Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2023; 10(3): 238-259
10.34172/jhbmi.2023.29

Optimizing the KNN Algorithm to Diagnose Obstructive Pulmonary Diseases

Pouramirarsalani Shahrzad[1]*, Vahdani manaf Nader[2], Rajebi Saman[3], Makouei Somaye[4]

• **Received:** 27 May 2023 • **Accepted:** 6 Nov 2023

Introduction: According to the World Health Organization, lung diseases are the third cause of death in the world. These diseases are chronic, so early diagnosis of these diseases is very important. Pulmonary function tests are important tools in examining and monitoring patients with respiratory injuries. This research aimed to optimize the K-Nearest Neighbor algorithm, which facilitates and accelerates self-assessment and interpretation of spirometry test results with higher accuracy.
Method: In this study, a method is proposed that improves the limitations of the basic algorithm by optimizing, valuing features, and weighted voting. Using this method, obstructive pulmonary diseases are detected based on the data set of spirometry tests, and general parameters are classified into three categories, namely, asthma, chronic bronchitis, and emphysema.
Results: In determining the appropriate method for calculating the data distance, the Minkowski method was chosen, and by applying the coefficients of the feature values, the accuracy of the classification increased. Weighted voting was done in the final part of the algorithm based on the Gaussian kernel, based on which a constant performance was obtained for changing the parameter of the number of neighbors. The results of the evaluations were carried out in the form of mutual validation. 95.4% accuracy and 93.2% precision were obtained in 3.12 seconds.
Conclusion: The use of machine learning algorithms can be effective in the analysis of medical data. Therefore, in this study, these approaches were used to provide a new method of classification, so that the proposed algorithm could improve the basic method, and also, had better accuracy and performance than other previous methods.

Keywords: Classification, Obstructive Pulmonary Diseases, Fisher's Discriminant Ratio, K Nearest Neighbor, Grasshopper Optimization

• **Citation:** Pouramirarsalani S, Vahdani manaf N, Rajebi S, Makouei S. Optimizing the KNN Algorithm to Diagnose Obstructive Pulmonary Diseases. Journal of Health and Biomedical Informatics 2023; 10(3): 238-59. [In Persian] doi: 10.34172/jhbmi.2023.29

1. Ph.D. Student in Biomedical Engineering, Faculty of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran
2. Ph.D. in Biomedical Engineering, Assistant Professor, Electrical Engineering Faculty, Seraj Higher Education Institute, Tabriz, Iran
3. Ph.D. in Telecommunications Engineering, Assistant Professor, Electrical Engineering Faculty, Seraj Higher Education Institute, Tabriz, Iran
4. Ph.D. in Electronic Engineering, Associate Professor, Faculty of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran
***Corresponding Author:** Shahrzad Pouramirarsalani
Address: Faculty of Electrical and Computer Engineering, Tabriz University, 29 Bahman Blv., Tabriz, East Azerbaijan
• **Tel:** 04133340081 • **Email:** pouramirarsalani@tabrizu.ac.ir

مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره دهم شماره سوم. ص ۲۵۹-۲۳۸، ۱۴۰۲
10.34172/jhbmi.2023.29

بهینه‌سازی الگوریتم KNN در راستای تشخیص بیماری‌های انسدادی ریوی

شهرزاد پورامیرارسلانی[1*]، نادر وحدانی مناف[2]، سامان راجبی[3]، سمیه ماکویی[4]

• دریافت مقاله: ۱۴۰۲/۳/۶ • پذیرش مقاله: ۱۴۰۲/۸/۱۵

مقدمه: به گزارش سازمان بهداشت جهانی، بیماری‌های ریوی سومین علت مرگ و میر در جهان می‌باشند. این بیماری‌ها ماهیت مزمن داشته، بنابراین تشخیص زودهنگام اهمیت بالایی دارد. تست‌های عملکردی ریوی ابزار مهمی در بررسی و پایش بیماران مبتلا به آسیب‌های تنفسی می‌باشند. هدف از این پژوهش بهینه‌سازی الگوریتم پایه K نزدیک‌ترین همسایه می‌باشد که با دقت بالاتری خودارزیابی و تفسیر نتایج تست اسپیرومتری را تسهیل و تسریع می‌کند.

روش: در این پژوهش کاربردی روشی پیشنهاد شده است که محدودیت‌های الگوریتم پایه را با بهینه‌سازی، ارزش‌گذاری ویژگی‌ها و رأی‌گیری وزن‌دار بهبود بخشیده و با به کارگیری آن بیماری‌های انسدادی ریوی را بر اساس مجموعه داده تشکیل یافته از تست‌های تنفس‌سنجی و پارامترهای عمومی، در سه دسته آسم، برونشیت مزمن و آمفیزم کلاس‌بندی کرده است.

نتایج: در تعیین روش مناسب برای محاسبه فاصله داده‌ها، روش مینوکوفسکی انتخاب شد و با اعمال ضرایب ارزش ویژگی‌ها در این رابطه دقت کلاس‌بندی افزایش یافت. رأی‌گیری وزن‌دار در قسمت نهایی الگوریتم بر اساس کرنل گوسی صورت گرفت که بر این اساس عملکرد ثابتی به ازای تغییر پارامتر تعداد همسایگان به دست آمد. نتایج ارزیابی‌ها در قالب اعتبارسنجی متقابل انجام شد که دقت ۹۵/۴ درصد و ۹۳/۲ درصد صحت در زمان ۳/۱۲ ثانیه به دست آمد.

نتیجه‌گیری: بکارگیری الگوریتم‌های یادگیری ماشین می‌تواند در تجزیه و تحلیل داده‌های پزشکی مؤثر واقع گردد؛ لذا در این مطالعه از این رویکردها برای ارائه روشی جدید در کلاس‌بندی، کمک گرفته شد، به طوری که الگوریتم پیشنهادی توانست روش پایه را بهبود ببخشد و همچنین دقت و عملکرد بهتری نسبت به روش‌های پیشین، داشته باشد.

کلیدواژه‌ها: کلاس‌بندی، بیماری‌های انسدادی ریوی، نرخ جداپذیری فیشر، K نزدیک‌ترین همسایه، الگوریتم بهینه‌سازی ملخ

•**ارجاع:** پورامیرارسلانی شهرزاد، وحدانی مناف نادر، راجبی سامان، ماکویی سمیه. بهینه‌سازی الگوریتم KNN در راستای تشخیص بیماری‌های انسدادی ریوی. مجله انفورماتیک سلامت و زیست پزشکی ۱۴۰۲؛ ۱۰(۳): ۵۹-۲۳۸. doi: 10.34172/jhbmi.2023.29

۱. دانشجوی دکتری تخصصی مهندسی پزشکی، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران
۲. دکتری تخصصی مهندسی پزشکی، استادیار گروه مهندسی پزشکی، مؤسسه آموزش عالی سراج، تبریز، ایران
۳. دکتری تخصصی مخابرات، استادیار گروه مهندسی برق، موسسه آموزش عالی سراج، تبریز، ایران
۴. دکتری تخصصی مهندسی برق- الکترونیک، دانشیار دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران
*** نویسنده مسئول:** شهرزاد پورامیرارسلانی
آدرس: آذربایجان شرقی، تبریز، بلوار ۲۹ بهمن، دانشگاه تبریز، دانشکده مهندسی برق و کامپیوتر
•**شماره تماس:** ۰۴۱-۳۳۳۴۰۰۸۱ • **Email:** pouramirarsalani@tabrizu.ac.ir

مقدمه

بیماری‌های ریوی با ۴ میلیون مرگ زودرس سالانه در جهان، یکی از علت‌های اصلی مرگ و میر و ناتوانی در جهان است [۱]. گزارش سازمان بهداشت جهانی پیش بینی می‌کند که تا سال ۲۰۳۰ بیماری‌های مزمن انسدادی ریوی (Chronic Obstructive Pulmonary Disease) COPD به سومین علت مرگ و میر در سراسر جهان تبدیل شود [۲،۳]. کاهش کیفیت هوا با افزایش آلودگی، مردم را در برابر بیماری‌های ریوی آسیب‌پذیرتر می‌کند. بیشتر بیماری‌های ریوی ماهیت مزمن دارند و درمان آن‌ها بار مالی بر دوش بیمار تحمیل می‌کند. بیماری‌های ریوی را می‌توان به دو دسته انسدادی و غیر انسدادی طبقه‌بندی کرد [۴]. برخی بیماری‌های ریوی مانند COPD تا حد زیادی غیرقابل برگشت هستند، بنابراین تشخیص زودهنگام چنین بیماری‌هایی بسیار مهم است. بیماری‌های انسدادی با محدودیت‌های جریان هوا ناشی از تغییرات ساختاری و/یا عملکردی در دیواره یا لومن راه هوایی مشخص می‌شوند [۵]. از بیماری‌های انسدادی ریوی می‌توان آسم، برونشیت و آمفیزم را نام برد [۶].

برای تشخیص بیماری‌های ریوی آزمایش‌های پزشکی مختلفی انجام می‌شود که مهم‌ترین و اصلی‌ترین روش تنفس‌سنجی، تست عملکرد ریوی است [۷]. تست‌های عملکردی ریوی ابزار مهمی در بررسی و پایش بیماران مبتلا به آسیب‌های تنفسی می‌باشند. پارامترهای به دست آمده از روش تنفس‌سنجی، شرح حال، معاینه فیزیکی و سایر نتایج پاراکلینیکی توسط متخصص تفسیر می‌شود. روند تفسیر نتایج به صورت دستی، زمان‌بر و وابسته به فرد متخصص می‌باشد [۸]. پس تجزیه و تحلیل داده‌های به دست آمده از تست‌های عملکرد ریوی به کمک الگوریتم‌های هوش مصنوعی می‌تواند برای پزشکان، محققان و بیماران مفید باشد [۹،۱۰].

کلاس‌بندی، که گاهی اوقات به عنوان شناسایی الگوی نظارت شده از آن یاد می‌شود، فرآیند مرتبط کردن نمونه‌های ناشناخته با یک کلاس نمونه از قبل تعیین‌شده بر اساس الگوی ویژگی‌های مشاهده شده آن‌ها است. به بسیاری از روش‌های تشخیص الگو، مدل‌های یادگیری ماشینی (Machine learning) ML نیز گفته می‌شود. ML زیرمجموعه‌ای از هوش مصنوعی است که معمولاً به عنوان استفاده و توسعه یک سیستم کامپیوتری با قابلیت خودآموزی و سازگاری توصیف می‌شود. ML معمولاً با استفاده از الگوریتم‌ها و مدل‌های ریاضی، با توجه به مسئله تعیین‌شده از الگوهای داده‌ها، ارزیابی و استنباط می‌کند [۱۱].

اتکا به سیستم‌های مراقبت‌های بهداشتی پزشکی مبنی بر فناوری و هوش مصنوعی امروزه بیش از پیش مورد توجه واقع گردیده است. این سیستم‌ها در زمینه‌های تشخیص، ارزیابی خطر و پیش‌بینی کاربرد دارند، بنابراین سیستم‌های مراقبت‌های بهداشتی هوشمند مبتنی بر مدل‌های هوش مصنوعی می‌توانند در زمینه کمک به بیماران مبتلا به بیماری‌های مزمن، تسریع روند تشخیص و خودارزیابی تست‌های پزشکی سودمند باشند [۹].

