University of Qom

Engineering Management and Soft Computing

JEMSC.QOM
Online ISSN: 2538-2675

Provide a method to diagnose and optimize diabetes using data mining methods and firefly algorithm

Reza Molaee Fard

MSC. Department of industrial engineering, Azad university of dezfull. Email: rezamolae4@gmail.com

Article Info

Article type:
Research Article

Article history:
Received 2023 March 23
Received in revised form 2023 May 27
Accepted 2023 May 20
Published online 2023 September 16

Keywords:
Data Mining,
DBSCAN Algorithm,
Diabetes Diagnosis,
Firefly Algorithm,
SVM Algorithm.

ABSTRACT

Diabetes is one of the most common, dangerous and costly diseases in the world today, which is increasing at an alarming rate. The use of data mining methods can help in the early diagnosis of diabetes, which prevents the progression of this disease and many of its complications such as cardiovascular disease, vision problems and kidney disease. Providing care and health services to people with diabetes provides useful information that can be used to identify, treat, follow-up care and even prevent diabetes. In this study, a new method is presented to improve the diagnosis and prevention of diabetes using data mining methods. In this research, the DBSCAN clustering algorithm is used to cluster the data. Then, using SVM, we classify the data to identify useful data, and finally, with the firefly algorithm, we increase the obtained data to increase we optimize performance with this algorithm. The results of this study indicate that the DBSCAN algorithm is more efficient than other clustering algorithms. Also, the SVM algorithm can achieve 98% accuracy, which compared to other data mining algorithms could achieve a higher accuracy percentage.

Cite this article: Molaee Fard, R. (2023). Provide a method to diagnose and optimize diabetes using data mining methods and firefly algorithm. *Engiiiring Management and Soft Computing, 9* (1). 36-48. **DOI**: https://doi.org/10.22091/JEMSC.2022.6575.1147

DOI: https://doi.org/10.22091/JEMSC.2022.6575.1147

Publisher: University of Qom

JEMSC.QOM
شاپا الکترونیکی: ۲۶۷۵-۲۵۳۸

مدیریت مهندسی و رایانش نرم

دانشگاه قم

ارائه روشی به‌منظور تشخیص و بهینه‌سازی بیماری دیابت با استفاده از روش‌های داده‌کاوی والگوریتم کرم شب‌تاب

رضا مولایی فرد

کارشناس ارشد، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد دزفول. رایانامه: rezamolae4@gmail.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۱/۰۳

تاریخ بازنگری: ۱۴۰۲/۰۳/۰۶

تاریخ پذیرش: ۱۴۰۲/۰۳/۰۹

تاریخ انتشار: ۱۴۰۲/۰۶/۲۵

کلیدواژه‌ها:
الگوریتم کرم شب‌تاب،
الگوریتم DBSCAN،
الگوریتم SVM،
تشخیص دیابت،
داده‌کاوی.

چکیده

بیماری دیابت یکی از شـایع‌ترین، خطرناک‌ترین و پرهزینه‌ترین بیماری‌های حال حاضـر دنیاسـت که با نرخ هشـداردهنده‌ای در حال افزایش اسـت. اسـتفاده از روش‌های داده‌کاوی می‌تواند به تشـخیص زودهنگام دیابت کمک کند که باعث جلوگیری از پیشـرفت این بیماری و خیلی از عوارض آن مانند بیماری‌های قلبی و عروقی، مشـکلات بینایی و بیماری‌های کلیوی می‌شـود. ارائه خدمات مراقبتی و بهداشـتی افراد مبتلا به بیماری دیابت اطلاعات مفیدی ایجاد می‌کند که با استفاده از این اطلاعات می‌توان برای شناسایی، درمان، مراقبت‌های بعدی و حتی پیشـگیری از بیماری دیابت اسـتفاده نمود. در این تحقیق به ارائه روش جدیدی به منظور بهبود تشـخیص و پیشـگیری از بیماری دیابت با اسـتفاده از روش‌های داده‌کاوی پرداخته می‌شـود. در این تحقیق از الگوریتم خوشـه‌بندی DBSCAN جهت خوشـه‌بندی داده‌ها اسـتفاده می‌شـود سـپس با اسـتفاده از SVM، داده‌ها را جهت تشخیص داده‌های مفید، دسته‌بندی می‌کنیم و در نهایت با الگوریتم کرم شب‌تاب داده‌های به‌دست آمده را جهت افزایش کارآیی با این الگوریتم بهینه می‌کنیم. نتایج حاصـل از این تحقیق حاکی از کارآیی بالاتر الگوریتم DBSCAN نسبت به سایر الگوریتم‌های خوشه‌بندی است. همچنین الگوریتم SVM می‌تواند دقت ۹۸ درصد را به‌دست آورد که در مقایسه با سایر الگوریتم‌های داده‌کاوی توانست درصد دقت بیشتری را کسب کند.

استناد: فتحی هفشجانی، کیامرث؛ باقری سرخی، مجید و مدیری، محمود. (۱۴۰۲). «ارائه روشی به‌منظور تشخیص و بهینه‌سازی بیماری دیابت با استفاده از روش‌های داده‌کاوی والگوریتم کرم شب‌تاب». *مدیریت مهندسی و رایانش نرم*، دوره ۹ (۱). صص: ۴۸-۳۶. https://doi.org/10.22091/JEMSC.2022.6575.1147

