqrt{(x_1-y_1)+(x_2-y_2)^2+\dots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2} \qquad (۷)$$

در الگوریتم k نزدیک ترین همسایه، هر نمونه بر مبنای فاصله تشابه به یک دسته تعلق می گیرد. در این روش داده های مشابه در نزدیکی یکدیگر قرار می گیرند، از این رو فاصله بین داده ها بر اساس نزدیکی آن ها

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

اندازه گیری می گردد. بر این اساس داده هایی که در کنار یکدیگر قرار می گیرند، همسایه نامیده شده و هر داده جدیدی که به الگوریتم داده شود، فاصله آن با دیگر داده ها محاسبه و در دسته ای قرار می گیرد که در نزدیک ترین فاصله قرار دارد.

بخش چهارم: معیارهای ارزیابی

در این مقاله، جهت ارزیابی کارایی طبقه بندی از پنج معیار دقت، بازخوانی، اندازه گیری، صحت و نرخ خطا استفاده شده است(۲۰،۲۱). دقـــت نشان دهنده تعداد نمونه های درست به تعداد کل نمونه ها است. معیار بازخوانی نشان می دهد که اگر نمونه ای، نوع A تشخیص داده شد با چه احتمالی نوع A می باشد. درصد صحت یک روش طبقه بندی بر روی مجموعه داده های آموزشی، درصد مشاهداتی از مجموعه آموزش است که به درستی توسط روش مورد استفاده طبقه بندی شده است(جدول شماره ۱).

جدول شماره ۱. پارامترهای مورد نیاز جهت بررسی داده

شماره رفرنس	توضیحات	معادلات
(۸)	دقت	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
(۹)	بازخوانی	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
(۱۰)	F-Measure	$\text{F} - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} * \text{Recall}}$
(۱۱)	صحت	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$
(۱۲)	نرخ خطا	$\text{Error Rate} = 1 - \text{Accuracy}$

پارامترهای درست مثبت(TP)، درست منفی(TN)، کاذب مثبت(FP)، کاذب منفی(FN) از پارمترهای اصلی برای معیارهای ارزیابی هستند. پارامتر درست مثبت، بیانگر تعداد رکوردهایی است که دسته واقعی آن ها مثبت بوده و الگوریتم دسته بندی نیز دسته آن ها را به درستی مثبت تشخیص داده است. پارامتر درست منفی، بیانگر تعداد رکوردهایی است که دسته واقعی آن ها منفی بوده و الگوریتم دسته بندی نیز دسته آن ها را به درستی منفی تشخیص داده است. پارامتر کاذب مثبت، بیانگر تعداد رکوردهایی است که دسته واقعی آن ها منفی بوده و الگوریتم دسته بندی دسته آن ها را به اشتباه مثبت تشخیص داده است. پارامتر کاذب منفی، بیانگر تعداد رکوردهایی است که دسته واقعی آن ها مثبت بوده و الگوریتم دسته بندی دسته آن ها را به اشتباه منفی تشخیص داده است.

یافته های پژوهش

در این مقاله به منظور انجام مقایسه ای عادلانه برای تشخیص بیماری هپاتیت، از مجموعه داده های یکسان در آموزش الگوریتم k نزدیک ترین همسایه استفاده شده است. پیاده سازی در محیط ویژوال سی شارپ ۲۰۱۷ انجام شده است. تعداد تکرارها و تعداد k مهم ترین فاکتورها در ارزیابی مدل پیشنهادی می باشند. برای ارزیابی مدل پیشنهادی از ۸۰ درصد داده ها برای آموزش و از ۲۰ درصد داده ها برای تست استفاده گردیده است. برای این منظور از مجموعه داده هپاتیت که از مخزن UCI که مرجعی برای یادگیری ماشین می باشد، انتخاب گردیده است. مجموعه داده هپاتیت شامل ۱۹ ویژگی و ۱ کلاس می باشد. مجموعه کلاس ها به دو گروه مبتلا به بیماری سالم و ناسالم تقسیم می شود که کلاس اول شامل ۱۲۳ نفر افراد سالم و کلاس دوم شامل ۳۲ نفر بیمار می باشند. جدول شماره ۲، ۲۰ ویژگی مجموعه داده هپاتیت را نشان می دهد.