Swaminathan و همکاران [۱۲] الگوریتم‌های مختلف ML را برای طبقه‌بندی COPD و غیر COPD با استفاده از رگرسیون لجستیک (Logistic regression)، درخت تصمیم، LDA (Linear discriminant analysis)، k-nearest neighbors) KNN، Support Vector Machine) SVM استفاده نمودند. مجموعه داده از اطلاعات جمعیت‌شناختی، بالینی، مرتبط با بیماری‌ها و اندازه‌گیری‌های اسپیرومتری و همچنین پارامترهای صدای تنفسی ۱۰۱ بیمار تشکیل یافته بود. طبقه‌بندی‌کننده KNN دقت ۷۵ درصد و SVM هنگام استفاده از مهم‌ترین پارامترهای صدای ریه، یعنی فرکانس متوسط و پارامترهای پیش‌بینی خطی، به حداکثر دقت طبقه‌بندی ۸۳/۶ درصد دست یافت. Wu و همکاران [۱۳] در یک مطالعه آینده‌نگر از بیماران مبتلا به COPD داده‌های مربوط به سبک زندگی، عوامل محیطی نظیر دما و رطوبت را جمع‌آوری کرده و با این ویژگی‌ها، عملکرد پیش‌بینی مدل‌های یادگیری ماشین از جمله جنگل تصادفی، درخت‌های تصمیم، k-نزدیک‌ترین همسایه، تجزیه و تحلیل تشخیص خطی، تقویت تطبیقی و یک مدل شبکه عصبی عمیق را ارزیابی نمودند و مدل پیش‌بینی نهایی به دقت ۹۲/۱٪، حساسیت ۹۴٪، و تشخیص‌پذیری ۹۰/۴٪ دست یافت. Ioachimescu و همکاران [۱۴] بر اساس مجموعه داده متشکل از ۱۵۳۰۸ نمونه شامل تست اسپیرومتری متوالی، بهترین کارآزمایی قابل قبول، پیش از گشادکننده برونش بود که در آزمایشگاه (Pulmonary function tests) PFTs کلینیک Cleveland انجام شد. یک مدل شبکه عصبی مصنوعی که اندازه‌گیری‌های اسپیرومتری سنتی و پارامترهای به دست آمده از منحنی جریان حجم بازدمی را ترکیب می‌کرد، به خوبی بین کلاس نرمال و اختلالات انسدادی، محدودکننده و مختلط تمایز قائل شد. در

[DOI: 10.34172/jhbmi.2023.29]

مطالعه Haider و همکاران [۱۵] مسئله طبقه‌بندی افراد عادی و COPD بر اساس آنالیز صدای تنفسی با استفاده از تکنیک‌های یادگیری ماشین مورد بررسی قرار گرفت. طبقه‌بندی بر اساس پارامترهای اسپیرومتری و همچنین پارامترهای صدای تنفسی ارزیابی شد. همچنین از طبقه‌بندی‌کننده‌های متمایز خطی و درجه دوم استفاده کردند. طبقه‌بندی کننده تشخیص خطی دقت طبقه‌بندی ۹۴/۴٪ را به دست آورد. Hussain و همکاران [۱۶] مجموعه داده‌ای شامل ۲۹۰۰ بیمار مبتلا به COPD از بیمارستان دانشگاه Paik، بوسان کره را جمع‌آوری کرده‌اند. پنج الگوریتم یادگیری ماشین برای طبقه‌بندی بیماران انسدادی ریوی به کار برده‌اند. نتیجه دقت کلاس‌بندی با ۲۴ ویژگی برای جنگل‌های تصادفی ۸۷/۲۱٪، ماشین بردار پشتیبان ۸۸/۱۸٪، ماشین تقویت‌کننده گرادیان ۹۰/۲۲٪، XGboost ۸۸/۰۷٪ و الگوریتم KNN ۸۶/۳۵٪ به دست آمد. Spathis و همکاران [۱۷] مجموعه داده را از کلینیکی در یونان با ۱۳۲ نمونه و متشکل از ۲۲ ویژگی مانند سیگار کشیدن، سن، حجم یک بازدم اجباری، نبض، سرفه و تنگی نفس برای تصمیم‌گیری بالینی برای تشخیص بیماران COPD و آسم جمع‌آوری کردند. نتایج یادگیری ماشینی نشان می‌دهد که در مورد بیماری مزمن انسدادی ریه، طبقه‌بندی‌کننده Random Forest با دقت ۹۷/۷ درصد از سایر تکنیک‌ها بهتر عمل می‌کند، در حالی که برجسته‌ترین ویژگی‌ها برای تشخیص سیگار کشیدن، حجم بازدم اجباری در یک ثانیه، سن و ظرفیت حیاتی اجباری است. در مورد آسم، بهترین دقت، ۸۰/۳ درصد، دوباره با طبقه‌بندی جنگل تصادفی به دست می‌آید، در حالی که برجسته‌ترین ویژگی MEF25(Maximal Expiratory Flow) است. Siddiqui و همکاران [۱۸]، داده‌های به دست آمده از سیگنال تنفسی و ترکیب ویژگی‌های اضافی مانند سن، جنسیت و سابقه مصرف سیگار برای تشخیص COPD را به کار بردند. طبقه‌بندی‌کننده‌های مختلف یادگیری ماشین، از جمله Naïve Baye، ماشین بردار پشتیبان، جنگل تصادفی، k نزدیک‌ترین همسایه (KNN)، Adaboost و شبکه عصبی برای تشخیص استفاده شده است. نتایج تجربی نشان می‌دهد که KNN دقت ۸۶ درصد را به دست آورد. LSTM از همه مدل‌های به کار گرفته شده بهتر عمل کرده و دقت ۹۳ درصد را به دست آورد. Bhattacharjee و همکاران [۱۹] بر اساس ۱۲ ویژگی اسپیرومتری ۱۱۶۳بیمار، شبکه عصبی پرسپترون چند لایه (Multilayer perceptron) MLP با استفاده از اعتبارسنجی متقاطع ۵ برابری (Cross-validation) آموزش دادند که با دقت ۸۳/۷ درصد بیماری‌های انسدادی و غیرانسدادی ریه پیش‌بینی شد. Raghavan و همکاران [۲۰] تجزیه و تحلیل رگرسیون لجستیک برای شناسایی متغیرهای مربوط به وجود انسداد راه هوایی را به کار بردند. مجموعه داده از عوامل شناخته شده COPD مانند سابقه سیگار کشیدن، سن، اسپیرومتری پس از گشادکننده برونش و اطلاعات دموگرافیک و فیزیولوژیک پایه کلیه شرکت‌کنندگان تشکیل دادند. مدل رگرسیون لجستیک دقت متوسطی را به صورت امتیاز AUC (Area under the Curve) ۷۷٪ به دست آورد. به طور خلاصه، سه گانه سابقه سیگار کشیدن، سن حداقل ۵۵ سال و وجود تنگی نفس در هنگام فعالیت، عناصر کلیدی مدل پیشنهادی در شناسایی بیماران در معرض خطر COPD که آزمایش اسپیرومتری برای آن‌ها توصیه می‌شود، بودند.

Vora و Shah [۲۱] با داده‌های جمع‌آوری شده از بیمارستان پرچنا هند، الگوریتم SVM و KNN را برای بررسی سطح بیماری COPD پیشنهاد کردند که KNN دقت ۹۰/۳۰ درصد را به دست آورد. Tarakci و Ozkan [۲۲] روش پیشنهادی W-KNN به هر یک از نقاط داده آموزشی وزنی اختصاص می‌دهد. وزن‌دهی با در نظر گرفتن مربع معکوس فاصله ساخته می‌شود. هدف این تخصیص وزن، دادن وزن بیشتر به نقاط نزدیک‌تر و وزن کمتر به نقاط دور بود، بنابراین پس از انتخاب عدد همسایه k مناسب و تابع فاصله، استفاده از kNN وزنی (WKNN) که وزن شاخص ویژگی را در نظر می‌گیرد، به بهبود عملکرد طبقه‌بندی کمک می‌کند. الگوریتم WKNN با سه مجموعه داده اجرا شده و باتوجه به نتایج، طبقه‌بندی موفق‌تری، نسبت به KNN انجام داد.

Mullick و همکاران [۲۳] گونه‌ای از kNN به نام kNN تطبیقی (Ada-kNN) را پیشنهاد شده است. طبقه‌بندی‌کننده Ada-kNN از چگالی و توزیع همسایگی یک نقطه آزمایشی استفاده می‌کند و با کمک شبکه‌های عصبی مصنوعی یک k مخصوص نقطه‌ای مناسب برای آن می‌آموزد. Kumbure و همکاران [۲۴] یک نسخه تعمیم‌یافته جدید از طبقه‌بندی‌کننده k نزدیک‌ترین همسایه فازی (FKNN) ارائه کردند که از بردارهای میانگین محلی و میانگین بونفرونی استفاده می‌کند. روش جدید پیشنهادی را طبقه‌بندی‌کننده K-نزدیک‌ترین همسایه فازی مبتنی بر میانگین بونفرونی (BM-FKNN) نامیدند که در موقعیت‌هایی که عدم تعادل واضح در توزیع

[DOI: 10.34172/jhbmi.2023.29]

کلاسی داده‌ها وجود داشته باشد، عملکرد خوبی دارد. عملکرد طبقه‌بندی‌کننده پیشنهادی با شش مجموعه داده دنیای واقعی و با یک مجموعه داده مصنوعی آزمایش شد. نتایج با نتایج طبقه‌بندی به‌دست‌آمده با k–نزدیک‌ترین همسایه کلاسیک، محک زده می‌شوند. نتایج نشان می‌دهد که طبقه‌بندی‌کننده جدید BM-FKNN پیشنهاد شده این پتانسیل را دارد که از معیارها در دقت طبقه‌بندی بهتر عمل کند.

هدف از این پژوهش بهبود عملکرد و رفع محدودیت‌های الگوریتم پایه K نزدیک‌ترین همسایه KNN می‌باشد که رویکردهای مبتنی بر یادگیری ماشین را در تفسیر نتایج تست اسپیرومتری به کار برده و مدلی که بتوانند با دقت و عملکرد مطلوبی بیماری‌های انسدادی ریوی را کلاس‌بندی کند ارائه کند.

روش

روش پژوهش این مطالعه کاربردی بوده و پایه اصلی آن بر اساس داده‌کاوی می‌باشد که با پیشنهاد مدلی مبتنی بر بهینه سازی الگوریتم K همسایه نزدیک به کلاس‌بندی بیماری‌های انسدادی ریوی می‌پردازد. داده‌ها از مراکز خصوصی طب تخصصی ریه جمع‌آوری گردید. مدل‌سازی الگوریتم و ارزیابی‌ها در محیط نرم افزار MATLAB 2021b انجام شد.

مشخصات مجموعه داده

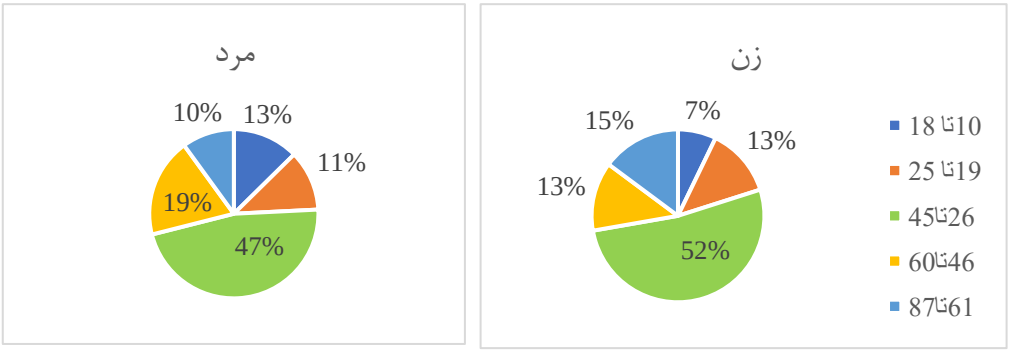
مجموعه داده شامل اطلاعات به دست آمده از روش اسپیرومتری و ویژگی‌های مرتبط با بیماری‌های ریوی، ۴۳۱ فرد می‌باشد. مجموعه داده جمع‌آوری شده از ۱۹۰ مرد و ۲۴۱ زن در محدوده سنی ۱۰ تا ۷۰ سال، به سه کلاس بیماری‌های انسدادی شامل آسم، بروشیت مزمن، آمفیزم تعلق دارد. جدول ۱ مرتبط با ویژگی‌های موجود در دادگان می‌باشد. توزیع نمونه‌های مجموعه داده در هرکلاس به صورت متعادل است. به طوری که ۳۳/۳۳ درصد به بیماری آسم و آمفیزم و ۳۴/۳۴ درصد به برونشیت مزمن تعلق دارد.