ناشر: دانشگاه قم

۱) مقدمه

امروزه دیابت تبدیل به یک بیماری گسترده در سراسر دنیا شده‌است که میلیون‌ها نفر به آن مبتلا هستند. طبق آخرین آمارهای منتشرشده سازمان بهداشت جهانی و فدراسیون بین‌المللی دیابت، از هر ۴ نفر بالای ۶۰ سال یک نفر به این بیماری مبتلا می‌باشد. از این رو این بیماری به یکی از چالش‌های مهم نظام بهداشت و درمان کشورهای مختلف تبدیل شده‌است. عدم تشخیص به‌موقع و یا ضعف در تشخیص این بیماری ازجمله مشکلات عمده دیگری است که در رابطه با این بیماری وجود دارد. این امر تا حد زیادی به دلیل عدم انتخاب الگوی مناسب توسط پزشک یا عدم استفاده از الگوهای استاندارد موجود است. بنابراین پیاده‌سازی روشی که بتواند هرفرد را در تشخیص صحیح ابتلا یا عدم ابتلا به این بیماری یاری رساند، می‌تواند گام مهمی در جهت پیشگیری و کنترل این بیماری بخصوص در مراحل ابتدایی آن تلقی گردد. بیماری دیابت یکی از شایع‌ترین، خطرناک‌ترین و پرهزینه‌ترین بیماری‌های حال حاظر دنیاست که با نرخ هشداردهنده‌ای در حال افزایش است. استفاده از روش‌های داده‌کاوی می‌تواند به تشخیص زودهنگام دیابت کمک کند که باعث جلوگیری از پیشرفت این بیماری و خیلی از عواض آن مانند بیماری‌های قلبی و عروقی، مشکلات بینایی و بیماری‌های کلیوی می‌شود. ارئه خدمات مراقبتی و بهداشتی افراد مبتلا به بیماری دیابت اطلاعات مفیدی ایجاد می کند که با استفاده از این اطلاعات می‌توان برای شناسایی، درمان، مراقبت‌های بعدی و حتی پیشگیری از بیماری دیابت استفاده نمود. از طرفی کاوش و بررسی حجم زیادی از این اطلاعات، نیازمند استفاده از روش‌های مؤثر و کارآمدی برای یافت الگوهای مربوط در این اطلاعت است که استفاده از تکنیک‌های مختلف داده‌کاوی بخصوص دسته‌بندی و الگوهای تکرارشونده می‌تواند کمک بسیار زیادی در این زمینه باشد. آنچه مسلم است با افزایش سیستم‌های الکترونیک سلامت، حجم داده‌های پزشکی روزبه‌روز در حال افزایش است. اما این مجموعه داده‌های بزرگ به‌طور خام هیچ کاربردی ندارند. برای آنکه بتوان از این داده‌ها، اطلاعات مفیدی استخراج کرد، نیاز به تحلیل داده‌ها و تبدیل این اطلاعات به اطلاعات مفید است. با توجه به حجم بالای اطلاعات، استفاده از عامل انسانی به‌عنوان تشخیص‌دهنده الگو و تحلیلگر داده‌ها پاسخگو نمی‌باشد لذا داده‌کاوی روی داده‌های پزشکی از اهمیت بالایی برخوردار است. داده‌کاوی را می‌توان از جنبه‌های مختلف در پیشگیری یا تشخیص انواع بیماری، انتخاب روش‌های درمان و مدت زمان بستری بیمار بکار برد. داده‌کاوی برای بررسی داده‌های زیاد و استخراج ویژگی‌ها از میان حجم انبوه اطلاعات مورد استفاده قرار می‌گیرد. از بین الگوریتم‌های داده‌کاوی، الگوریتم‌های دسته‌بندی بهترین کارآیی را در بین این الگوریتم‌ها دارا می‌باشند زیرا در این الگوریتم‌ها نمونه ها دارای برچسب کلاس هستند و هدف از تعیین برچسب کلاس، ایجاد یک نمونه جدید می‌باشد. داده‌کاوی روشی برای کشف الگوهای پنهان و استخراج اطلاعات معنی دار از مجموعه داده‌های بزرگ است. درواقع داده‌کاوی بخشی از فرآیند استخراج دانش است که هدف آن دستیابی به دانش نهفته در داده‌ها با کمترین دخالت انسانی است. در سال‌های اخیر تحقیقات زیادی در زمینه پیشگیری و درمان بیماری دیابت صورت گرفته‌است. در این تحقیق نیز با بهره‌گیری از روش‌های پیشین به ارائه روش جدیدی به منظور بهبود و پیشگیری درمان بیماری دیابت پرداخته می‌شود که با رفع ایرادات روش‌های قبلی، سیستم پیشنهادی می‌تواند نتایج دقیق‌تری کسب کند.

۲) کارهای پیشین

پراسد و همکاران در مقاله خود در سال ۲۰۲۰ به ارائه روشی به‌منظور بهبود تشخیص دیابت با استفاده از الگوریتم‌های داده‌کاوی پرداختند. این محققان برای تشخیص و پیشگیری بیماری دیابت از چهار نوع الگوریتم داده‌کاوی استفاده کردند. این الگوریتم‌ها برای استفاده در مجموعه‌های آموزشی و آزمایشی و تولید جزئیات پیش‌بینی استفاده شده‌است. این الگوریتم‌ها با استفاده از معیارهایی مانند دقت، میانگین قدر مطلق خطا (MAE) و میانگین مربع خطای نقاط (RMSE) ارزیابی شدند. نتایج حاصل از تحقیق نشان داد که همه الگوریتم‌های یادگیری ماشین قادر به پیش‌بینی بیماری‌ها خواهند بود. با این حال؛ آنها از نظر سطح دقت متفاوت هستند. طبق تحقیقات صورت‌گرفته، جنگل تصادفی بالاترین دقت را در بین این الگوریتم‌ها داشته‌است و این الگوریتم می‌تواند پیش‌بینی دقیقی از تخصیص انواع بیماری‌ها داشته‌باشد[1] .

کازرونی و همکاران در مقاله خود در سال ۲۰۲۰ به یافتن و مقایسه روش‌های استخراج و پیش‌بینی اطلاعات مربوط به بیماری دیابت با استفاده از روش‌های داده‌کاوی پرداختند. این محققان به ارزیابی ۶ مورد از ابزارهای داده‌کاوی برای پیش‌بینی بیماری دیابت پرداختند که هدف از این ارزیابی و مقایسه، یافتن بهترین روش استخراج داده برای پیش‌بینی T2DM با استفاده از روش‌های داده‌کاوی بود. نتایج حاصل از تحقیق حاکی از میزان بالای AUC اختصاص داده‌شده به SVM و رگرسیون لجستیک و همچنین دو الگوریتم KNN و ANN بیشترین حساسیت متوسط و بیشترین ویژگی‌های تشخیص را نسبت به سایر الگوریتم‌های موجود دارا می‌باشد [2].

کومار و همکاران در مقاله خود در سال ۲۰۲۰ به بررسی روش‌های داده‌کاوی برای پیش‌بینی بیمارها پرداختند. این محققان بر این عقیده هستند که روش‌های داده‌کاوی برای تشخیص بیماری‌ها دارای اهمیت زیادی هستند و بر کسی پوشیده نیست که این اطلاعات می‌تواند برای متخصصان مراقبت‌های بهداشتی و درمان جهت اهداف مختلف مورد استفاده قرار گیرد. این محققان هدف از مقاله خود را ارزیابی و تجزیه و تحلیل داده‌ها با استفاده از سه روش منحصربه‌فرد Naïve Bayes (NB)، Support Vector Machine (SVM) و درخت تصمیم قرار دادند و همچنین برای تصمیم‌گیری رویکردهای بالقوه برای پیش‌بینی احتمال بیماری قلبی برای دیابتی‌ها بیماران وابسته به دقت پیش‌بینی آنها هستند [3].

