Sci J Iran Blood Transfus Organ 2024; 21 (2): 151-159

Original Article

Predicting pre-collection umbilical cord blood clotting using advanced machine learning algorithms

Esmaielpour A.H.[1], Ameli M.[1], Mozdgir A.[1], Ahmadi O.[1], Zarabi M.[2]

[1]Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran
[2]Department of Regenerative Medicine, Royan Institute for Stem Cell Biology and Technology, Tehran, Iran

Abstract

Background and Objectives

Umbilical cord blood is a valuable source of stem cells used in transplants to treat various diseases including leukemia, lymphoma and genetic disorders. However, cord blood clotting during the collection process can reduce sample quality and quantity and impact its efficacy in cord blood banking. This article aims to predict pre-collection cord blood clotting in donors using advanced machine learning techniques.

Materials and Methods

In this retrospective study, data was gathered using 928127 samples available in the fetal cord blood bank, and with using supervised machine learning classification algorithms, including decision tree, naïve Bayes, K-Nearest Neighbors, Support vector machine, Random forest, Majority voting and Multilayer perceptron, prediction of cord blood clotting was performed on the Royan cord blood bank database and their performance was compared using evaluation metrics such as Accuracy, Precision, Recall, and F1 Score.

Results

In this study, the algorithm accuracy of Decision Tree was 0.80, Naive Bayes was 0.63, K-Nearest Neighbors was 0.83, Support Vector Machine was 0.65, Random Forest was 0.84, Majority Voting Classifier was 0.81, and Multilayer Perceptron was 0.74.

Conclusions

In this study, the performance of Random Forest and K-Nearest Neighbors algorithms demonstrated the best accuracy showing that machine learning algorithms can predict prenatal cord blood clotting with high accuracy which can help prevent sampling of clotted specimens in order to reduce costs and storage problems.

Key words: Stem Cells, Machine Learning, Umbilical Cord Blood, Bioanformatics

Received: 8 Jan 2024
Accepted: 10 Jun 2024

Correspondence: Ameli M., PhD in Industrial Engineering. Assistant Professor of Department of Industrial Engineering, Faculty of Engineering, Kharazmi University.
Postal Code: 1571914911, Tehran, Iran. Tel: (+9826) 34523124; Fax: (+9821) 88825580
E-mail: *m.ameli @khu.ac.ir*

خون
فصلنامه پژوهشی
دوره ۲۱ شماره ۲ تابستان ۱۴۰۳ (۱۵۹-۱۵۱)

مقاله پژوهشی

پیش‌بینی لخته شدن خون بند ناف پیش از جمع‌آوری با کمک الگوریتم‌های یادگیری ماشین پیشرفته

امیرحسین اسمعیل‌پور[۱]، مریم عاملی[۲]، اشکان مزدگیر[۳]، اُرد احمدی[۳]، مرتضی ضرابی[۴]

چکیده

سابقه و هدف

خون بند ناف منبع ارزشمندی از سلول‌های بنیادی است که در پیوند برای درمان بیماری‌های مختلف از جمله لوسمی، لنفوم و اختلالات ژنتیکی مورد استفاده قرار می‌گیرد. با این حال، لخته شدن خون بند ناف در فرآیند جمع‌آوری می‌تواند کیفیت نمونه را کاهش دهد و بر اثر بخشی آن در ذخیره‌سازی خون بند ناف در بانک‌ها تأثیر بگذارد. در این مقاله با استفاده از روش‌های پیشرفته یادگیری ماشین، لخته‌شدن خون بند ناف قبل از جمع‌آوری نمونه‌ها از اهداکنندگان پیش‌بینی شده است.