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

جدول شماره ۲. ویژگی های مجموعه داده هپاتیت

ردیف	ویژگی‌ها		دامنه
۱	AGE	سن بیمار	۱۰ تا ۸۰
۲	SEX	جنسیت	زن(۱)؛ مرد(۲)
۳	STEROID	استروئید	خیر(۱)؛ بله(۲)
۴	ANTIVIRALS	ضد ویروس	خیر(۱)، بله(۲)
۵	FATIGUE	خستگی	خیر(۱)، بله(۲)
۶	MALAISE	احساس مرض	خیر(۱)، بله(۲)
۷	ANOREXIA	بی اشتهایی	خیر(۱)، بله(۲)
۸	LIVER BIG	بزرگی کبد	خیر(۱)، بله(۲)
۹	LIVER FIRM	سفتی کبد	خیر(۱)، بله(۲)
۱۰	SPLEEN PALPABLE	طحال کبد	خیر(۱)، بله(۲)
۱۱	SPIDERS	عنکبوت	خیر(۱)، بله(۲)
۱۲	ASCITES	آسیت	خیر(۱)، بله(۲)
۱۳	VARICES	واریس	خیر(۱)، بله(۲)
۱۴	BILIRUBIN	بیلی روبین	۰/۳۹,۰/۸۰,۱/۲۰,۲/۰۰,۳/۰۰,۴/۰۰
۱۵	ALK PHOSPHATE	آلک فسفات	۳۳,۸۰,۱۲۰,۱۶۰,۲۰۰,۲۵۰
۱۶	SGOT	سگات	۱۳,۱۰۰,۲۰۰,۳۰۰,۴۰۰,۵۰۰
۱۷	ALBUMIN	آلبومین	۲/۱,۳/۰,۳/۸,۴/۵,۵/۰,۶/۰
۱۸	PROTIME	پروترومبین تایم	۱۰ تا ۹۰
۱۹	HISTOLOGY	هیستولوژی	خیر(۱)، بله(۲)
۲۰	Class	کلاس	زنده (۲)؛ مرده (۱)

متغیر هدف(کلاس) در این مطالعه، وجود یا عدم وجود بیماری هپاتیت است که در مورد هر کدام از افراد مورد بررسی یکی از این دو حالت ثبت گردیده است. مقدار عددی این ویژگی برابر با ۱ نشان دهنده وجود بیماری و مقدار عددی ۲ نشان دهنده عدم وجود بیماری می‌باشد.

در جدول شماره ۳، نتایج مدل پیشنهادی بر مبنای تعداد تکرار و تعداد k نشان داده شده است. تعداد تکرارها در مدل پیشنهادی برابر با ۱۰۰، ۲۰۰ و ۳۰۰ می باشند. هم چنین تکرار k در مدل پیشنهادی در محدوده ۳ تا ۵ می باشد. بیشترین درصد صحت با ۱۰۰ بار تکرار برابر با ۹۴/۸۹ درصد است. هم چنین بیشترین درصد صحت با ۲۰۰ و ۳۰۰ بار تکرار برابر با ۹۵/۹۳ درصد و ۹۶/۴۵ درصد است. لذا بیشترین درصد صحت در مدل پیشنهادی برابر با ۹۶/۴۵ درصد است. هم چنین اگر تعداد k کمتر باشد نرخ خطا هم کمتر است.