علیرضایی و همکاران در مقاله خود در سال ۲۰۱۹ به ارائه روشی به‌منظور پیشگیری از بیماری دیابت با استفاده از الگوریتم‌های فرابتکاری پرداختند. این محققان در مقاله خود از روشی مبتنی بر الگوریتم خوشه‌بندی k-means برای شناسایی و حذف نقاط دور استفاده کردند. سپس به‌منظور انتخاب ویژگی‌های قابل توجه با بالاترین دقت طبقه‌بندی با استفاده از ماشین‌های بردار پشتیبانی (SVM) استفاده نمودند. علاوه بر این؛ برای اعتبارسنجی مدل ساخته‌شده از روش اعتباردهی ضربدری ۱۰ برابر (CV) استفاده کردند. نتایج حاصل از این تحقیق حاکی از آن بود که الگوریتم کرم شب‌تاب چندهدفه (MOFA) و الگوریتم رقابتی امپریالیستی چندهدفه (MOICA) با دقت طبقه‌بندی ۱۰۰ درصد از الگوریتم ژنتیک مرتب‌سازی غیرسلطه (NSGA-II) و چندهدفه پیشی می‌گیرند و بهینه‌سازی ازدحام ذرات (MOPSO) دقت ۹۴/۶ درصد را به‌دست آورد[4].

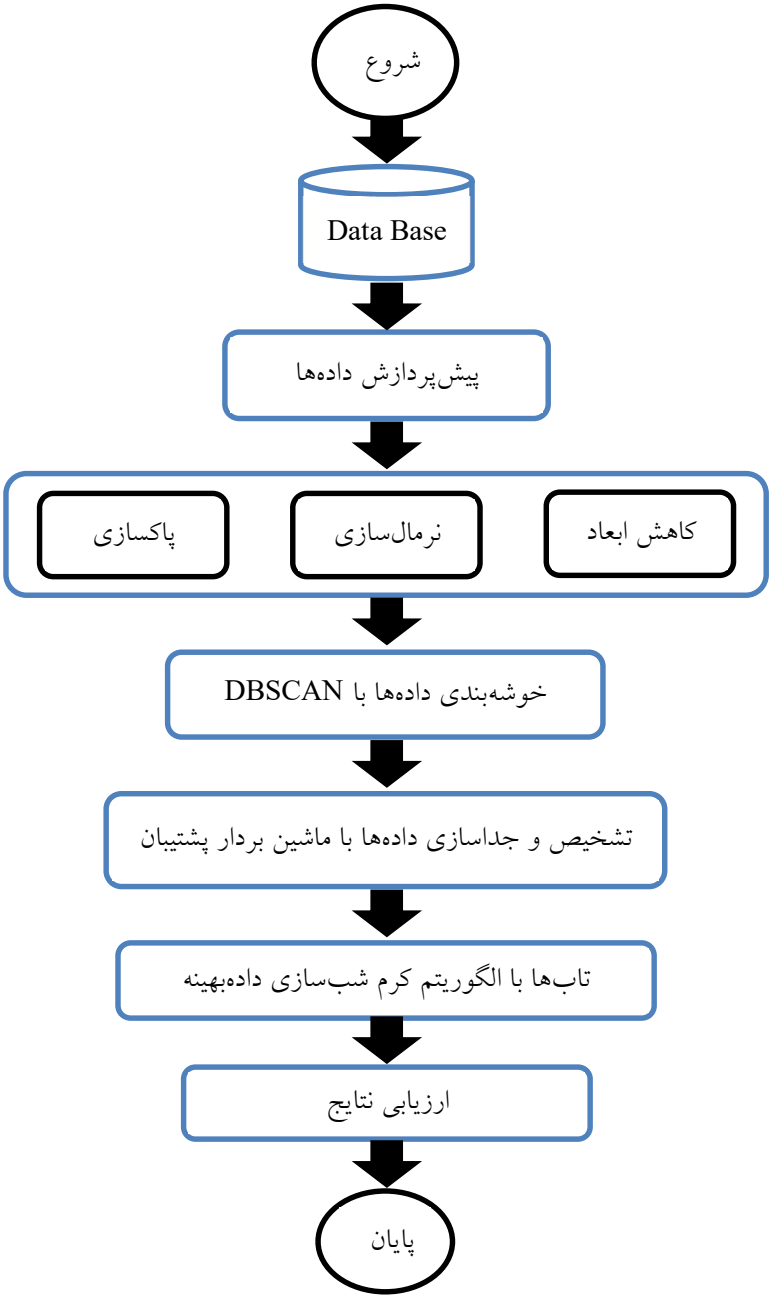
کائور وشارما در مقاله خود در سال ۲۰۱۸ به بررسی روش‌های پیشگبری و درمان بیماری دیابت با استفاده از ابزارهای داده‌کاوی پرداختند. این محققان از الگوریتم ژنتیک برای پیش‌بینی و پیشگیری از بیماری دیابت استفاده کردند. این

محققان در تحقیقات خود به این نکته پی‌بردند که استفاده از ابزارهای داده‌کاوی ۷۴ تا ۱۰۰ درصد می‌تواند به بهبود بیماری دیابت کمک نماید. روش کار این محققان استفاده از یک روش فرابتکاری ترکیبی با استفاده از دو الگوریتم ژنتیک و الگوریتم بهینه‌سازی ذرات بود. نتایج حاصل از این تحقیق حاکی از دقت ۸۷ درصدی روش پیشنهادی بود که می‌تواند تا حدود زیادی به فرآیند تشخیص بیماری دیابت کمک نماید [5].