جدول ۱: ویژگی‌های موجود در مجموعه داده

ویژگی	مخفف	توصیف
Forced expiratory volume in 1 second	FEV1	حجم بازدم اجباری در یک ثانیه
Forced vital capacity	FVC	ظرفیت حیاتی اجباری
Forced expiratory volume /Forced vital capacity	FEV1/FVC	نسبت حجم بازدم اجباری در یک ثانیه بر ظرفیت حیاتی اجباری
mean expiratory flow	MEF 25	جریان بازدمی اجباری در نیمی از وسط FVC
Respiratory Rate	RR	نرخ تنفس تعداد دم و بازدم درد دقیقه
Peak expiratory flow	PEF	حداکثر سرعت جریان هوا هنگام بازدم (اوج جریان بازدمی)
Total lung capacity	TLC	ظرفیت کل ریه
Residual capacity	RLC	ظرفیت باقی مانده ریه
Age	–	سن
Gender	–	جنسیت
Smoking	–	سابقه مصرف سیگار
Drinking	–	سابقه مصرف الکل
Body-Mass-Index	BMI	شاخص توده بدنی
Severity	–	درجه شدت بیماری شامل: خفیف، متوسط، شدید
Disease	–	برچسب‌ها شامل: آسم، برونشیت و آمفیزم

شکل ۱ توزیع بیماری در بین گروه‌های سنی به تفکیک جنسیت را نشان می‌دهد. گروه سنی ۲۶ تا ۴۵ سال بیشترین میزان درگیری با بیماری‌های انسدادی ریوی را دارد.

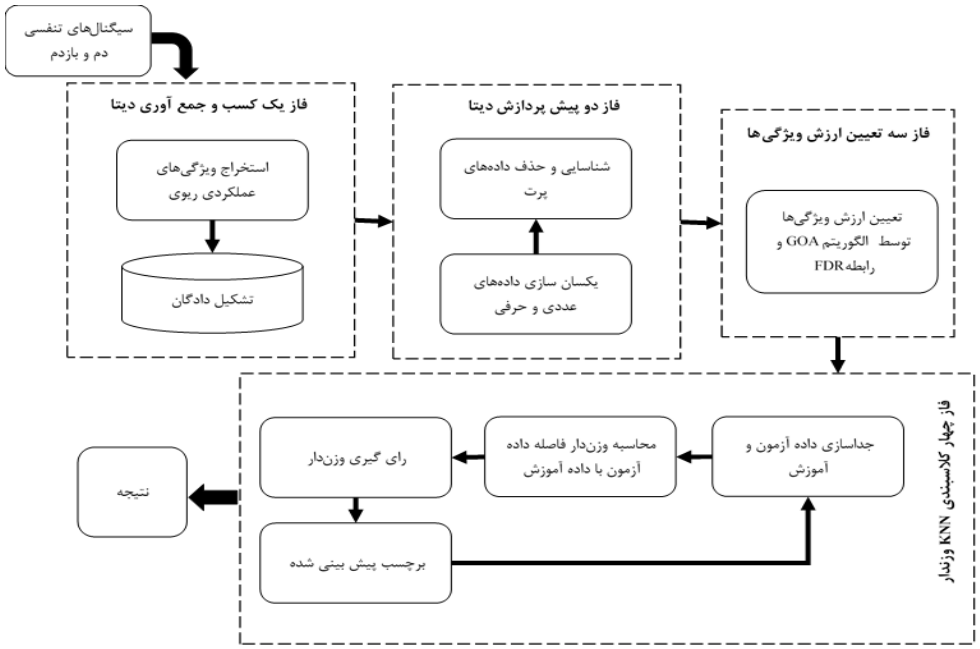
[DOI: 10.34172/jhbmi.2023.29]



شکل ۱: توزیع بیماری در گروه‌های سنی به تفکیک جنسیت

روش پیشنهادی

روش به کار رفته در این مطالعه بر پایه الگوریتم K همسایه نزدیک در چهار بخش به کلاس‌بندی بیماری انسدادی ریوی می‌پردازد. طبق فلوچارت شکل ۲ در فاز اول داده‌ها کسب و جمع‌آوری می‌شوند. فاز دوم داده‌ها پیش‌پردازش می‌شوند. در فاز سوم ارزش ویژگی‌های موجود در دادگان تعیین شده و در فاز چهارم کلاس‌بندی با روش پیشنهادی KNN وزن‌دار انجام می‌گردد.



شکل ۲: فلوچارت روش پیشنهادی

فاز یک کسب و جمع‌آوری داده

ویژگی‌های عملکردی ریوی توسط تست اسپیرومتری از سیگنال‌های تنفسی ثبت می‌شود. این ویژگی‌ها توسط دو منحنی حجم زمان و حجم جریان به دست می‌آید. ویژگی‌های سن، جنسیت، شاخص توده بدنی، سابقه مصرف الکل و سیگار نیز از هر بیمار ثبت می‌شود.

فاز دو پیش پردازش مجموعه داده

پیش‌پردازش‌هایی که بر روی دادگان صورت می‌گیرد شامل یکسان‌سازی داده‌های حرفی و عددی و حذف داده‌های پرت می‌باشد. اکثر الگوریتم‌های داده‌کاوی، داده‌های عددی را به عنوان ورودی دریافت می‌کنند و ساختار یادگیری آن‌ها بر اساس یادگیری از ماتریس‌های عددی است؛ بنابراین برای یکسان‌سازی داده‌های حرفی و عددی، قبل از ورود مجموعه داده به الگوریتم کلاس‌بند، ویژگی‌های موجود در دادگان که به

[DOI: 10.34172/jhbmi.2023.29]

صورت متغیرهای اسمی هستند به صورت عددی تبدیل شوند. داده‌های پرت، مقادیری هستند که خارج از محدوده طبیعی یک متغیر قرار دارند. برای شناسایی و حذف داده‌های پرت در این پژوهش از روش چارک و نمودار جعبه‌ای استفاده شد. در این روش باید چارک اول و سوم را محاسبه کرد و سپس با استفاده از آن حد بالا (U) و پایین (L)، مطابق فرمول‌های ۱ تا ۳ محاسبه نمود. سپس محدوده بین ربعی (که ۵۰ درصد داده‌ها در چارک اول و سوم است) محاسبه می‌شود در اصل (Inter Quartile Range) IQR همانند فرمول ۱، تفاوت بین چارک سوم و چارک اول است. داده‌هایی پرت هستند که در بین حد بالا و پایین نباشند به عبارتی کوچک‌تر از L و بزرگ‌تر از U باشند [۲۵].

$$IQR=Q3-Q1 \quad (۱)$$

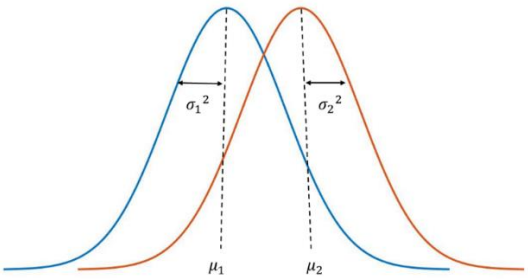
$$L=Q1-1.5*(IQR) \quad (۲)$$

$$U=Q3+1.5*(IQR) \quad (۳)$$

فاز سه تعیین ارزش ویژگی‌ها

- نسبت جداسازی فیشر

نسبت جداسازی فیشر (Fisher Discriminant Ratio) FDR روشی آماری است که در مسائل یادگیری ماشین و تشخیص الگو کاربرد دارد. FDR در مسائل کاهش ابعاد داده، انتخاب ویژگی، افزایش جداسازی، تمایز کلاس‌های مختلف داده و یافتن بهترین تقریب خطی از بردارهای ویژگی داده به کار می‌رود. اگر داده‌های مربوط به دو کلاس به طور شکل گوسی فرض شود. به دلیل وجود داده‌های مشابه در بین اعضای دو کلاس ناحیه مشترکی مابین توزیع گوسی کلاس‌ها به وجود می‌آید که در شکل ۳ قابل مشاهده است. وجود این ناحیه مسئله تفکیک و تشخیص دو کلاس را از یکدیگر دچار مشکل می‌کند.



شکل ۳: نمودار گوسی داده

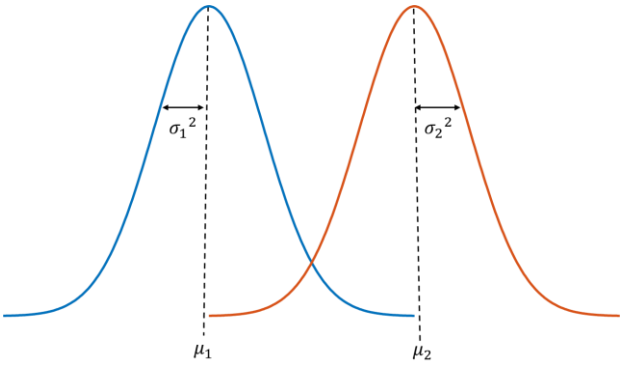
هرچه این ناحیه مشترک کوچک‌تر باشد، داده‌های دو کلاس بیشتر قابل تفکیک خواهد بود. برای کمتر شدن اثر این ناحیه و تمایز بیشتر کلاس‌ها از هم از نسبت جداسازی فیشر FDR استفاده می‌شود. FDR برای مسئله دو کلاسه به صورت معادله ۴ تعریف می‌شود که μ میانگین و σ واریانس هر کلاس می‌باشد.

$$FDR=\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} \quad (۴)$$

باتوجه به مفهوم پارامترهای توابع گوسی، هرچه اختلاف میانگین کلاس‌ها از هم بیشتر بوده و واریانس‌های دو کلاس کوچک‌تر باشد، ناحیه مشترک کمتر می‌شود. در نتیجه برای جداسازی دو کلاس از یکدیگر مقدار FDR باید ماکزیمم شود. مقدار به دست

[DOI: 10.34172/jhbmi.2023.29]

آمده از رابطه ۴، یک عدد خواهد بود که نرخ جداپذیری فیشر نامیده می‌شود و می‌توان آن را در مقام مقایسه، ارزیابی نمود. به‌طوری که هر چه مقدار آن بیشتر شود، کلاس‌بندی دو کلاس از هم ساده‌تر شده و نرخ تشخیص صحیح کلاس‌بندی بهبود خواهد یافت. می‌توان با تکنیک‌هایی نحوه توزیع داده‌ها را با افزایش FDR بهینه کرد. رسم توزیع گوسی داده‌های مربوط به دو کلاس، بعد از بهینه‌سازی مطابق با شکل ۴ خواهد بود.