جدول شماره ۳. نتایج مدل پیشنهادی بر مبنای تعداد تکرار و تعداد k

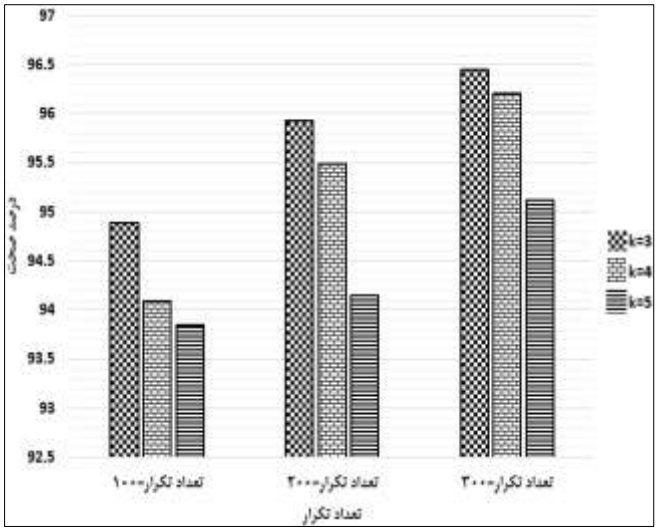
تعداد تکرار	تعداد k	دقت	بازخوانی	F-Measure	صحت	نرخ خطا
۱۰۰	۳	۹۳/۶۵	۹۴/۷۲	۹۴/۱۸	۹۴/۸۹	۵/۱۱
	۴	۹۳/۱۴	۹۴/۱۱	۹۳/۶۲	۹۴/۰۹	۵/۹۱
	۵	۹۲/۹۸	۹۳/۰۷	۹۳/۰۲	۹۳/۸۵	۶/۱۵
۲۰۰	۳	۹۴/۳۵	۹۴/۶۱	۹۴/۴۸	۹۵/۹۳	۴/۰۷
	۴	۹۴/۰۲	۹۴/۱۵	۹۴/۰۸	۹۵/۴۹	۴/۵۱
	۵	۹۳/۹۶	۹۴/۰۳	۹۳/۹۹	۹۴/۱۵	۵/۸۵
۳۰۰	۳	۹۶/۲۳	۹۶/۷۴	۹۶/۴۸	۹۶/۴۵	۳/۵۵
	۴	۹۵/۸۹	۹۶/۱۶	۹۶/۰۲	۹۶/۲۱	۳/۷۹
	۵	۹۵/۳۲	۹۵/۴۶	۹۵/۳۹	۹۵/۱۲	۴/۸۸

در شکل شماره ۲، نمودار مقایسه درصد صحت مدل پیشنهادی بر مبنای تعداد تکرار برای مجموعه داده هپاتیت نشان داده شده است. شکل شماره ۲ نشان می دهد که تعداد تکرار با درصد صحت رابطه مستقیم دارد. یعنی اگر تعداد تکرار بیشتر باشد درصد صحت بیشتر خواهد بود. در شکل شماره ۲، محور افقی نمودار،

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

بیانگر تعداد تکرار و محور عمودی بیانگر درصد صحت است. نمودار برای تکرارهای ۱۰۰، ۲۰۰ و ۳۰۰ بار ترسیم شده است. بیشترین درصد صحت در حالت ۳۰۰ بار تکرار است.



شکل شماره ۲. نمودار درصد صحت مدل پیشنهادی برمبنای تعداد تکرار و k

در جدول شماره ۴، نتایج مدل پیشنهادی بر مبنای انتخاب ویژگی نشان داده شده است. نتایج جدول شماره ۴ بر مبنای ۳۰۰ بار تکرار حاصل شده است. اگر تعداد ویژگی ها کمتر باشند درصد صحت مدل پیشنهادی بیشتر است. برای مثال اگر تعداد ویژگی ها برابر با ۷ باشند آن گاه درصد صحت برابر با ۹۸/۳۶ درصد است و اگر تعداد ویژگی ها برابر با ۱۹ باشند آن گاه درصد صحت برابر با ۹۶/۴۵ درصد است.

جدول شماره ۴. نتایج مدل پیشنهادی بر مبنای انتخاب ویژگی

ردیف	تعداد ویژگی	دقت	بازخوانی	F-Measure	صحت	نرخ خطا
۱	۶	۹۸/۸۲	۹۸/۹۲	۹۸/۸۷	۹۸/۳۶	۱/۶۴
۲	۸	۹۸/۵۶	۹۸/۷۰	۹۸/۶۳	۹۸/۲۸	۱/۷۲
۳	۹	۹۸/۴۱	۹۸/۶۸	۹۸/۵۴	۹۸/۱۳	۱/۸۷
۴	۱۰	۹۸/۲۲	۹۸/۳۵	۹۸/۲۸	۹۸/۰۷	۱/۹۳
۵	۱۱	۹۷/۶۳	۹۷/۹۱	۹۷/۷۷	۹۷/۸۹	۲/۱۱
۶	۱۲	۹۶/۴۷	۹۷/۰۶	۹۶/۷۶	۹۷/۶۵	۲/۳۵
۷	۱۴	۹۶/۲۵	۹۷/۱۴	۹۶/۶۹	۹۷/۳۱	۲/۶۹
۸	۱۵	۹۶/۷۹	۹۶/۸۵	۹۶/۸۲	۹۶/۹۲	۳/۰۸
۹	۱۶	۹۶/۶۱	۹۶/۹۳	۹۶/۷۷	۹۶/۸۵	۳/۱۵
۱۰	۱۷	۹۵/۵۲	۹۶/۰۷	۹۵/۷۹	۹۶/۵۳	۳/۴۷
۱۱	۱۹	۹۵/۴۲	۹۵/۷۹	۹۵/۶۰	۹۶/۴۵	۳/۵۵

در این مقاله، تعداد همـــسایه ها در الگوریتم k نزدیک ترین همسایه در بهینه ترین حالت برابر ۳ در نظر گرفته شد تا بیشترین درصد صحت حاصل شود. هم چنین تعداد ویژگی ها در کاهش نرخ خطا بسیار موثر هستند اگر تعداد ویژگی ها کمتر باشد، نرخ خطا هم کمتر می شود.