بریک و همکاران در مقاله خود در سال ۲۰۲۱ به بررسی ابزارهای مفید داده‌کاوی برای پیشگیری و درمان بیماری دیابت پرداختند. بررسی این محققان نشان می‌دهد که حدود ۳۴۷ میلیون نفر از جمعیت جهان تحت تأثیر دیابت قرار دارند. دیابت نه تنها در فرد مسن بلکه بر روی نسل جوان نیز تأثیر می‌گذارد. تشخیص دیابت در مراحل اولیه نیز یک چالش بزرگ است. این تشخیص برای روند تصمیم‌گیری سیستم پزشکی مفید خواهد بود. پیش‌بینی اولیه دیابت به ما کمک می‌کند تا زندگی انسان را از دیابت نجات دهیم. طولانی‌شدن دیابت منجر به خطر آسیب در اندام‌های حیاتی بدن انسان می‌شود. بنابراین پیش‌بینی اولیه دیابت برای نجات انسان از دیابت بسیار مهم است. در این مقاله دو مجموعه از روش یادگیری ماشین برای پیش‌بینی دیابت ارائه شده‌است. یکی از آنها الگوریتم مبتنی بر طبقه‌بندی و دیگری الگوریتم ترکیبی است. در طبقه‌بندی، الگوریتم جنگل تصادفی مورد استفاده قرار گرفته‌است. برای رویکرد ترکیبی نیز، الگوریتم XGBoostاستفاده شده‌است. این دو الگوریتم به‌منظور بررسی دقت پیش‌بینی در دیابت برای دو روش یادگیری ماشین مختلف اجرا و مقایسه شدند و میانگین نمره ۷۴/۱۰ را کسب کردند که از الگوریتم Random Forest بهتر است [6].

۳) روش پیشنهادی

در روش پیشنهادی به ارائه روش جدیدی به‌منظور شناسایی، تشخیص و پیشگیری بیماری دیابت با استفاده از روش‌های داده‌کاوی پرداخته می‌شود. روش پیشنهادی بدین صورت می‌باشد که در ابتدا اطلاعات مربوط به بیماران را جمع‌آوری می‌کنیم. سپس برای این که دقت اطلاعات را از جهات مختلف بالا ببریم عملیات پیش‌پردازش داده‌ها را روی داده‌ها انجام می‌دهیم. عملیات داده‌کاوی از آن جهت مورد استفاده قرار می‌گیرد که داده‌های خام را نمی‌توان به‌طور مستقیم به ابزارهای داده‌کاوی تزریق کرد. در مراحل بعد و پس از پیش‌پردازش داده‌ای باید عملیات خوشه‌بندی را بر روی داده‌ها مورد استفاده قرار دهیم. روش خوشه‌بندی موردنظر استفاده از الگوریتم خوشه‌بندی DBSCAN می‌باشد. خوشه‌بندی داده به‌منظور یافتن بیشترین شباهت میان داده‌ها مورد استفاده قرار می‌گیرد. در مرحله بعد باید داده‌های به‌دست آمده از مراحل قبل را جهت تشخیص داده‌ها با استفاده از ماشین‌بردار پشتیبان دسته‌بندی کنیم و در مرحله آخر نیز جهت بهینه‌سازی و بهبود اطلاعات، داده‌ها را با استفاده از الگوریتم کرم شب‌تاب، بهینه می‌کنیم تا کارآیی الگوریتم به بالاترین حد خود برسد و در نهایت به ارزیابی داده‌ها می‌پردازیم.



شکل ۱. نمایی از روش پیشنهادی

۴) جمع‌آوری داده‌ها

برای شروع بکار و پیاده‌سازی روش پیشنهادی ابتدا باید پایگاه داده‌ای از لیست افراد مبتلا به بیماری دیابت و علائم این افراد را داشته باشیم تا بتوانیم روش پیشنهادی را پیاده‌سازی کنیم. برای انجام این کار یک فایل شامل اطلاعات ۸۰۴ بیمار را جمع‌آوری کردیم. بعد از جمع‌آوری اطلاعات مربوط به بیماران، کار را بر روی این داده‌ها آغاز کردیم.

۵) پیش‌پردازش داده‌ها

پس از مرحله جمع‌آوری داده‌ها باید عملیات پیش‌پردازش داده‌ها را بر روی داده‌ها انجام دهیم. در فرآیند داده‌کاوی نیاز

داریم تا داده‌ها برای الگوریتم آماده شوند زیرا معمولا نمی‌توان داده‌ها را به‌صورت خام به الگوریتم‌های داده‌کاوی و یادگیری ماشین تزریق کرد. برای آماده‌سازی داده‌ها نیاز است تا آنها را از شکل و حالت اولیه خارج کرده و به شکلی که برای الگوریتم مناسب باشد، تبدیل کرد. همچنین داده‌های موجود معمولا دارای زواید مختلفی هستند که ممکن است الگوریتم را دچار خطا کنند[7, 8]. در داده‌کاوی نیاز داریم تا داده‌های اضافی را که به مسئله و الگوریم کمکی نمی‌کنند، حذف کنیم. عملیات پیش‌پرازش داده‌ها معمولا قبل از عملیات اصلی الگوریتم‌های داده‌کاوی انجام می‌گیرند و باعث تسهیل و کمک به الگوریتم‌ها می‌شوند[9].

۶) پاکسازی داده‌ها

در این مرحه باید داده‌های موجود را پاکسازی کنیم. پاکسازی داده‌ها فرآیند از بین بردن خطاها و ناسازگاری‌ها در داده‌هاست و درواقع مرحله کنترل کیفی قبل از انجام تحلیل داده‌ها می‌باشد. اغلب به جهت خطاهای عملیاتی و پیاده‌سازی سیستم‌ها، داده‌های برآمده از منابع دنیای واقعی پرغلط، ناقص و ناسازگار هستند. لازم است در ابتدا چنین داده‌هایی تمیز شوند. این کار شامل برخی عملیات پایه مانند نرمال‌سازی، حذف نویز یا اغتشاش، مواجهه با داده‌های مفقوده، کاهش افزونگی و برطرف کردن داده‌هاست.