شکل ٤: توزیع گوسی داده بعد بهینه‌سازی

به دلیل وجود ویژگی‌های متعدد و چندکلاسه بودن مسئله مورد مطالعه، FDR برای تمام ویژگی‌ها محاسبه می‌شود؛ بنابراین FDR برای مسائل چندکلاسه به صورت زیر تعریف می‌شود:

$$Multi\ Class\ FDR = \frac{\sum_{j=1}^{L}(\mu_j - \mu_T)^2}{\sum_{j=1}^{L}\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ij} - \mu_j)^2} \quad (۵)$$

که در معادله ۵، L تعداد کلاس‌ها، μj میانگین هر کلاس، μT میانگین همه کلاس‌ها، nj تعداد عضوهای هر کلاس و x عناصر هر کلاس می‌باشد. در این مرحله ارزش ویژگی‌های موجود در دادگان توسط الگوریتم بهینه‌سازی ملخ و نسبت جداسازی فیشر به دست آمد. باید برای هر ویژگی یک وزن در جهت افزایش دقت کلاس‌بندی اختصاص داده شود. بدین صورت که در ابتدا ذره‌ها یا همان ملخ‌ها در جهت یافتن وزن‌های تصادفی برابر با تعداد ویژگی‌ها تلاش می‌کنند. پارامترهای

$$a = [a_1, a_2, \dots, a_n] \quad (۶)$$

ضرائب به دست آمده از الگوریتم بهینه‌سازی ملخ به صورت بردار ۶ می‌باشد. این ضرائب متناسب با ویژگی‌های موجود در داده می‌باشد و به صورت یک ترکیب خطی، ارزش ویژگی‌ها را

الگوریتم ملخ از قبیل مکان ذره و سرعت ذره و همچنین بهترین ذره در هر تکرار به روزرسانی می‌شوند. این روند تا پیدا شدن بهترین وزن برای ویژگی‌ها که منجر به افزایش تفکیک‌پذیری کلاس‌های مختلف موجود در مجموعه داده شود، ادامه می‌یابد. تابع برازندگی برای الگوریتم ملخ بر اساس معادله ۴ به کار رفته در نسبت جداسازی فیشر تعریف می‌شود. برای بررسی اثر FDR می‌توان با ضرب کردن ضرایب مجهول در ویژگی‌ها مقادیر μ و σ را تغییر داد.

در کنار هم نشان می‌دهد که در صورت و مخرج معادله ۴ ضرب می‌شود. معادله ۵ به فرم ماتریسی به صورت زیر بازنویسی می‌شود و بردار a در صورت و مخرج رابطه ضرب می‌شود.

[DOI: 10.34172/jhbmi.2023.29]

$$Multi\ Class\ FDR = \frac{\sum_{j=1}^{L} a^T(\mu_j - \mu_T) \times (\mu_j - \mu_T)^T a}{\sum_{j=1}^{L} a^T \left(\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ij} - \mu_j)(x_{ij} - \mu_j)^T\right) a} \quad (۷)$$

در هر تکرار از الگوریتم ملخ، ضرایب a در فرمول FDR اعمال شده و نتیجه به صورت یک عدد به دست می‌آید که نرخ جداسازی فیشر نامیده می‌شود. ماکزیمم‌ترین عددی که براساس روابط موجود در تابع برازندگی به دست بیاید، باعث پایان الگوریتم بهینه‌سازی خواهد شد. ضرایبی که باعث ماکزیمم شدن FDR شوند جواب نهایی مسئله می‌باشند. این ضرایب به عنوان بهترین ترکیب ارزش ویژگی‌ها در مراحل بعدی روش پیشنهادی به کار خواهند رفت.

الگوریتم بهینه‌سازی ملخ

الگوریتم بهینه‌سازی ملخ از رفتار دسته جمعی ملخ‌ها در طبیعت الهام گرفته شده است [۲۶]. این الگوریتم مبتنی بر جمعیت است. دو مرحله اساسی بهینه‌سازی الگوریتم عبارت‌اند: از اکتشاف و بهره‌برداری از فضای جستجو، که ملخ این دو مرحله را در طول جستجوی غذا از طریق تعاملات اجتماعی فراهم می‌کند [۲۷]. مدل ریاضی به کار گرفته شده برای شبیه‌سازی رفتار ملخ‌ها در ابتدا به شکل زیر بوده است:

$$X_i = S_i + G_i + A_i \qquad i = 1,2,\ldots,N \quad (۸)$$

که در آن Xi موقعیت ملخ i ام، S_i تعامل اجتماعی، Gi نیروی گرانش اعمال شده به ملخ i ام و Ai جهت باد می‌باشد. مقدار Si یعنی تعامل اجتماعی برای ملخ iام با توجه به رابطه زیر محاسبه می‌شود. که در آن dij فاصله بین ملخ i ام با ملخ jام می‌باشد.

$$\boldsymbol{S_i = \sum_{j=1}^{N} s(d_{ij})\widehat{d_{ij}}} \quad (۹)$$

s یک تابع برای تعریف فشار نیروی اجتماعی می‌باشد همان‌طور که در رابطه ۹ نشان داده شده است و $\widehat{d_{ij}}$ یک بردار واحد از i امین ملخ به j امین ملخ می‌باشد. تابع S، که نیروی اجتماعی را تعریف می‌کند همانند زیر محاسبه می‌شود:

$$\boldsymbol{s(r) = fe^{\frac{-r}{l}} - e^{-r}} \quad (۱۰)$$

که در آن f نشان دهنده شدت جاذبه و I نشان دهنده طول مقیاس جاذبه می‌باشد. تابع S چگونگی تأثیر بر روی تعامل اجتماعی (جاذبه و دافعه) ملخ‌ها را نشان دهد [۲۶].

فاز چهار کلاس‌بندی KNN وزن‌دار

در این فاز الگوریتمی برای تشکیل یک مدل کلاس‌بند پیشنهاد می‌گردد که با این الگوریتم بیماری‌های انسدادی ریوی تشخیص داده خواهد شد. مراحل الگوریتم به ترتیب زیر تا زمانی که تمام داده‌های موجود در پایگاه داده در هر دو بخش آموزش و آزمون به کار گرفته شوند، تکرار می‌گردد.

الگوریتم k-نزدیک‌ترین همسایه وزن‌دار

الگوریتم K همســـایه نزدیک (KNN) جزء الگوریتم‌های یادگیری نظارت شده است که در مسائل کلاس‌بندی استفاده می‌شود. این الگوریتم به دلیل سادگی و اثربخشی آن شناخته شده است. اگرچه این طبقه‌بندی‌کننده ساده است، اما KNN برای مطالعه‌ای که در آن هیچ دانش قبلی در مورد داده‌های مورد استفاده وجود ندارد، کاربرد دارد [۲۷].

کلاس‌بندی داده‌ها بر اساس نمونه‌های آموزشی نزدیک یا همسایه در یک منطقه خاص انجام می‌شود. برای یک ورودی جدید، K نزدیک‌ترین همسایگان محاسبه می‌شوند و اکثریت در میان داده‌های همسایه، کلاس ورودی جدید را تعیین می‌کنند

[DOI: 10.34172/jhbmi.2023.29]

[۲۸]. جهت تعیین نزدیک‌ترین همسایگان باید فاصله داده آزمایش از مجموعه داده آموزشی محاسبه شود روش‌های محاسبه فاصله و فرمول آن‌ها [۲۹] به شرح جدول ۲ می‌باشد.

جدول ۲: روش‌های محاسبه فاصله

ردیف	نام روش	فرمول
۱	فاصله اقلیدسی	$dis(A,B) = \sqrt{(A-B)^2}$
۲	فاصله بلوک شهری	$dis(A,B) = \sum\lvert A-B\rvert$
۳	فاصله چبیشف	$dis = max_j\{\lvert A-B\rvert\}$
۴	فاصله مینکوفسکی	$dis(A,B) = \sqrt[p]{\sum\lvert A-B\rvert^p}$
۵	فاصله کسینوسی	$dis = 1 - \frac{A'B}{\sqrt{(A'A)*(B'B)}}$
۶	فاصله همبستگی	$\rho(A,B) = \frac{cov(A,B)}{\sigma_A\sigma_B}$ $dis(A,B) = 1-\rho(A,B)$

الگوریتم k–نزدیک‌ترین همسایه وزن‌دار یک نسخه اصلاح شده از الگوریتم KNN است. این کلاس‌بند با اضافه شدن مقادیری با عنوان وزن به الگوریتم KNN به دست می‌آید [۳۰]. در این روش فاصله همسایه‌های نزدیک از یک کرنلی عبور می‌کنند تا به وزن تبدیل شوند و رأی‌گیری به صورت وزن‌دار انجام شده و نمونه آزمایش بر این اساس دسته‌بندی می‌شود [۳۰]. رویکردهای مختلفی جهت تبدیل فاصله به شباهت وجود دارد. توسط کرنل‌های موجود در جدول ۳ وزن همسایه‌های نزدیک محاسبه می‌شود.

جدول ۳: کرنل‌های محاسبه شباهت

شماره	فرمول
۱	$\frac{1}{d(x_i,x)}$
۲	$\frac{1}{d(x_i,x)^2}$
۳	$\frac{1}{c+d(x_i,x)^2}$
۴	$exp\left(-\frac{d(x_i,x)^2}{\sigma^2}\right)$
۵	$\frac{1}{\sqrt{2*pi}} * exp\left(-\frac{d(x_i,x)^2}{\sigma^2}\right)$
۶	$\frac{1}{d(x_i,x)^{\frac{2}{m-1}}}$

میانگین موزون

میانگین موزون، میانگین حسابی یک مجموعه داده‌های نابرابر و ناموزون می‌باشد. میانگین ساده یا حسابی زمانی به کار می‌رود که مشاهدات (داده‌های آماری) دارای اهمیت مساوی باشند؛ اما در محاسبه میانگین یک مجموعه مشاهده‌های نابرابر، برای هر یک از عامل‌ها، وزن یا ارزش معینی طبق

[DOI: 10.34172/jhbmi.2023.29]

معادله ۱۱ به صورت Wi در نظر گرفته می‌شود و سپس آن عامل Xi در وزن معین ضرب می‌گردد. آن‌گاه جمع این ارقام به‌دست آمده بر مجموع وزن‌ها تقسیم می‌شود [۳۱].

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad (۱۱)$$

جداسازی داده آزمون و آموزش

یکی از مراحل اولیه در مسائل کلاس‌بندی جداسازی داده‌ها به دو بخش آزمون و آموزش می‌باشد. در کلاس‌بند KNN باید میزان شباهت داده‌های آزمون با کل دادگان آموزش محاسبه شود. انواع روش‌های مختلفی برای تقسیم‌بندی داده‌ها وجود دارد که در این مقاله ۷۰ درصد داده‌ها به عنوان داده آموزش و ۳۰ درصد باقی‌مانده به عنوان داده آزمون در نظر گرفته شد. این تقسیم بندی داده در چهارچوب اعتبارسنجی متقابل ۵ گانه به صورت رندوم انتخاب شد و نتیجه ارزیابی‌ها به صورت میانگین محاسبه گردید. به دلیل تعداد محدود داده‌ها، نتیجه نهایی معیارهای ارزیابی در قالب اعتبارسنجی متقابل Leave one out محاسبه شد.

محاسبه وزن‌دار فاصله

در الگوریتم KNN برای دسته‌بندی از همه ویژگی‌ها به یک اندازه استفاده می‌شود. درصورتی‌که همه ویژگی‌ها ممکن است در دسته‌بندی نقش یکسان نداشته باشند و دسته‌بندی به درستی انجام نگیرد به منظور رفع این مشکل، باید با توجه به درجه اهمیت ویژگی‌ها، فاصله نمونه‌ها از هم به صورت وزن‌دار محاسبه شود.