مواد و روش‌ها

در یک مطالعه گذشته‌نگر، تعداد ۹۲۸۱۲۷ نمونه از بانک خون بند ناف رویان از سال ۱۳۸۴ تا ۱۴۰۰ بررسی شدند. داده‌ها با استفاده از نمونه‌های موجود در بانک خون بند ناف رویان و با استفاده از الگوریتم‌های طبقه‌بندی یادگیری نظارت شده، از جمله درخت تصمیم، بیزین ساده، K- نزدیک‌ترین همسایه، ماشین‌بردار پشتیبان، جنگل تصادفی، طبقه‌بندی رأی اکثریت و پرسپترون چند لایه برای پیش‌بینی لخته‌شدن خون بند ناف بر روی داده‌های بانک خون بند ناف رویان اجرا و عملکرد آن‌ها با استفاده از معیارهای ارزیابی دقت، صحت، بازخوانی و امتیاز F1 مقایسه شد.

یافته‌ها

در این مطالعه دقت الگوریتم درخت تصمیم ۰/۸۰، بیزین ساده ۰/۶۳، K- نزدیک‌ترین همسایه ۰/۸۳، ماشین‌بردار پشتیبان ۰/۶۵، جنگل تصادفی ۰/۸۴، طبقه‌بندی رأی اکثریت ۰/۸۱ و پرسپترون چند لایه ۰/۷۴ اندازه‌گیری شده است.

نتیجه گیری

در این مطالعه عملکرد دو الگوریتم جنگل تصادفی و K- نزدیک‌ترین همسایه بهترین کارآیی را از خود نشان دادند و بیانگر آن است که می‌توان با کمک الگوریتم‌های یادگیری ماشین، با دقت بالایی بروز لخته پیش از زایمان را در نوزاد پیش‌بینی کرد و به کمک آن می‌توان از نمونه‌برداری نمونه‌های دارای لخته به منظور کاهش هزینه و مشکلات ذخیره‌سازی آن‌ها جلوگیری نمود.

کلمات کلیدی: سلول‌های بنیادی، یادگیری ماشین، خون بند ناف، بیوانفورماتیک

تاریخ دریافت: ۱۴۰۲/۱۰/۱۸
تاریخ پذیرش: ۱۴۰۳/۰۳/۲۱

۱- دانشجوی دکترای مهندسی صنایع ـ دانشکده فنی و مهندسی دانشگاه خوارزمی ـ تهران ـ ایران
۲- مؤلف مسئول: دکترای مهندسی صنایع ـ استادیار گروه مهندسی صنایع ـ دانشکده فنی و مهندسی دانشگاه خوارزمی ـ تهران ـ ایران ـ کدپستی: ۱۵۷۱۹۱۴۹۱۱
۳- دکترای مهندسی صنایع ـ استادیار گروه مهندسی صنایع ـ دانشکده فنی و مهندسی دانشگاه خوارزمی ـ تهران ـ ایران
۴- پزشک عمومی ـ گروه پزشکی بازساختی ـ پژوهشکده زیست شناسی و فناوری سلول‌های بنیادی ـ پژوهشگاه رویان ـ تهران ـ ایران

مقدمه

خون بند ناف (Umbilical Cord Blood: UCB) منبع غنی از سلول‌های بنیادی خونساز است که می‌تواند در درمان اختلالات ژنتیکی، نقص ایمنی و اختلالات خونی مورد استفاده قرار گیرد. پیوند خون بند ناف یک روش پزشکی است که در آن سلول‌های بنیادی خون بند ناف یک نوزاد سالم به بیمار پیوند زده می‌شود. این پیوند برای درمان بیماری‌های مختلفی از جمله سرطان‌های خون، اختلالات خونی، سندرم‌های نارسایی مغز استخوان و اختلالات ایمنی کاربرد دارد (۱).