۷) نرمال‌سازی داده‌ها

پس از پیش‌پردازش داده‌ها، باید داده‌ها را نرمال کنیم. در نرمال‌سازی داده‌ها، تغییر داده‌ها به گونه‌ای است که آنها را به یک دامنه کوچک و معین مانند فاصله بین ۱- و ۱ نگاشت کنند. هدف نرمال‌سازی حذف افزونگی داده و باقی نگهداشتن وابستگی بین داده‌های مرتبط می‌باشد. این فرآیند اغلب باعث ایجاد جداول بیشتر می‌شود ولی اندازه‌گیری پایگاه داده را کاهش داده و بهبود کارآیی را تضمین می‌کند. روش‌های مختلفی جهت نرمال‌سازی داده‌ها وجود دارد که معروف‌ترین آنها می‌توان به روش Min – Max Normalization اشاره کرد. در این روش هرکدام از داده‌ها را می‌توان به یک بازه دلخواه تبدیل کرد. فرمول کلی این روش برای تبدیل داده‌ها به بازه بین ۰ تا ۱ به‌صورت زیر می‌باشد:

$$z = \frac{x - \min(x)}{\max(x) - min} \qquad (۱)$$

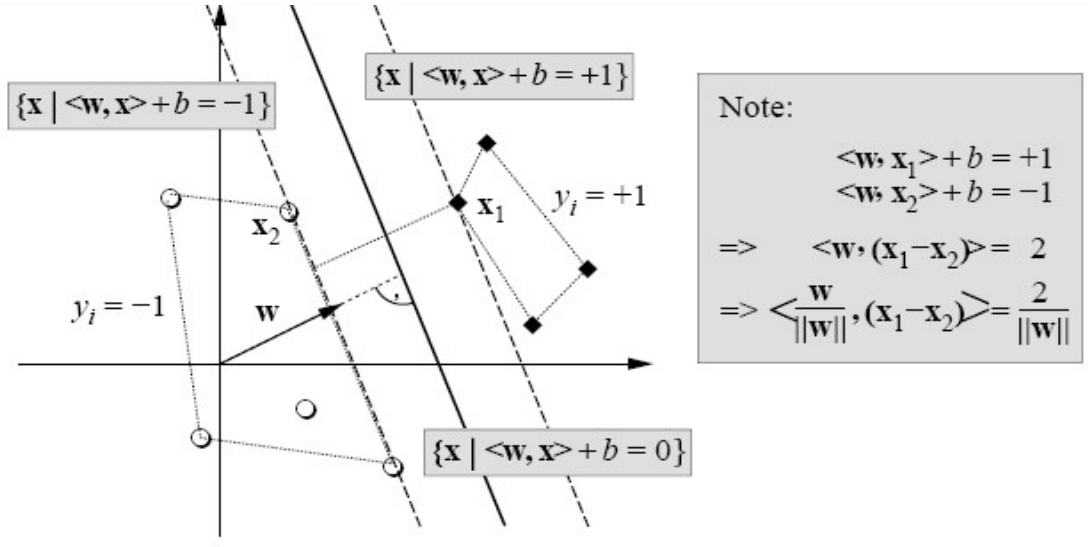
۸) خوشه‌بندی داده‌ها با الگوریتم DBSCAN

پس از پیش‌پردازش داده‌ها باید عملیات خوشه‌بندی داده‌ها را انجام دهیم. خوشه‌بندی از آن جهت مورد استفاده قرار خواهد گرفت که ما بتوانیم شباهت بین داده‌ها را پیدا کنیم. روش خوشه‌بندی موردنظر الگوریتم خوشه‌بندی DBSCAN می‌باشد که در بین الگوریتم‌های خوشه‌بندی موجود بالاترین میزان کارآیی را دارد. روش کار الگوریتم DBSCAN به این صورت می‌باشد که در این الگوریتم دو پارامتر وجود دارد. یکی از این پارامترها شعاع است که به آن Epsilon می‌گویند و دومین پارمتر، حداقل نقاط موجود در یک خوشه است که به آن Min point می‌گویند[10]. این الگوریتم ابتدا یک نمونه را انتخاب می‌کند و با توجه به شعاع Epsilon به دنبال همسایه برای این نقطه در فضا می‌گردد. اگر الگوریتم در آن شعاع مشخص Epsilon حداقل توانست به تعداد Min point نقطه پیدا کند، آنگاه همه آن نقاط با هم به یک خوشه

تعلق می‌گیرند. الگوریتم سپس به دنبال یکی از نقطه‌های همجوار نقطه فعلی می‌رود تا دوباره با شعاع Epsilon در آن نقطه به دنبال نقاط همسایه بگردد و اگر باز هم تعداد نقاط همسایه جدید پیدا شوند، این الگوریتم دوباره همه آن نقاط جدید را با نقاط قبلی به یک خوشه تعلق می‌دهد و اگر نقطه جدیدی در همسایگی پیدا نکرد این خوشه تمام شده‌است و برای پیدا کردن خوشه‌های دیگر در نقاط دیگر به‌صورت تصادفی یک نقطه دیگر را انتخاب کرده و شروع به یافتن همسایه و تشکیل خوشه جدید برای آن نقطه می‌کند. این کار آنقدر ادامه پیدا می‌کند تا تمام نقاط بررسی شوند[11, 12].

۹) تشخیص و دسته‌بندی داده‌ها با SVM

در مرحله بعد باید داده‌های خود را جهت تشخیص نمونه‌های سالم و نمونه‌های بیمار دسته‌بندی کنیم. روش پیشنهادی موردنظر استفاده از الگوریتم SVM می‌باشد که می‌تواند دسته‌بندی دقیقی را بر روی داده‌ها انجام دهد. در روش ماشین‌بردار پشتیبان، بردارهای ورودی به یک فضای چندبعدی نگاشت می‌شوند. پس از آن، یک ابرسطح ساخته خواهد شد که با حداکثر فاصله ممکن، بردارهای ورودی را از هم جدا خواهد کرد [13]. به این ابرسطح، ابرسطح با حداکثر مرز جداکننده گفته می‌شود. همانگونه که در شکل 2 نشان داده شده‌است، دو ابرسطحِ موازی در دو سمتِ ابرسطح با حداکثر مرز جداکننده ساخته خواهد شد که داده‌های مربوط به دو طبقه را به‌گونه‌ای از هم مجزا می‌کنند که هیچ داده‌ای در مرز بین این دو ابرسطح قرار نمی‌گیرد. ابرسطح با حداکثر مرز جداکننده، ابرسطحی است که فاصله بین دو ابرسطحِ موازی را به حداکثر می‌رساند [14,15]. فرض بر این است که هرچقدر مرز جداکننده یا درواقع فاصله بین دو ابرسطح موازی بیشتر باشد، خطای طبقه‌بندی هم کمتر خواهد بود. در پایان این مرحله، داده‌ها دسته‌بندی و به دو طبقه مجزا تقسیم می‌شوند که از طریق این تقسیم‌بندی می‌توان داده‌ها را جهت داده‌های صحیح و غیرصحیح تقسیم نمود.