به این منظور باید برای هر ویژگی یک وزن تعریف شود. هر چه قدر وزن یک ویژگی بزرگ‌تر باشد، میزان تأثیر آن در محاسبه فاصله بیشتر است. اگر برای n همسایه موجود در پایگاه داده‌ها یک بردار از وزن‌ها به صورت a تعریف شود. آن‌گاه با تلفیق بردار a و فرمول محاسبه فاصله بین دو نمونه A و B رابطه محاسبه معیار شباهت وزن‌دار طبق رابطه ۹ به دست خواهد آمد.

$$A = [A_1, A_2, \dots, A_n]$$
$$B = [B_1, B_2, \dots, B_n]$$
$$\boldsymbol{dis(A,B)} = \sqrt[p]{\sum_{i=1}^{n} (\boldsymbol{a} * |\boldsymbol{A} - \boldsymbol{B}|)^p} \quad (۱۲)$$

بردار وزن‌ها ترکیبی خطی از ارزش تمام ویژگی‌ها می‌باشد. هر چه این وزن‌ها دقیق‌تر تعیین شود، موجب افزایش دقت کلاس‌بندی می‌شود. برای تولید بهترین بردار وزن از الگوریتم بهینه‌سازی ملخ استفاده شد.

رأی‌گیری وزن‌دار

پس از محاسبه وزن‌دار فواصل و یافتن k-نزدیک‌ترین همسایه‌ها، باید از یک الگوریتم رأی‌گیری برای تعیین کلاس پیش‌بینی شده استفاده شود. با استفاده از برچسب‌های داده‌های آموزش، تعداد فراوانی هر کلاس در K همسایه نزدیک شمرده می‌شود. در الگوریتم KNN پایه کلاسی که بیشتر تعداد عضو را داشته باشد به عنوان خروجی معرفی می‌شود؛ اما استفاده از رویکرد قانون اکثریت در رأی‌گیری، مشکلات مهمی دارد. در شمارش فراوانی برچسب‌ها اگر دو کلاس تعداد عضو برابری داشته باشند تصمیم‌گیری به‌صورت شانسی خواهد بود. امکان دارد که همسایه‌های دورتری که مربوط به کلاس مخالف هستند به علت تعداد بیشتر موجب تشخیص اشتباه شود. برای غلبه بر مشکلات مطرح شده، رأی‌گیری به صورت وزن‌دار پیشنهاد می‌شود.

فواصلی که در بخش اول الگوریتم KNN محاسبه شده و نزدیک‌ترین همسایه‌ها بر اساس آن‌ها انتخاب می‌شود، به معیارهای شباهت تبدیل شده و به عنوان وزن مورد استفاده قرار می‌گیرد. وزن‌های اختصاص داده شده تعیین می‌کنند که هر همسایه چقدر بر عملیات طبقه‌بندی تأثیر می‌گذارد. رویکردهای مختلفی جهت تبدیل فاصله به شباهت وجود دارد. با استفاده از

[DOI: 10.34172/jhbmi.2023.29]

کرنل‌های جدول ۳ وزن همسایه‌های نزدیک محاسبه شده و در قسمت رأی‌گیری اعمال می‌شوند؛ بنابراین در پیش‌بینی برچسب داده ورودی، علاوه بر تعداد همسایگان، میزان شباهت

$$\hat{y} = Max\left(\frac{\sum_{i=1}^{k} W_i \; I(y_i = r)}{\sum_{i=1}^{k} W_i}\right) \quad (۱۳)$$

که در رابطه Wi وزن‌های K همسایه نزدیک، y برچسب داده‌های آموزش، r برچسب کلاس‌ها (نوع بیماری‌ها)، و $\hat{y}$ برچسب خروجی پیش بینی شده می‌باشد. در رابطه ۱۳ برای تعیین برچسب خروجی از میانگین موزون وزن‌ها استفاده می‌شود. برچسب خروجی برای تمام داده‌های آزمایش بر اساس مراحل گفته شده به دست می‌آید.

روش‌های ارزیابی

روش‌های ارزیابی به کار رفته در این مطالعه شامل ماتریس درهم ریختگی و نمودار (Receiver Operation Characteristic) ROC می‌باشد. به منظور تعیین کیفیت و ارزیابی روش پیشنهادی از ماتریس درهم ریختگی استفاده شد. این ماتریس براساس پارامترهای مثبت صحیح (True Positive) TP، منفی صحیح (True Negative) TN، مثبت کاذب (False Positive) FP و منفی کاذب (False Negative) FN تعریف شده طبق مسئله مورد مطالعه تشکیل می‌شود.

در طبقه‌بندی چند کلاسه تمامی پارامترهای ارزیابی برای هر کلاس به صورت جداگانه محاسبه شد. در نهایت مجموع نتایج

$$TP=TP_A + TP_B + TP_C \quad (۱۴)$$

$$TN=TN_A + TN_B + TN_C \quad (۱۵)$$

$$FP=FP_A + FP_B + FP_C \quad (۱۶)$$

$$FN=FN_A + FN_B + FN_C \quad (۱۷)$$

معیارهای ارزیابی دقت (Accuracy)، صحت (precision)، حساسیت (Sensitivity)، تشخیص‌پذیری (Specificity) طبق فرمول‌های ۱۸ تا ۲۱ محاسبه می‌شود. فرآیند ارزیابی بر اساس اعتبارسنجی متقابل Leave-One-Out می‌باشدکه در هر بار یک نمونه از دادگان برای آزمون کنار گذاشته شده و بقیه برای آموزش استفاده می‌شود. این کار تا زمان به کارگیری تمام دادگان در هر دو مرحله آموزش و آزمون ادامه می‌یابد.

به داده آزمایش نیز دخیل می‌شود. برچسب داده ورودی طبق رابطه زیر محاسبه می‌شود.

به دست آمده برای هر پارامتر گزارش شد. در پژوهش حاضر که یک مسئله سه کلاسه است. اگر سه دسته بیماری‌های آسم، برونشیت مزمن و آمفیزم، به صورت برچسب‌های A، B و C در نظر گرفته شود، پارامترهای مرتبط با معیارهای ارزیابی به عنوان نمونه برای برچسب A به صورت زیر محاسبه می‌شود:

TP_A: مواردی با برچسب A که به درستی در کلاس A قرار گرفته‌اند.

TN_A: مواردی بدون برچسب A که در کلاس B یا C قرار گرفته‌اند.

FP_A: مواردی بدون برچسب A که به اشتباه در کلاس A قرار گرفته اند.

FN_A: مواردی با برچسب A که به اشتباه در کلاس B یا C قرار گرفته‌اند.

محاسبات برای برچسب‌های B و C به همین منوال تکرار خواهد شد و مجموع نتایج دسته‌ها به عنوان پارامتر مربوطه حاصل خواهد شد. پارامترهای نهایی طبق فرمول‌های ۱۴ تا ۱۷ می‌باشد.

[DOI: 10.34172/jhbmi.2023.29]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱۸)$$

$$precision = \frac{TP}{TP + FP} \quad (۱۹)$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \quad (۲۰)$$

$$specificity = \frac{TN}{TN + FP} \quad (۲۱)$$

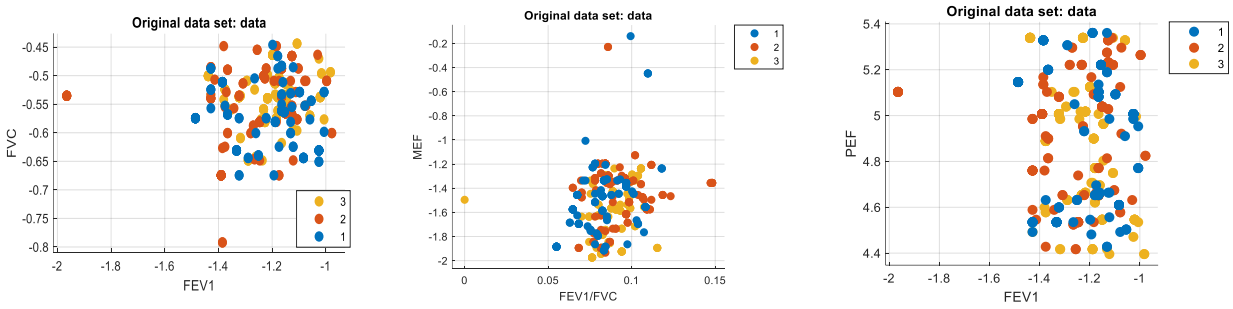
نتایج

در این بخش نتایج و ارزیابی‌هایی بر اساس مراحل ذکر شده انجام شد. نتایج شبیه‌سازی و مقایسه الگوریتم ارائه شده با الگوریتم‌های پایه بررسی گردید. جدول ۴ خلاصه‌ای از اطلاعات آماری ویژگی‌های موجود در مجموعه داده را شرح می‌دهد.

جدول ٤: اطلاعات آماری ویژگی‌های دادگان

	مینیمم			ماکزیمم			میانگین			انحراف معیار		
کلاس	۱	۲	۳	۱	۲	۳	۱	۲	۳	۱	۲	۳
حجم بازدم اجباری	۲/۱۰	۲/۰۴	۲/۰۵	۳/۱۰	۴/۱۰	۳	۲/۵۵	۲/۶۵	۲/۵۵	۰/۲۹	۰/۴۱	۰/۲۳
ظرفیت حیاتی اجباری	۲/۰۵	۲/۰۶	۲/۰۴	۳/۱۰	۳/۶۴	۲/۹۸	۲/۶۳	۲/۵۴	۲/۴۵	۰/۲۴	۰/۳۱	۰/۱۹
جریان بازدم اجباری ۲۵	۰/۱۴	۰/۲۳	۱/۲۴	۱/۹۲	۱/۹۴	۱/۹۸	۱/۵۱	۱/۴۳	۱/۵۴	۰/۲۸	۰/۲۶	۰/۲۳
ظرفیت کل ریه	۲/۲۴	۲/۲۵	۲/۳۱	۴/۹۲	۴/۳۲	۳/۹۰	۲/۹۳	۲/۹۸	۳/۱۱	۰/۶۰	۰/۶۱	۰/۵۵
سن	۱۹	۱۷	۱۰	۸۷	۷۴	۷۶	۴۴/۸۳	۳۸/۷۸	۳۳/۲۰	۱۶/۷۴	۱۲/۹۱	۰/۱۶
شاخص توده بدنی	۱۸/۴۰	۱۸/۴۲	۱۹/۴۳	۲۰۱۵	۲۳/۸۰	۲۴/۵۰	۵۲/۳۶	۲۱/۲۳	۲۱/۹	۲۴۷/۲۵	۱/۵۳	۱/۵۴

از ویژگی‌های موجود در دادگان، ۵ ویژگی به عنوان نمونه برای نشان دادن نحوه توزیع داده‌ها در بین کلاس‌های مختلف انتخاب شد. شکل ۵ نحوه پراکندگی داده‌ها را نشان می‌دهد. نوع بیماری‌ها بر اساس رنگ مشخص شده است. رنگ آبی بیماری آسم، رنگ قرمز بیماری برونشیت مزمن و رنگ زرد نشانگر بیماری آمفیزم می‌باشد.



شکل ٥: پراکندگی دادگان به ازای ویژگی‌های مختلف

[DOI: 10.34172/jhbmi.2023.29]

فاز دو پیش پردازش داده

یکسان‌سازی داده‌های حرفی و عددی

ویژگی‌های دسته‌ای categorical موجود در پایگاه داده به صورت عددی تبدیل شدند. به عنوان مثال: ویژگی درجه شدت بیماری به صورت خفیف =۱، متوسط=۲ و شدید=۳ کد گذاری می‌شود.