بحث و نتیجه گیری

با توجه به توسعه سیستم های تصمیم یار در حوزه پزشکی، در این مطالـــعه، هدف ارائه یک سیستم تصمیم یار برای تشخیص بیماری هپاتیت و یاری رساندن به پزشک برای تشخیص بیماری هپاتیت است. بدین منظور از الگوریتم جلبک مصنوعی باینری برای انتخاب ویژگی و از k نزدیک ترین همســـایه برای دسته بندی افراد بیمار و سالم استفاده شده است. درصد

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

صحت مدل پیشنهادی با ۳۰۰ بار تکرار برابر با ۹۶/۴۵ درصد و درصد صحت با ۶ ویژگی برابر با ۹۸/۳۶ درصد بوده است. لذا انتخاب ویژگی در افزایش دقت و کاهش نرخ خطا موثر است.

پژوهش های گذشته در حوزه داده کاوی بیماری هپاتیت، مربـــوط به پیش بینی و طبقه بندی بیماران می باشند. گزارش های منتشر شده از این پژوهش ها در رابطه با دقت و صحت الگوریتم های پیـــش بینی می باشند. در حوزه تشخیص بیماری هپاتیت، چند نمونه از کارهایی که نزدیک به کار ما هستند را بررسی کردیم(۲۲). از شبکه عصبی مصنوعی احتمالاتی برای تشخیص بیماری هپاتیت استفاده شده است. نتایج نشان داده که دقت تشخیص مدلشان برابر با ۹۱/۲۵ درصد است. مدل پیشنهادی در مقایسه با شبکه عصبی مصنوعی احتمالاتی درصد صحت را به ۹۶/۴۵ درصد رسانده است. دوگانتکین و همکاران(۲۳) از سیستم استنتاج فـــازی برای تشخیص بیماری هپاتیت بهره گرفته اند. دقت تشخیص سیستم استنتاج فازی برابر با ۹۴/۱۶ درصد است. سلیمی و همکاران(۲۴) با استفاده از ترکیب ماشین بردار پشتیبان و الگوریتم شبیه سازی آنالینگ به تشخیص هپاتیت پرداخته اند و این مدل دارای دقت ۹۶/۲۰ درصد است که به عنوان مدل مناسبی جهت تـــشخیص مـــعرفی شده است. از ماشین بردار پشتیبان برای طبقه بندی و از الگوریتم آنالینگ برای بهینه سازی پارامترهای ماشین بردار پشتبان استفاده شده است.

محققین از تکنیک های داده کاوی برای تشخیص بیماری هپاتیت استفاده کرده اند(۲۵). درصد صحت در الگوریتم های نیوی بیز، k نزدیک ترین همسایه و درخت تصمیم به ترتیب برابر با ۸۴/۵۲ درصد، ۸۱/۹۴ درصد و ۷۶/۷۷ درصد است. دو محقق زورارپاسی و اوزل(۲۶) از ترکیب الگوریتم تکاملی تفاضلی و کلونی زنبور مصنوعی برای تشخیص بیماری هپاتیت استفاده کرده اند. درصد صحت در مدل ترکیبی برابر با ۹۴/۴۰ درصد است. در (۲۷) از الگوریتم ژنتیک و ماشین بردار پشتیبان برای تشخیص بیماری هپاتیت استفاده شده است. در مدل ترکیبی از الگوریتم ژنتیک برای انتخاب ویژگی استفاده شده است. درصد صحت الگوریتم ژنتیک و ماشین بردار پشتیبان برابر با ۸۹/۶۷ درصد است. درصد صحت مدل پیشنهادی در مقایسه با مدل ترکیبی الگوریتم ژنتیک و ماشین بردار پشتیبان بیشتر است.