شکل۲. نحوه ساخت ابرسطح جداکننده بین دو طبقه داده در فضای دو بعدی

۱۰) بهینه‌سازی داده‌ها با الگوریتم کرم شب‌تاب

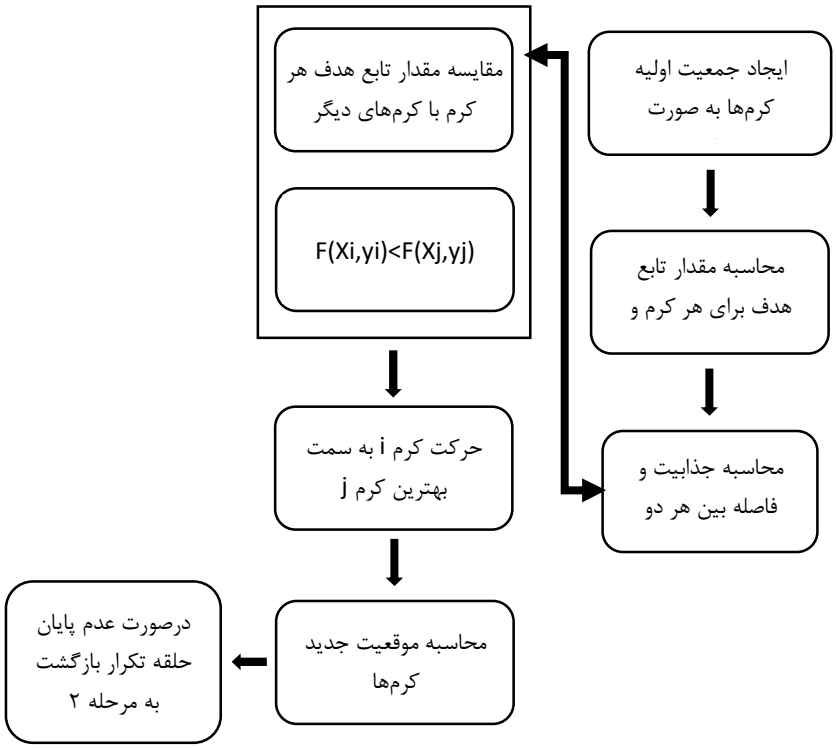
در این مرحله باید با استفاده از الگوریتم کرم شب‌تاب داده‌های خود را بهینه کنیم. ایده الگوریتم کرم شب‌تاب از ارتباط نوری میان کرم‌های شب‌تاب الهام گرفته شده‌است [16]. این الگوریتم را می‌توان از مظاهر هوش ازدحامی دانست که در آن از همکاری و احتمالا رقابت اعضای ساده و کم‌هوش، مرتبه بالاتری از هوشمندی ایجاد می‌شود که قطعا توسط هیچ‌یک از اجزا قابل حساب نیست. این الگوریتم یک الگوریتم هوش‌جمعی است [17]. هر فرد (که در این الگوریتم کرم شب‌تاب نامیده می‌شود) در جمعیت بیانگر یک راه‌حل بالقوه در یک فضای چندبعدی است [18]. به دلیل قدرت جذب بین کرم‌های شب‌تاب، آنها به سمت بقیه مکان‌ها حرکت می‌کنند تا راه‌حل‌های بهتری را بیابند [19,20]. در الگوریتم کرم شب‌تاب، میزان جاذب بودن براساس قدرت تابش نور و روشنایی آن فرد تعیین می‌شود. این میزان در اصل متناسب با میزان شایستگی یک فرد است. یک مسئله پیداکردن بیشینه است. برای کرم شب‌تاب x، ارتباط میان شایستگی و میزان تابش نور و روشنایی آن کرم می‌تواند به‌صورت $I(x) \propto f(x)$ بیان شود. میزان جاذب بودن که با β نشان داده می‌شود، متناسب با r خواهد بود. با افزایش این فاصله، میزان جاذب بودن به‌تدریج کاهش خواهد یافت [21]. درنظر بگیرید که X_i بیانگر i امین فرد در جمعیت باشد، در این صورت میزان جاذب بودن فرد i ام توسط فرد j ام با استفاده از رابطه زیر بیان خواهد شد.

$$\boldsymbol{\beta(r_{i,j}) = \beta_0 e^{-r^2_{ij}} \; where \; r_{ij} = \|X_i - X_j\|} \qquad (۲)$$

که در آن $\|X_i - X_j\|$ فاصله اقلیدسی دو کرم شب‌تاب i و j و پارامتر β_0 میزان جاذب بودن در فاصله صفر را بیان می‌کند. Y نیز ضریب جذب نور را بیان می‌کند. بهترین مقدار برای y برابر $\frac{1}{r^2}$ است که در آن r مقیاس طول برای متغیرهای طراحی‌شده خواهد بود. برای هر کرم شب‌تاب X_i، در مقابسه با کرم X_j، اگر X_j روشنایی بیشتری نسبت به X_i براساس میزان جاذب بودن X_j به سمت X_j حرکت خواهد کرد. میزان این حرکت براساس معادله زیر به‌دست می‌آید.

$$\boldsymbol{x_{id}(t+1) = x_{id}(t) + \beta_0 e^{-r^2_{ij}} \left(x_{jd}(t) - x_{id}(t)\right) + \; a\varepsilon_i} \qquad (۳)$$

که در آن، x_{id} مولفه d ام فرد iام در جمعیت، d بعد راه‌حل، α ضریب اهمیت به حرکت تصادفی، ε_i یک عدد تصادفی و t شاره نسل می‌باشند. چهارچوب الگوریتم کرم شب‌تاب را در شکل زیر مشاهده می‌کنید.