جدول ۵: تبدیل ویژگی‌های دسته‌ای به عددی

نام ویژگی	نام گذاری حرفی	نام گذاری عددی
سن	زن/ مرد	۰-۱
سابقه مصرف سیگار	بله/خیر	۰-۱
سابقه مصرف الکل	بله/خیر	۰-۱
شدت	خفیف/ متوسط/ شدید	۱-۲-۳
نام بیماری	برچسب(آسم، برونشیت، آمفیم)	۱-۲-۳

شناسایی و حذف داده‌های پرت

با استفاده از روش چارک و نمودار جعبه‌ای، داده‌های پرت شناسایی شدند که در این دادگان حدود ۸/۵ درصد داده پرت وجود دارد. ویژگی‌هایی که به صورت دسته‌ای هستند به صورت کامل ثبت شده و مقادیر پرت و از دست رفته ندارد. در جدول ۶ تعداد داده‌های پرت در بقیه ویژگی‌ها مشخص شده است.

جدول ٦: تعداد داده‌های پرت

ویژگی	FEV1	FVC	FEV1/FVC	MEF 25	RR	PEF	TLC	RLC	BMI
تعداد	۷	۲	۸	۶	۰	۰	۴	۵	۲

فاز سه تعیین ارزش ویژگی‌ها

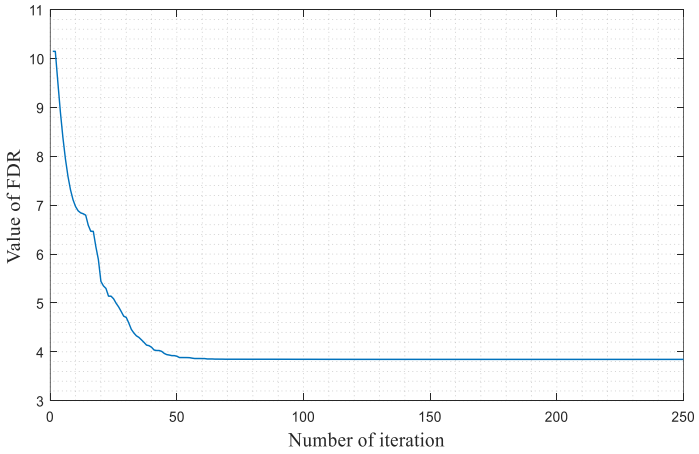
در این فاز برای تعیین ارزش ویژگی‌ها از الگوریتم بهینه‌سازی ملخ استفاده شد. این روش بهینه‌سازی با جمعیت اولیه ۳۰ ملخ شروع می‌شود. شرط خاتمه این الگوریتم رسیدن به بالاترین نرخ جداپذیری فیشر در ۲۵۰ تکرار می‌باشد که توسط تابع برازندگی تا ۴ رقم اعشار محاسبه شد. پارامترهای انتخابی این الگوریتم مطابق با جدول ۷ می‌باشد.

جدول ۷: پارامترهای انتخابی الگوریتم ملخ

پارامتر	مقدار انتخابی
اندازه جمعیت	۳۰
تعداد تکرار	۲۵۰
تعداد متغیرها	به تعداد ویژگی‌ها
محدوده متغیرها	۲۰- تا ۲۰+
ضریب کاهشی پارامتر C	Cmin=0 Cmax=1
تابع برازندگی	رابطه FDR

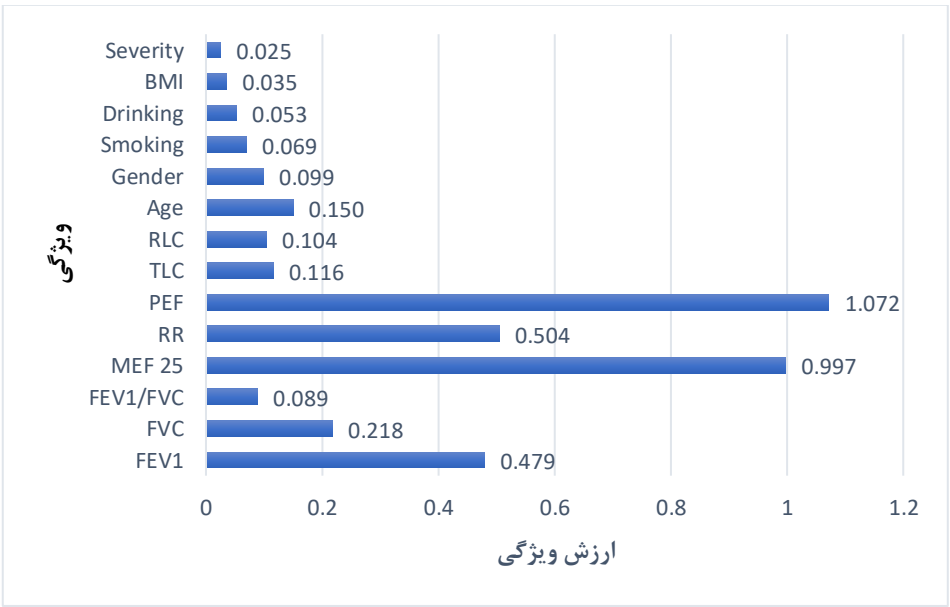
[DOI: 10.34172/jhbmi.2023.29]

نمودار شکل ۶ بر اساس مقادیر FDR محاسبه شده توسط تابع برازندگی، رسم شد. این نمودار نشانگر به جواب رسیدن الگوریتم ملخ می‌باشد. روند صاف شدن و عدم تغییر نمودار نشانگر این است که الگوریتم به ماکزیممم‌ترین مقدار ممکن fdr رسیده است.



شکل ٦: نمودار مینیمم‌سازی تابع برازندگی در الگوریتم GOA

تعیین ضرایب ویژگی‌ها در راستای افزایش FDR، توسط الگوریتم بهینه‌سازی ملخ فرآیندی تصادفی می‌باشد؛ بنابراین برای اطمینان از نتایج به دست آمده، الگوریتم چندین بار تکرار گردید. طبق نتایج به دست آمده، دامنه نوسان تغییرات در هر بار اجرا کم می‌باشد، ولی برای حصول نتیجه دقیق و مطمئن، میانگین نتایج به دست آمده به عنوان نتیجه نهایی گزارش می‌شود. نتیجه به صورت شکل ۷ قابل مشاهده است. این ضرایب، ارزش ویژگی‌های موجود در دادگان را در کنار یکدیگر نشان می‌دهد.



شکل ۷: ارزش ویژگی‌ها

[DOI: 10.34172/jhbmi.2023.29]

مقادیر ضرائب بهینه شده طبق شکل ۷ می‌باشد. با اعمال این ضرایب در بردار ویژگی هر نمونه، ویژگی‌های بهینه به دست می‌آید. الگوریتم KNN با ویژگی‌های بهینه تشکیل می‌شود.

فاز چهار کلاس‌بندی KNN وزن‌دار

بعد از این که KNN به عنوان روش کلاس‌بندی انتخاب شد. لازم است قسمت‌های مختلف این الگوریتم مطابق با دادگان انتخاب و تنظیم شود. ایده‌ها و روش پیشنهادی در مراحل این الگوریتم اعمال شود تا بتواند به بهترین نحو به هدف این مطالعه که کلاس‌بندی بیماری‌های انسدادی ریوی می‌باشد، دست یابد.

انتخاب نحوه محاسبه فاصله (شباهت)

روش‌های مختلفی برای محاسبه فاصله داده آزمایش از مجموعه داده‌های آموزش وجود دارد. این روش‌ها در جدول ۱ ذکر شده است. برای انتخاب روشی مناسب برای محاسبه میزان شباهت داده‌ها، روش‌های مختلفی انتخاب و ارزیابی شد. جدول ۸ نتایج معیارهای مختلف محاسبه شباهت اعمال شده در الگوریتم KNN را نشان می‌دهد.

جدول ۸: نتایج معیارهای مختلف محاسبه فاصله

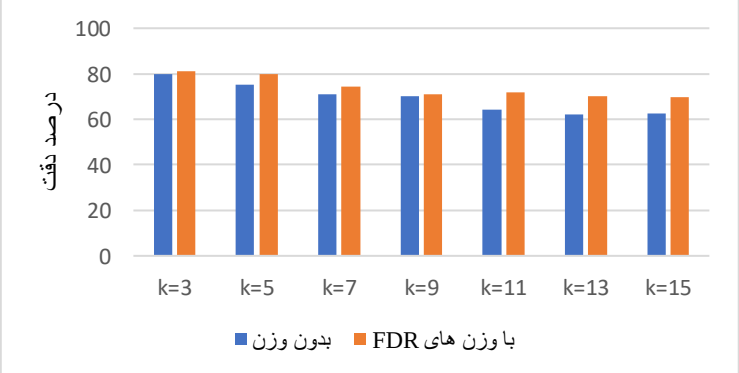
روش	میانگین دقت
فاصله اقلیدسی	۹۱/۹۷۹۹۵
فاصله بلوک شهری	۹۱/۷۲۹۳۲
فاصله چبیشف	۹۱/۹۰۸۳۴
فاصله مینکوفسکی	۹۲/۳۳۷۹۹
فاصله کسینوسی	۹۱/۵۸۶۱۱
فاصله همبستگی	۹۲/۲۳۰۵۸

با تغییر مقدار K به ازای مقادیر مختلف، دقت محاسبه گردید. طبق نتایج جدول، روش Minkowski با میانگین ۹۲/۳۳ درصد دقت، برای محاسبه میزان شباهت داده‌ها مناسب‌ترین روش می‌باشد.

محاسبه وزن‌دار فاصله داده آزمون با داده آموزش

ارزش ویژگی‌های به دست آمده در فاز به صورت یک ترکیب خطی در برداری به نام a قرار گرفت. ضرایب در فرمول فاصله مینوکوفسکی اضافه شده و در محاسبه فاصله وزن‌دار نمونه‌ها از هم به کار برده شد.

شکل ۸ نمودارهای دقت کلاس‌بندی الگوریتم KNN ساده و KNN بهینه شده با ضرایب FDR، به ازای K های مختلف می‌باشد. با میانگین گرفتن از دقت‌های به دست آمده، KNN بهینه عملکرد بهتری به نسبت KNN کلاسیک دارد که نتایج در جدول ۹ قابل مشاهده است؛ بنابراین ثابت می‌شود که تمام ویژگی‌های موجود در دادگان سهم یکسانی در توصیف و کلاس‌بندی مسئله مورد نظر ندارند و متفاوت در نظر گرفتن تأثیر ویژگی‌ها در محاسبه میزان شباهت داده آزمایش به دادگان آموزش، در جهت افزایش دقت می‌تواند سودمند باشد.



شکل ۸: مقایسه دقت KNN کلاسیک و KNN بهینه به ازای k مختلف

[DOI: 10.34172/jhbmi.2023.29]

جدول ۹: مقایسه محاسبه فاصله به صورت ساده و وزن‌دار

ردیف	نحوه محاسبه فاصله	میانگین دقت
۱	بدون ضرایب	۶۹/۲۸
۲	با ضرایب	۷۳/۹۳

رأی‌گیری وزن‌دار

در بخش پایانی الگوریتم KNN باید برای تعیین کلاس داده مورد نظر، رأی‌گیری انجام شود. KNN ساده با توجه به میزان فراوانی نوع برچسب همسایگان نزدیک، تصمیم گیری می‌کند؛ اما در KNN بهینه علاوه بر تعداد هر کلاس، میزان شباهت همسایگان با داده آزمایش به صورت وزن‌هایی در نظر گرفته می‌شوند. این وزن‌ها در واقع همان فواصل محاسبه شده همسایگان از داده آزمایش می‌باشد که از طریق کرنل‌های معرفی شده در جدول ۳، به دست آمد و این در مرحله رأی‌گیری اعمال می‌شوند.

انواع مختلف این کرنل‌ها در الگوریتم اعمال شده و نتایج به صورت جدول۱۰ به دست آمد. کرنل شماره ۵ با الگوی گوسی، مناسب‌ترین روش برای تبدیل فاصله به میزان شباهت می‌باشد.