یکی از چالش‌های اصلی در جمع‌آوری خون بند ناف، بروز لخته‌های خون در بند ناف است که مانع از جریان خون و کاهش تولید سلول‌های بنیادی می‌شود. بدین ترتیب جمع‌آوری و ذخیره‌سازی UCB با وجود لخته در بند ناف می‌تواند پیچیده شده و عملکرد سلول‌های بنیادی را کاهش داده و کیفیت آن‌ها را به خطر بیندازد؛ بنابراین، پیش‌بینی احتمال لخته‌شدن بند ناف قبل از جمع‌آوری می‌تواند به اطمینان از عملکرد بالاتر سلول‌های بنیادی و نتایج بهتر برای بیماران کمک کند و به پزشکان در تصمیم‌گیری آگاهانه کمک نماید (۲).

بند ناف از دو شریان و یک ورید تشکیل شده است که اکسیژن و مواد مغذی را به جنین می‌رساند و مواد زائد را دفع می‌کند. بند ناف هم‌چنین حاوی ماده ژلاتینی به نام ژله وارتون است که از فشرده شدن رگ‌ها جلوگیری می‌کند. با این حال، در حین زایمان، بند ناف می‌تواند فشرده یا پیچ خورده شود و منجر به تشکیل لخته گردد. عوامل متعدد دیگری نیز ممکن است در شکل‌گیری لخته در خون بند ناف نقش‌آفرین باشند (۳). برخی از این عوامل که ریشه در بیماری‌های مادر و نوزاد دارند، مانند دیابت مادر در زمان بارداری، رابطه مستقیم با لخته‌شدن خون نوزاد دارند (۴). حیفض و همکاران در پژوهش خود ۵۲ مورد لخته شدن بند ناف از ۳ جمعیت مختلف را مورد تجزیه وتحلیل قرار داده و با ۶۸ مورد از ادبیات موضوع مقایسه کرد که بدین ترتیب مجموعه‌ای از عوامل و بیماری‌های مختلف را در شکل‌گیری لخته مورد بررسی قرار داد و به این نتیجه رسید که در نوزادان پسر احتمال بیشتری برای لخته‌شدن وجود دارد (۵). در این پژوهش عواملی هم‌چون عوارض مامایی مانند انواع عفونت و شرایط سیستمیک جنین مانند دیابت و خونریزی جنینی مهم ارزیابی شدند. از جمله عامل دیگری که در ایجاد لخته تاثیرگذار است، کم‌خونی مادر می‌باشد (۶). تغییرات انعقاد خون بند ناف در فشار خون مادر در پژوهش لاکس مورد بررسی قرار گرفت (۷). این پژوهش هم‌چنین به این نتیجه رسید که آسیب کبدی به عنوان یکی از اصلی‌ترین علل انعقاد خون بند ناف نوزاد مطرح است.

با توجه به این که پیش‌بینی لخته شدن خون بند ناف پیش از زایمان می‌تواند منجر به بهبود کمیت و کیفیت شود و به علاوه صرفه جویی اقتصادی به همراه داشته باشد، لذا استفاده از مدل‌های پیش‌بینی کننده در این زمینه، امری ضروری است. امروزه روش‌های یادگیری ماشین (Machine Learning: ML) به عنوان ابزاری قدرتمند برای پیش‌بینی پدیده‌ها و متغیرهای مختلف در علوم گوناگون به کار می‌روند. هر چند هنوز مطالعه‌ای به منظور پیش‌بینی لخته‌شدن خون بند ناف با این ابزار صورت نگرفته است. از این رو در ادامه پژوهش‌هایی که از ML برای کنترل کیفیت سلول‌های بنیادی استفاده کرده‌اند، مورد بررسی قرار می‌گیرند.