اهمیت تشخیص به موقع بیماری ها با استفاده از مدل های پیش بینی از دهه های گذشته توسط محققین درک شده است. جهت مدیریت بهتر سلامتی و کاهش مصرف خدمات سلامت نیاز است که میزان خطر ابتلا به بیماری یا پیشرفت بیماری هپاتیت در میان کلیه افراد یک جمعیت، شناســـایی شود. یافته های این مطالعه می تواند در راستای تهیه نرم افزارهای مدیریت بیماری و هم چنین تشخیص نوع بیماری در آینده مورد استفاده قرار گیرد. این مطالعه دارای محدودیت هایی نیز هست. الگوریتم جلبک مصنوعی و الگوریتم های فراابتکاری در مقایسه با مدل های داده کاوی نیازمند زمـــــان طولانی تری برای یافتن الگوهای مفید هستند. هم چنین جهت آمـــوزش بهتر مـــــدل به پردازشگرهای قدرتمندی نیاز اســـــت تا این که چرخه تکرار برای حالت های بیشتر اجرا شود و دقت تشخیص افزایش یابد. هم چنین داده های مفقود شده در نمونه ها و احتمال خطـــــا در ثبـــــت داده به دلـــــیل به کارگیری روش های سنتی از دیگر محـــــدودیت های این مطالعه است.

References

1.Venkatesh AG, Brickner H, Looney D, Hall DA, Aronoff spencer clinical detection of hepatitis C viral infection by yeast secreted HCV core gold binding peptide. Biosens Bioelectron 2018; 119: 230-6. doi. 10.1016/j.bios.2018.07.026

2.Hsu WY. A decision making mechanism for assessing risk factor significance in cardiovascular diseases. Dec Sup Sys2018; 115: 64-77. doi.10.1016/j.dss.2018.09.004

3.Malmir B, Amini M, Chang SI. A medical decision support system for disease diagnosis under uncertainty. Exp Sys Appl 2017; 88: 95-108. doi.10.1016/j.eswa.2017.06.031

4.Nazari S, Fallah M, Kazemipoor H, Salehipour A. A fuzzy inference fuzzy analytic hierarchy process based clinical decision support system for diagnosis of

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

heart diseases. Exp Sys Appl2018; 95: 261-271. doi.10.1016/j.eswa.2017.11.001
5.Rouhani M, Haghighi MM. The diagnosis of hepatitis diseases by support vector machines and artificial neural networks, international association of computer science and information technology. Spring Con2009; 2:456-8. doi.10.1109/IACSIT-SC.2009.25
6.Ozyilmaz L, Yildirim T. Artificial neural networks for diagnosis of hepatitis disease Proceedings. Inte Con Neural Net2003; 1: 586-9. doi.10.1109/IJCNN.2003.1223422
7.Zhang X, Wu C, Li J, Wang X, Jung KH. Binary artificial algae algorithm for multidimensional knapsack problems. Appl Soft Comput2016; 43: 583-95. doi.10.1016/j.asoc.2016.02.027
8.Martin B. Instance based learning nearest neighbour with generalization.2 th ed. Doct Disser Uni Waikato Publication. 1995;P.132-9.
9.Behravesh A, Sadeghzadeh M. Diagnosis of hepatitis disease with cuckoo search binary method and least squares support vector. Norwe Sci Technol Res 2016;3:32-7.
10.Ebrahimi M, Roosta M , Farjami Y. [Review and comparison of data based techniques in the diagnosis of hepatitis international conference on nonlinear systems and optimization of electrical and computer engineering]. Dubai Pand Andish Rahpo 2015;1:123-6. (Persian)
11.Khodarahmi R, Rouhani M. Detection of hepatitis disease using combination of k-means and weed optimization algorithms. Inte Con Mod Res Eng Sci 2016;1:43-7.
12.Kharshadizadeh N, Rezaei H. Diagnosis of hepatitis disease using rave collection and machine learning. Comput Eng Inf Technol Manage Sci 2014;2:162-7.
13.Mousavirad SJ, Komala HE. [Intelligent diagnosis of hepatitis disease by analyzing primary components and Fusion of classifiers]. Koomesh 2013;16:149-58. (Persian)
14.Nilashi M, Ahmadi H, Shahmoradi L, Ibrahim O, Akbari E. A predictive method for hepatitis disease diagnosis using ensembles of neuro fuzzy technique. J Inf Publ Health 2018;2:23-8. doi. 10.1016/j.jiph.2018.09.009
15.Uzer MS, Yilmaz N, Inan O. Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification.4th ed. Hindawi Publication Corp.2013; P.1-10. doi.10.1155/2013/419187
16.Nilashia M, Ahmadi H, Shahmoradi L, Othman I, Akbari E. A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. J Inf Publ Health. Volume 12, 2019;1:13-20. doi.10.1016/j.jiph.2018.09.009
17. Harper PR. A review and comparison of classification algorithms for medical decision making. Health Pol 2005; 71:315-31. doi. 10.1016/j.healthpol.2004.05.002
18.AlJarullah A. Decision tree discovery for the diagnosis of type II diabetes. Int Con Inform Technol 2011; 3: 303-7. doi.10.1109/INNOVATIONS.2011.5893838
19.Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Patt Rec 2008; 41:3692-705. https://doi.org/10.1016/j.patcog.2008.05.019
20.Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlock 2018; 10: 100-107. doi.10.1016/j.imu.2017.12.006
21.Edla DR, Cheruku R. Diabetes finder a bat optimized classification system for type-2 diabetes. Procedi Comput Sci 2017; 115: 235-42. doi.10.1016/j.procs.2017.09.130
22.Bascil MS, Oztekin H. A study on hepatitis disease diagnosis using probabilistic neural network, J Med Sys2012;36: 1603-6. doi.10.1007/s10916-010-9621-x
23.Dogantekin E, Dogantekin A, Avci D. Automatic hepatitis diagnosis system based on Linear discriminant analysis and adaptive network based on Fuzzy inference system. Exp Sys Appl2009; 36: 11282-11286. doi.10.1016/j.eswa.2009.03.021
24.Sartakhti JS, Zangooei MH, Mozafari K.Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing. Compute Meth Prog Biomed2012; 108: 570-9. doi.10.1016/j.cmpb.2011.08.003
25.Ashraf M, Chetty G, Tran D, Sharma D.Hybrid approach for diagnosing thyroid, hepatitis, and breast cancer based on correlation based feature selection and naive