جدول ۱۰: نتایج روش‌های مختلف تبدیل فاصله به میزان شباهت

شماره	فرمول	میانگین دقت
۱	$\frac{1}{d(x_i,x)}$	۹۲/۳۳
۲	$\frac{1}{d(x_i,x)^2}$	۹۲/۱۹
۳	$\frac{1}{c+d(x_i,x)^2}$	۹۲/۱۹
۴	$exp\left(-\frac{d(x_i,x)^2}{\sigma^2}\right)$	۹۱/۷۶
۵	$\frac{1}{\sqrt{2*pi}}*exp\left(-\frac{d(x_i,x)^2}{\sigma^2}\right)$	۹۲/۷۳
۶	$\frac{1}{d(x_i,x)^{\frac{2}{m-1}}}$	۴۳/۸۲

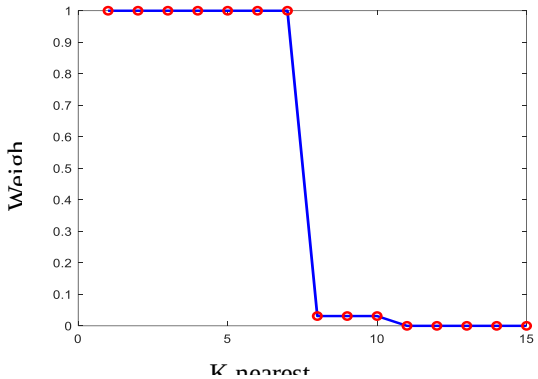
جدول ۱۱: نتایج روش‌های تبدیل فاصله به میزان شباهت به ازای K مختلف

شماره کرنل	k=۳ تعداد	۵	۷	۹	۱۱	۱۳	۱۵
۱	۹۱/۹۷۹۹۵	۹۲/۲۳۰۵۸	۹۲/۴۸۱۲	۹۲/۴۸۱۲	۹۱/۹۷۹۹۵	۹۲/۷۳۱۸۳	۹۲/۴۸۱۲
۲	۹۱/۹۷۹۹۵	۹۱/۹۷۹۹۵	۹۲/۲۳۰۵۸	۹۲/۴۸۱۲	۹۲/۲۳۰۵۸	۹۱/۹۷۹۹۵	۹۲/۴۸۱۲
۳	۹۱/۹۷۹۹۵	۹۱/۹۷۹۹۵	۹۲/۲۳۰۵۸	۹۲/۴۸۱۲	۹۲/۲۳۰۵۸	۹۱/۹۷۹۹۵	۹۲/۴۸۱۲
۴	۹۱/۹۷۹۹۵	۹۱/۷۲۹۳۲	۹۱/۷۲۹۳۲	۹۱/۷۲۹۳۲	۹۱/۷۲۹۳۲	۹۱/۷۲۹۳۲	۹۱/۷۲۹۳۲
۵	۹۲/۷۳۱۸۳	۹۲/۷۳۱۸۳	۹۲/۷۳۱۸۳	۹۲/۷۳۱۸۳	۹۲/۷۳۱۸۳	۹۲/۷۳۱۸۳	۹۲/۷۳۱۸۳
۶	۳۷/۰۹۲۷۳	۳۸/۳۴۵۸۶	۴۳/۱۰۷۷۷	۵۱/۶۲۹۰۷	۴۶/۳۶۵۹۱	۴۷/۱۱۷۷۹	۴۳/۱۰۷۷۷

الگوریتم با تمام کرنل‌ها به ازای Kهای مختلف اجرا شد و دقت کلاس‌بندی محاسبه شد. نتایج به صورت جدول ۱۱ به دست آمد. کرنل گوسی شماره ۵ عملکرد ثابتی به ازای تغییر k را دارد؛ بنابراین با انتخاب این روش، میزان حساسیت الگوریتم به پارامتر K کم می‌شود. در نتیجه عملکرد الگویتم در خصوص کاهش زمان بهبود پیدا می‌کند. محاسبه فاصله همسایگی برحسب کرنل

[DOI: 10.34172/jhbmi.2023.29]

گوسی به دلیل ماهیت واریانسی آن علاوه بر فاصله بین داده‌ها، پراکندگی داده‌های کلاس‌ها را نیز در نظر می‌گیرد. نمودار شکل ۹ براساس وزن‌های محاسبه شده توسط کرنل گوسی انتخابی برای K همسایه نزدیک رسم شده است.

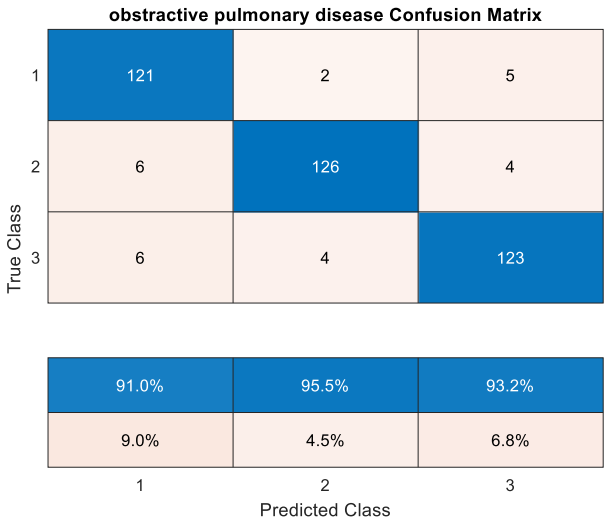


شکل ۹: وزن‌های ۱۵ همسایه نزدیک

کرنل مناسب برای محاسبه وزن‌ها انتخاب شد و طبق آن وزن‌ها محاسبه گردید. مقادیر این وزن‌ها در تعداد همسایگان هر کلاس اعمال شده و رأی نهایی برای تعیین برچسب نمونه‌های آزمایش طبق رابطه۱۰ مشخص شد. این فرآیند برای تمام دادگان آزمایش طبق روش اعتبارسنجی متقابل اجرا شد.

ماتریس درهم ریختگی

نتیجه ارزیابی نهایی کلاس‌بندی بر روی دادگان به صورت شکل ۱۰ ماتریس درهم ریختگی و شکل ۱۱ نمودار ROC به دست آمد.



شکل ۱۰: ماتریس درهم ریختگی

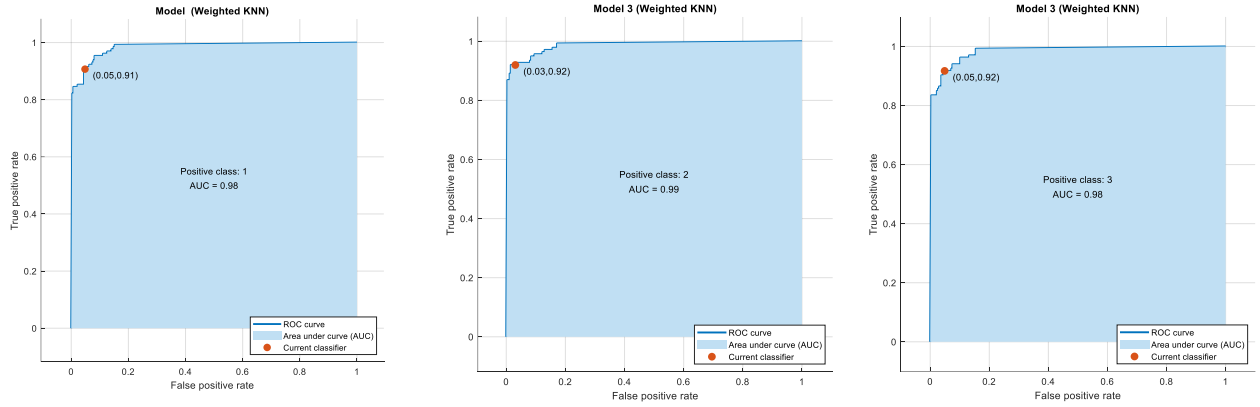
نتایج پارامترهای ارزیابی به تفکیک هر کلاس طبق جدول ۱۲ می‌باشد. براساس ماتریس درهم ریختگی و معیارهای ارزیابی دقت، صحت، حساسیت، تشخیص پذیری به ترتیب ۹۵/۴، ۹۳/۲، ۹۳/۲و ۹۶/۵۹ درصد به دست آمد.

[DOI: 10.34172/jhbmi.2023.29]

جدول ۱۲: نتایج پارامترهای ارزیابی به تفکیک هر کلاس

کلاس	TP	FP	TN	FN
آسم	۱۲۱	۷	۲۵۷	۱۲
برونشیت مزمن	۱۲۶	۱۰	۲۵۵	۶
آمفیزم	۱۲۳	۱۰	۲۵۵	۹
مجموع	۳۷۰	۲۷	۷۶۷	۲۷

نمودار ROC به تفکیک هر کلاس بر اساس نرخ تشخیص صحیح و غلط رسم شد مقدار AUC در کلاس اول و سوم ۰/۹۸ و در کلاس دوم ۰/۹۹ به دست آمده است و نشانگر این است که دقت تشخیص صحیح بالایی در هر کلاس وجود دارد.



ج) کلاس اول ب) کلاس دوم الف) کلاس سوم

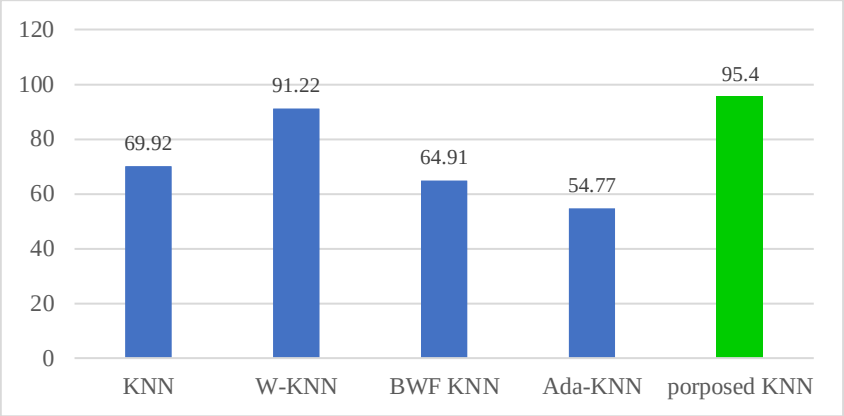
شکل ۱۱: نمودار ROC

مقایسه با روش‌های قبلی

عملکرد روش پیشنهادی از نظر دقت کلاس‌بندی با الگوریتم پایه و سایر الگوریتم‌های جدید ارائه شده در کارهای پیشین مقایسه شد. KNN پایه در ابتدا پیاده‌سازی و با ارزیابی انجام گرفته دقتی معادل ۶۹/۹۲ به دست آمد. روش WKNN با اعمال وزن توانست نسبت به الگوریتم پایه دقت بهتری با ۹۱/۲۲ درصد به دست آورد. دو روش جدید به کار رفته در مطالعات برای بهینه‌سازی الگوریتم پایه توسط شبکه‌های عصبی و فازی، بر روی مجموعه داده این مطالعه پیاده‌سازی شد.

روش BM-FKNN [۲۴] احتمال می‌رفت که با وجود استفاده از روش فازی و میانگین بونفرونی دقت خوبی نسبت به بقیه بر روی مجموعه داده این مطالعه داشته باشد؛ اما با پیاده‌سازی دقت ۶۴/۹۱ به دست آمد که در مقایسه با روش‌های قدیمی پایه، عملکرد ضعیف‌تری کسب کرد. روش Ada-KNN [۲۳] در ارزیابی‌ها پایین‌ترین دقت را به دست آورد. با توجه به نمودار شکل ۱۲ می‌توان مشاهده کرد که دقت کلاس‌بندی الگوریتم ارائه شده در این مطالعه، نسبت به سایر الگوریتم‌های KNN موجود، بیشتر می‌باشد.