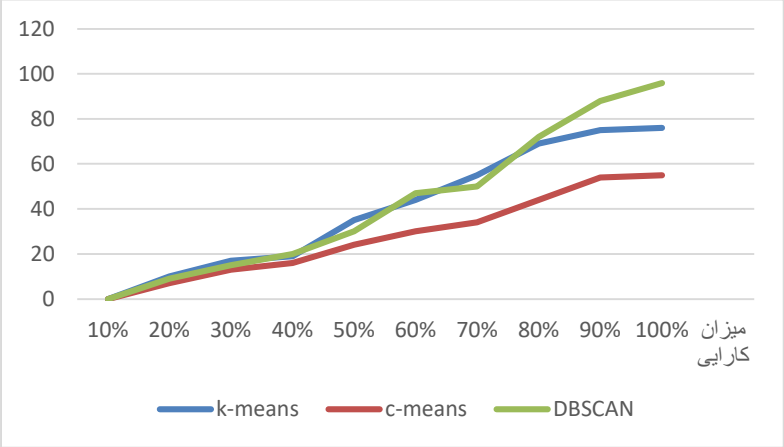
شکل۳. چهارچوب الگوریتم کرم شب‌تاب

۱۱) ارزیابی روش پیشنهادی

برای ارزیابی روش پیشنهادی ابتدا مقایسه‌ای بین الگوریتم خوشه‌بندی DBSCAN و سایر روش‌های خوشه‌بندی از لحاظ میزان کارآیی صورت گرفت. سپس مقایسه‌ای بین الگوریتم SVM و سایر الگوریتم‌های دسته‌بندی از لحاظ میزان دقت صورت گرفت و در نهایت نیز الگوریتم کرم شب‌تاب با سایر الگوریتم‌های بهینه‌سازی مقایسه گردید که نتایج حاصل از ارزیابی روش پیشنهادی را در نمودارهای زیر می‌توان مشاهده نمود. به‌منظور ارزیابی میزان کارآیی الگوریتم DBSCAN مقایسه‌ای بین این الگوریتم و دو الگوریتم k-means و c-means صورت گرفت که نتایج حاصل از ارزیابی حاکی از عملکرد بهتر الگوریتم استفاده‌شده در روش پیشنهادی تحقیق بود. در این ارزیابی الگوریتم DBSCAN توانست میزان کارآیی ۹۶ درصد را به‌دست آورد. این در حالی بود که الگوریتم k-means، کارآیی ۷۶ درصد و الگوریتم c-means، کارآیی ۵۵ درصد را به‌دست آورد که نتایج حاصل از این ارزیابی را می‌توان در جدول زیر مشاهده نمود.

جدول ۱. ارزیابی میزان کارآیی الگوریتم DBSCAN با سایر روش‌های خوشه‌بندی

نام الگوریتم	میزان کارآیی
K-Means	76 درصد
C-Means	۵۵ درصد
DBSCAN	۹۶ درصد

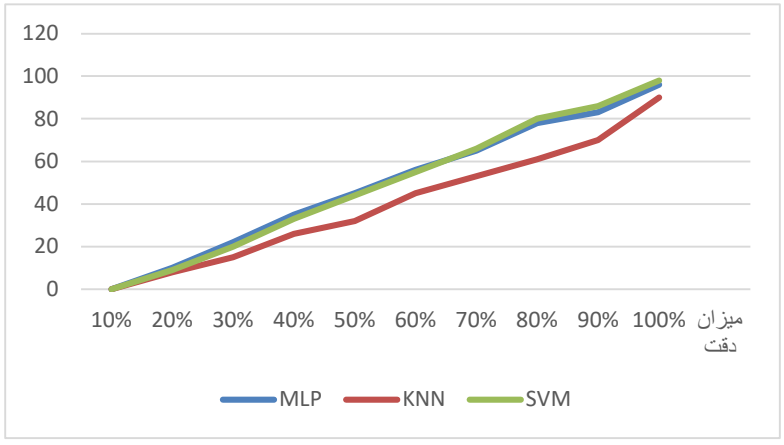


شکل٤. مقایسه میزان کارآیی الگوریتم خوشه‌بندی DBSCAN با دو الگوریتم K-means و C-means

به‌منظور ارزیابی میزان دقت الگوریتم SVM نیز مقایسه‌ای بین این الگوریتم و الگوریتم‌های KNN و MLP صورت گرفت. نتایج حاصل از این ارزیابی حاکی از درصد دقت بالاتر الگوریتم SVM بود. به این صورت که الگوریتم SVM توانست میزان دقت ۹۸ درصد را به‌دست آورد درحالیکه الگوریتم KNN میزان دقت ۹۰ درصد و الگوریتم MLP میزان دقت ۹۶ درصد را به‌دست آورد. نتایج حاصل از این مقایسه را در جدول زیر مشاهده می‌کنید.

جدول ۲. ارزیابی میزان دقت روش پیشنهادی و سایر روش‌ها

نام الگوریتم	میزان دقت
SVM	۹۸ درصد
KNN	۹۰ درصد
MLP	۹٦ درصد



شکل۵. مقایسه میزان دقت الگوریتم SVM با دو الگوریتم KNN وMLP

به‌منظور ارزیابی میزان روش‌های بهینه‌سازی نیز مقایسه‌ای بین الگوریتم کرم شب‌تاب و الگوریتم‌های PSO و الگوریتم گرگ خاکستری صورت گرفت. نتایج حاصل از میزان ارزیابی قسمت بهینه‌سازی حاکی از میزان بهینه‌سازی بالاتر الگوریتم کرم شب‌تاب بود تا جایی که الگوریتم کرم شب‌تاب توانست میزان دقت ۹۵ درصد و الگوریتم PSO دقت

۸۵ درصد و الگوریتم گرگ خاکستری دقت ۹۰ درصد را به‌دست آورد. نتایج حاصل از این مقایسه را در جدول زیر مشاهده می کنید.

جدول۳. ارزیابی و مقایسه میزان بهینه‌سازی الگوریتم کرم شب‌تاب و سایر روش‌ها

نام الگوریتم	میزان دقت روش بهینه سازی
الگوریتم PSO	۸۵ درصد
الگوریتم گرگ خاکستری	۹۰ درصد
الگوریتم کرم شب‌تاب	۹۵ درصد