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

bayes. ICONIP 2012; 3: 272-80. doi.10.1007/978-3-642-34478-7_34
26.Zorarpacı E, Ozel SA.A hybrid approach of differential evolution and artificial bee colony for feature selection, Exp Sys Appl2016; 62: 91-103. doi.10.1016/j.eswa.2016.06.004
27.Tan KC, Teoh EJ, Yu Q, Goh KC.A hybrid evolutionary algorithm for attribute selection in data mining. Exp Sys Appl2009; 36: 8616-30. doi. 10.1016/j.eswa.2008.10.013

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]

Analysis of Hepatitis Patient Data using Binary Artificial Algae Algorithm based on K-Nearest Neighbor

Biglarisaleh A[1], Soleimaniangharehchopogh F[1]*

(Received: December 31, 2018 Accepted: December 21, 2019)

Abstract

Introduction: The timely diagnosis and prediction of diseases are among the main issues in medical sciences. The use of decision-making systems to discover the underlying knowledge in the disease information package and patient records is one of the most effective ways of diagnosing and preventing disease. This study aimed to design a medical decision system that can detect hepatitis.

Materials & Methods: This study was conducted based on a descriptive-analytic design. Its dataset contains 155 records with 19 features in the University of California-Irvine machine learning database. This study utilized the Binary Artificial Algae Algorithm (BAAA) for Feature Selection (FS). Moreover, K-Nearest Neighbor (KNN) was used to classify hepatitis into two healthy and unhealthy classes. In total, 80% of the data was employed for training, and the remaining (20%) was used for testing. Furthermore, Precision, Recall, F-measure, and Accuracy were utilized to evaluate the model.

Findings: According to the results, the accuracy of the proposed model was estimated at 96.45%. After selecting the features with the BAAA, the percentage of the accuracy reached 98.36% in the best situation. In the proposed model with 300 repetitions, the Precision, Recall, F-Measure, and error rate were 96.23%, 96.74%, 96.48%, and 3.55%, respectively.

Discussion & Conclusions: Hepatitis is one of the most common diseases among females and males. A timely diagnosis of this disease not only reduces the costs but also increases the chance of successful treatment. In this study, the disease was diagnosed using the hybrid method, and a high accuracy level was obtained in disease diagnosis by FS.

Keywords: Binary artificial algae algorithm, Feature selection, Hepatitis disease diagnosis, K-nearest neighbor, Medical decision making system

1. 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

[Downloaded from sjimu.medilam.ac.ir on 2026-11-11]

[DOI: 10.29252/sjimu.28.2.59]