[DOI: 10.34172/jhbmi.2023.29]



شکل ۱۲: مقایسه دقت انواع الگوریتم‌های KNN

بحث و نتیجه‌گیری

در این مطالعه براساس دادگان تشکیل‌یافته از پارامترهای پاراکلینیکی و عمومی، روشی پیشنهاد شد که با بهینه‌سازی الگوریتم KNN بتواند بیماری‌های انسدادی ریوی را کلاس‌بندی کند. روش پیشنهادی در چهار مرحله اقدام به کلاس‌بندی سه گروه از بیماری‌های انسدادی ریوی نمود. الگوریتم KNN برای کلاس‌بندی داده‌های مربوط به یک دسته از بیماری که ویژگی‌های مشابه و نزدیک به هم دارند روشی مناسب می‌باشد که از سرعت اجرای بالایی برخوردار است [۳۲].

اما این الگوریتم محدودیت‌هایی از جمله: یکسان در نظر گرفتن تمام همسایگان و قضاوت بر اساس فراوانی هر کلاس، یکسان در نظر گرفتن تمام ویژگی در محاسبه فاصله دو داده و حساسیت به تعداد K دارد [۳۳].

که در این مطالعه روشی برای حل این مشکلات پیشنهاد شد. همه ویژگی‌های موجود در دادگان نمی‌توانند توزیع و سهمی برابر در توصیف و تشخیص مسئله مورد نظر داشته باشند؛ بنابراین ارزش‌گذاری ویژگی‌ها امری مهم می‌باشد. به دلیل این که مسئله مورد مطالعه مربوط به یک دسته از بیماری است که رنج ویژگی‌های نزدیک به هم بوده و تعداد ویژگی محدودی دارد و امکان همپوشانی بین این سه کلاس بیماری وجود دارد. آسم و COPD علائم مشابهی دارند و ابهام در تشخیص گاهاً وجود دارد. برای جلوگیری از ابهامات تشخیصی و افتراق دقیق این بیماری‌ها از هم، در نظر گرفتن تمامی ویژگی‌ها در تشخیص نهایی مهم است [۳].

با به کارگیری رابطه FDR و الگوریتم بهینه‌سازی ملخ تأثیر تمام ویژگی‌ها در کنار هم سنجیده شد و یک ترکیب خطی از ویژگی‌ها که به بهترین حالت داده‌های سه کلاس را از هم جدا کند به دست آورده شد. الگوریتم KNN وزن‌دار پیشنهادی بر اساس محاسبه فاصله مینوکوفسکی با ضرایب FDR و رأی‌گیری وزن‌دار بر پایه کرنل گوسی تشکیل شد که زمان اجرای الگوریتم ۳/۱۲ ثانیه می‌باشد. دقت به دست آمده از این روش ۹۵/۴ درصد می‌باشد. مقایسه انجام شده با سایر روش‌ها نشان داد که روش پیشنهادی دقت و عملکرد بهتری نسبت به بقیه دارد و چون روش‌های پیشین مطرح شده بر روی مجموعه داده متفاوتی آزمایش و ارزیابی شده بودند، در این پژوهش روش‌ها پیاده‌سازی شدند و همه آن‌ها با مجموعه داده حاضر اجرا و ارزیابی شده و سپس مقایسه شدند که نتایج در شکل ۱۲ آورده شده است.

تعداد نمونه‌ها و متغیرهای این مطالعه می‌تواند به عنوان یک محدودیت در نظر گرفته شود. در آینده، متغیرهای جدیدی که ممکن است برای پیش‌بینی و تشخیص بیماری مؤثر باشد، مانند ویژگی‌های آنتروپومتریک بیشتر، ویژگی‌های مرتبط با عادات و سبک زندگی و اندازه‌گیری‌های پزشکی با مشاهدات بیشتر اضافه شود. مطالعات بعدی می‌تواند بر اساس پیشنهاداتی مانند: به‌کارگیری الگوریتم پیشنهادی در جهت کلاس‌بندی

[DOI: 10.34172/jhbmi.2023.29]

سایر بیماری‌های ریوی، استفاده از دیگر روش‌های محاسبه ارزش ویژگی‌ها و استفاده از روش پیشنهادی در سیستم‌های تشخیصی آنلاین انجام گیرد.

تعارض منافع

نویسندگان اعلام می‌دارند که هیچ گونه تعارض منافعی در مطالعه حاضر وجود ندارد و این پژوهش فاقد حمایت مالی بوده است.

References

1. World Health Organization (WHO). Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach. World Health Organization; 2007 [cited 2022 Oct 5]. Available from: http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

2. Anakal S, Sandhya P. Clinical decision support system for chronic obstructive pulmonary disease using machine learning techniques. International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT); 2017 Dec 15-16; Mysuru, India: IEEE; 2017. p. 1-5. doi: 10.1109/ICEECCOT.2017.8284601

3. Haider NS, Behera AK. Computerized lung sound-based classification of asthma and chronic obstructive pulmonary disease (COPD). Biocybernetics and Biomedical Engineering 2022;42(1):42-59.

4. Srivastava A, Jain.S, Miranda R, Patil S, Pandya S, Kotecha K. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. PeerJ Computer Science 2021;7:e369. doi: 10.7717/peerj-cs.369

5. Hyatt RE, Scanlon PD, Nakamura M. Interpretation of pulmonary function tests. 4th ed. Philadelphia, Pennsylvania, United States: Lippincott Williams & Wilkins; 2014.

6. Saha S, Majumdar S, Bhattacharyya P. Obstructive Airway Diseases. InPulmonomics: Omics Approaches for Understanding Pulmonary Diseases. Singapore: Springer Nature Singapore. 2023. p. 21-30.

7. Ranu H, Wilde M, Madden B. Pulmonary function tests. Ulster Med J 2011;80(2):84-90.

8. Kocks JW, Cao H, Holzhauer B, Kaplan A, FitzGerald JM, Kostikas K, et al. Diagnostic Performance of a Machine Learning Algorithm (Asthma/Chronic Obstructive Pulmonary Disease [COPD] Differentiation Classification) Tool Versus Primary Care Physicians and Pulmonologists in Asthma, COPD, and Asthma/COPD Overlap. The Journal of Allergy and Clinical Immunology: In Practice 2023; 11(5):1463-74. https://doi.org/10.1016/j.jaip.2023.01.017

9. Kaplan A, Cao H, FitzGerald JM, Iannotti N, Yang E, Kocks JW, et al. Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. J Allergy Clin Immunol Pract 2021;9(6):2255-61. doi: 10.1016/j.jaip.2021.02.014.

10. Topalovic M, Das N, Burgel PR, Daenen M, Derom E, Haenebalcke C, et al. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. Eur Respir J 2019;53(4):1801660. doi: 10.1183/13993003.01660-2018

11. Hasbi NH, Bade A, Chee FP, Rumaling MI. Pattern recognition for human diseases classification in spectral analysis. Computation 2022;10(6):96. https://doi.org/10.3390/computation10060096

12. Swaminathan S, Qirko K, Smith T, Corcoran E, Wysham NG, Bazaz G, et al. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. PLoS One 2017;12(11):e0188532. https://doi.org/10.3390/computation10060096

13. Wu CT, Li GH, Huang CT, Cheng YC, Chen CH, Chien JY, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. JMIR Mhealth Uhealth 2021;9(5):e22591. doi: 10.2196/22591.

14. Ioachimescu OC, Stoller JK. An alternative spirometric measurement. area under the expiratory flow–volume curve. Ann Am Thorac Soc 2020;17(5):582-8. doi: 10.1513/AnnalsATS.201908-613OC.

15. Haider NS, Singh BK, Periyasamy R, Behera AK. Respiratory sound-based classification of chronic obstructive pulmonary disease: a risk stratification approach in machine learning paradigm. Journal of Medical Systems 2019;43:1-3.

16. Hussain A, Choi HE, Kim HJ, Aich S, Saqlain M, Kim HC. Forecast the exacerbation in patients of chronic obstructive pulmonary disease with clinical indicators using machine learning techniques. Diagnostics 2021;11(5):829. doi: 10.3390/diagnostics11050829

17. Spathis D, Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health Informatics J 2019;25(3):811-27. doi: 10.1177/1460458217723169.

18. Siddiqui HU, Saleem AA, Bashir I, Zafar K, Rustam F, Diez ID, et al. Respiration-Based COPD Detection Using UWB Radar Incorporation with Machine Learning. Electronics 2022;11(18):2875. https://doi.org/10.3390/electronics11182875

19. Bhattacharjee S, Saha B, Bhattacharyya P, Saha S. Classification of obstructive and non-obstructive pulmonary diseases on the basis of spirometry using machine learning techniques. Journal of Computational Science 2022;63:101768. https://doi.org/10.1016/j.jocs.2022.101768

[DOI: 10.34172/jhbmi.2023.29]

20. Raghavan N, Lam YM, Webb KA, Guenette JA, Amornputtisathaporn N, Raghavan R, et al. Components of the COPD Assessment Test (CAT) associated with a diagnosis of COPD in a random population sample. COPD 2012;9(2):175-83. doi: 10.3109/15412555.2011.650802.
21. Vora S, Shah C. COPD classification using machine learning algorithms. Int Res J Eng Technol 2019;6(4):608-11.
22. Tarakci F, Ozkan IA. Comparison of classification performance of kNN and WKNN algorithms. Selcuk University Journal of Engineering Sciences 2021;20(2):32-7.
23. Mullick SS, Datta S, Das S. Adaptive Learning-Based -Nearest Neighbor Classifiers with Resilience to Class Imbalance. IEEE Trans Neural Netw Learn Syst 2018;29(11):5713-25. doi: 10.1109/TNNLS.2018.2812279.
24. Kumbure MM, Luukka P, Collan M. A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. Pattern Recognition Letters 2020;140:172-8. https://doi.org/10.1016/j.patrec.2020.10.005
25. Vinutha HP, Poornima B, Sagar BM. Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. In book: Information and Decision Sciences. Singapore: Springer; 2018. p. 511-8. doi:10.1007/978-981-10-7563-6_53
26. Saremi S, Mirjalili S, Lewis A. Grasshopper optimisation algorithm: theory and application. Advances in Engineering Software 2017;105:30-47. https://doi.org/10.1016/j.advengsoft.2017.01.004
27. Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. International Conference on Intelligent Computing and Control Systems (ICCS); 2019 May 15-17; Madurai, India: IEEE; 2019. p. 1255-60.
28. Kramer O. Dimensionality reduction with unsupervised nearest neighbors. Berlin: Springer; 2013.
29. Abu Alfeilat HA, Hassanat AB, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big Data 2019;7(4):221-48. doi: 10.1089/big.2018.0175.
30. Gou J, Xiong T, Kuang Y. A Novel Weighted Voting for K-Nearest Neighbor Rule. Journal of Computers 2011;6(5):833-40. doi:10.4304/jcp.6.5.833-840
31. Gatz DF, Smith L. The standard error of a weighted mean concentration—I. Bootstrapping vs other methods. Atmospheric Environment 1995;29(11):1185-93.
32. Suguna N, Thanushkodi K. An improved k-nearest neighbor classification using genetic algorithm. International Journal of Computer Science Issues 2010;7(2):18-21.
33. Wang AX, Chukova SS, Nguyen BP. Ensemble k-nearest neighbors based on centroid displacement. Information Sciences 2023;629:313-23. https://doi.org/10.1016/j.ins.2023.02.004

[DOI: 10.34172/jhbmi.2023.29]