شکل٦. مقایسه الگوریتم کرم شب‌تاب و دو الگوریتم PSO و الگوریتم گرگ خاکستری

۱۲) نتیجه‌گیری

داده‌کاوی روی داده‌های پزشکی از اهمیت بالایی برخوردار است. داده‌کاوی را می‌توان از جنبه‌های مختلف در پیشگیری یا تشخیص انواع بیماری، انتخاب روش‌های درمان و مدت زمان بستری بیمار بکار برد. داده‌کاوی برای بررسی داده‌های زیاد و استخراج ویژگی‌ها از میان حجم انبوه اطلاعات مورد استفاده قرار می‌گیرد. از بین الگوریتم‌های داده‌کاوی، الگوریتم‌های دسته‌بندی بهترین کارآیی را در بین این الگوریتم‌ها دارا می‌باشند زیرا در این الگوریتم‌ها، نمونه‌های دارای برچسب کلاس هستند و هدف از تعیین برچسب کلاس، ایجاد یک نمونه جدید می‌باشد. داده‌کاوی روشی برای کشف الگوهای پنهان و استخراج اطلاعات معنی‌دار از مجموعه داده‌های بزرگ است. در این نوع از مطالعات، متغیرهای فیزیکی و خونی عده‌ای از بیماران دیابتی و افراد معمولی به الگوریتم‌های دسته‌بندی در داده‌کاوی داده می‌شود. این الگوریتم‌ها می‌توانند مدل‌هایی را برای رده‌بندی بیماران به دو رده بیمار دیابتی و افراد سالم ایجاد نماید. از مدل‌های ایجادشده می‌توان به‌منظور رده‌بندی مراجعین جدید و افراد مشکوک به بیماری دیابت استفاده نمود و بیماری یا سلامت افراد جدید را پیش‌بینی کرد. در این تحقیق سعی کردیم که با استفاده از الگوریتم‌های داده‌کاوی به پیش‌بینی و تشخیص بیماری دیابت کمک کنیم. نتایج حاصل از تحقیق حاکی از آن است که روش پیشنهادی موجود در این تحقیق می‌تواند تا بیش از ۹۸ درصد به درستی بیماری را تشخیص داده و توانسته دقت بیشتری را نسبت به سایر الگوریتم‌های موجود به‌دست آوریم.

منابع

Alirezaei, M., Niaki, S. T. A., & Niaki, S. A. A. (2019). A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. Expert Systems with Applications, 127, 47-57. https://doi.org/10.1016/j.trb.2017.04.003

Barik, S., Mohanty, S., Mohanty, S., & Singh, D. (2021). Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In Intelligent and Cloud Computing (pp. 399-409). Springer, Singapore. https://doi.org/1094/j.trb.2002.24.45

Dey, N. (2020). Applications of firefly algorithm and its variants. Springer Singapore. https://doi.org/1061/j.trb.2023.30.113

Gopi, A. P., Jyothi, R. N. S., Narayana, V. L., & Sandeep, K. S. (2020). Classification of tweets data based on polarity using improved RBF kernel of SVM. International Journal of Information Technology, 1-16. https://doi.org/1033/j.trb.2022.20.12

Kaur, P., & Sharma, M. (2018). Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review. Int. J. Pharm. Sci. Res, 9, 2700-2719. https://doi.org/1093/j.trb.2017.22.63

Kazerouni, F., Bayani, A., Asadi, F., Saeidi, L., Parvizi, N., & Mansoori, Z. (2020). Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches. BMC bioinformatics, 21(1), 1-13. https://doi.org/1019/j.trb.2018.14.13

Kouziokas, G. N. (2020). SVM kernel based on particle swarm optimized vector and Bayesian optimized SVM in atmospheric particulate matter forecasting. Applied Soft Computing, 93, 106410. https://doi.org/1086/j.trb.2010.32.80

Kumar, A., Kumar, P., Srivastava, A., Kumar, V. A., Vengatesan, K., & Singhal, A. (2020, April). Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients. In International Conference on Advances in Computing and Data Sciences (pp. 507-518). Springer, Singapore. https://doi.org/1083/j.trb.2012.35.14

Li, Z., Li, Y., Lu, W., & Huang, J. (2020). Crowdsourcing logistics pricing optimization model based on DBSCAN clustering algorithm. IEEE Access, 8, 92615-92626. https://doi.org/1074/j.trb.2018.9.61

Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). Big data preprocessing. Cham: Springer. https://doi.org/1031/j.trb.2011.35.6

Marie-Sainte, S. L., & Alalyani, N. (2020). Firefly algorithm based feature selection for Arabic text classification. Journal of King Saud University-Computer and Information Sciences, 32(3), 320-328. https://doi.org/1017/j.trb.2015.38.125

Manikannan, K., & Nagarajan, V. (2020). Optimized mobility management for RPL/6LoWPAN based IoT network architecture using the firefly algorithm. Microprocessors and Microsystems, 77, 103193. https://doi.org/1048/j.trb.2006.23.93

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. TrAC Trends in Analytical Chemistry, 116045. https://doi.org/1013/j.trb.2023.37.122

Pitchaimanickam, B., & Murugaboopathi, G. (2020). A hybrid firefly algorithm with particle swarm optimization for energy efficient optimal cluster head selection in wireless sensor networks. Neural Computing and Applications, 32(12), 7709-7723. https://doi.org/1071/j.trb.2013.7.138

Prasad, K. S., Reddy, N. C. S., & Puneeth, B. N. (2020). A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms. SN Computer Science, 1(2), 1-6. https://doi.org/1030/j.trb.2020.31.125

Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & De Albuquerque, V. H. C. (2020). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. The journal of supercomputing, 76(2), 1128-1143. https://doi.org/1061/j.trb.2022.13.65

Sheridan, K., Puranik, T. G., Mangortey, E., Pinon-Fischer, O. J., Kirby, M., & Mavris, D. N. (2020). An application of dbscan clustering for flight anomaly detection during the approach phase. In AIAA Scitech 2020 Forum (p. 1851). https://doi.org/1033/j.trb.2021.14.99

Tang, S., Yuan, S., & Zhu, Y. (2020). Data preprocessing techniques in convolutional neural network based on fault diagnosis towards rotating machinery. IEEE Access, 8, 149487-149496. https://doi.org/1060/j.trb.2017.29.82

Trachanatzi, D., Rigakis, M., Marinaki, M., & Marinakis, Y. (2020). A Firefly Algorithm for the Environmental Prize-Collecting Vehicle Routing Problem. Swarm and Evolutionary Computation, 100712. https://doi.org/1072/j.trb.2002.30.38

Wang, Y., Gu, Y., & Shun, J. (2020, June). Theoretically-efficient and practical parallel DBSCAN. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 2555-2571). https://doi.org/1092/j.trb.2012.36.31

Zhou, J., Nekouie, A., Arslan, C. A., Pham, B. T., & Hasanipanah, M. (2020). Novel approach for forecasting the blast-induced AOp using a hybrid fuzzy system and firefly algorithm. Engineering with Computers, 36(2), 703-712. https://doi.org/1080/j.trb.2010.32.57