هدف بسیاری از پژوهش‌ها در این حوزه تعریف پیش‌بینی‌کننده‌های بالینی قبل از تولد برای تعداد سلول‌های هسته‌دار (Total Nucleated Cell count: TNC) است که به شناسایی اهداکنندگان موفق واحدهای خون بند ناف قبل از شروع زایمان فعال کمک می‌کند. این پژوهش‌ها با پیش‌بینی TNC تلاش می‌کنند تا کارآیی بانک خون در ذخیره‌سازی نمونه‌های بهتر را افزایش دهند. این مطالعه‌ها از روش‌های یادگیری گروهی (Ensemble Learning) و روش‌های کلاسیک ML مانند درخت تصمیم استفاده کردند (۹، ۸). هم‌چنین هاره با کمک رگرسیون به این نتیجه رسید که سن حاملگی، نژاد مادر و وزن و جنس نوزاد با TNC ارتباط دارد (۱۱، ۱۰). با کمک یادگیری ماشین و روش‌های آماری مختلف مانند شبکه‌های عصبی پرسپترون چند لایه، رگرسیون لجستیک و درخت تصمیم، ارزش کیفی نمونه‌های خون و دسته‌بندی مناسب آن‌ها

(برای دور انداختن یا انجماد) را تعیین کرد.

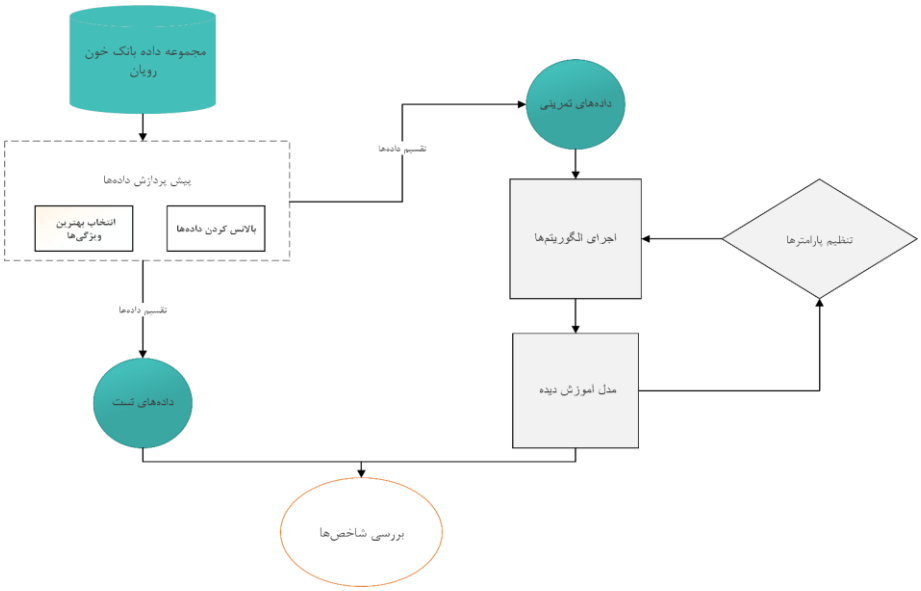
همان طور که پژوهش‌های بالا نشان می‌دهند، روش‌های یادگیری ماشین توانایی لازم برای ارائه پیش‌بینی کیفیت خون بند ناف را دارا هستند اما این روش‌ها برای پیش‌بینی وجود و یا بروز لخته در خون بند ناف پیش از زایمان استفاده نشده‌اند. هدف این مقاله، پیش‌بینی لخته‌شدن خون بند ناف پیش از جمع‌آوری نمونه‌ها از اهداکنندگان با کمک روش‌های یادگیری ماشین پیشرفته بود. عملکرد الگوریتم‌های مختلف یادگیری ماشینی برای پیش‌بینی لخته‌شدن خون بند ناف ارزیابی شده و در مورد پیامدهای بالقوه این یافته‌ها برای عمل بالینی بحث می‌شود.

مواد و روش‌ها

در این پژوهش از زبان برنامه‌نویسی پایتون (Python) استفاده شده است که یک زبان برنامه‌نویسی سطح بالا و همه منظوره است. پایتون یک زبان قدرتمند با طیف گسترده‌ای از کتابخانه‌ها و ابزارها محسوب می‌شود. تا سال ۲۰۲۰، ۸۹٪ پژوهش‌های مربوط به ML با کمک زبان پایتون نوشته شده است (۱۲).

توصیف مجموعه داده‌ها:

این پژوهش از نوع گذشته‌نگر بود. داده‌ها شامل ۹۲۸۱۲۷ نمونه جمع‌آوری شده در بانک خون بند ناف رویان از سال ۱۳۸۴ تا ۱۴۰۰ بود. این داده‌ها ۸۶ ستون (ویژگی) شامل اطلاعات مربوط به لخته شدن و برخی از ویژگی‌های هر اهداکننده هستند. معیار انتخاب این ویژگی‌ها در دسترس‌بودن و قابلیت اندازه‌گیری آن‌ها پیش از تولد نوزاد بود. ویژگی‌هایی که تعداد رکوردهای آن‌ها به طور قابل‌توجهی از دست رفته بود نیز حذف شدند. در نهایت تعداد ستون‌های مورد بررسی (بدون ستون هدف یعنی لخته بودن خون بند ناف) ۳۲ ویژگی را تشکیل دادند. لازم به ذکر است که ویژگی‌های مربوط به سابقه بیماری در خانواده، پیش از این و توسط مجموعه بانک خون رویان به کمک پرسش‌نامه‌ای آنلاین جمع‌آوری شده است. برای داده‌ها و رکوردهایی که ازدست رفته و یا پرت بودند، سطر مربوط به آن‌ها از بین مجموع سطرها حذف شد. در یادگیری ماشین، الگوریتم‌ها بر اساس ویژگی‌هایی که از داده‌ها استخراج می‌شوند، یاد می‌گیرند. این ویژگی‌ها در واقع پیش‌بینی کننده‌های لختگی خون بند ناف هستند یعنی نشانه‌هایی هستند که به الگوریتم کمک می‌کنند تا الگوهایی را در داده‌ها شناسایی کند و از آن‌ها برای پیش‌بینی یا تصمیم‌گیری استفاده کند (۱۳). در شکل گام‌های پژوهش به منظور پیش‌بینی بروز لخته به صورت شماتیک رسم شده‌اند (شکل ۱).



شکل ۱: گام‌های پژوهش به منظور پیش‌بینی بروز لخته در خون بند ناف

ستون هدف برای پیش‌بینی لخته بودن خون بند ناف پیش از اهدا شامل دو دسته به صورت باینری است. دسته اول شامل نمونه‌هایی است که دارای لخته هستند و دسته دوم شامل نمونه‌هایی است که لخته ندارند.

با توجه با این که دسته اول شامل ۲۵۰۸۹ نمونه و دسته دوم شامل ۱۰۰۷۶۵ نمونه بود، مشخص می‌شود داده‌ها نامتوازن هستند. در الگوریتم‌های طبقه‌بندی استاندارد، توزیع کلاس‌ها متوازن در نظر گرفته می‌شود و این دسته از الگوریتم‌ها در مواجهه با مجموعه‌داده‌های نامتوازن، عملکرد مناسبی را از خود ارائه نمی‌دهند؛ چرا که الگوریتم‌های معمول طبقه‌بندی به سمت نمونه‌های آموزشی کلاس بزرگ‌تر متمایل می‌شوند که این موضوع باعث افزایش خطا در شناسایی نمونه‌های اقلیت می‌شود (۱۴). به منظور مقابله با این مشکل، در این پژوهش از روش نمونه‌برداری بیش از حد اقلیت مصنوعی (Synthetic Minority Oversampling Technique: SMOTE) استفاده شد که امکان تولید داده‌های مصنوعی را فراهم می‌سازد. این روش با استفاده از همسایه‌های هر نمونه از کلاس اقلیت، نمونه‌های مصنوعی جدیدی می‌سازد. به این صورت که در مرحله اول، به‌ازای هر نمونه i از کلاس اقلیت، k همسایه نزدیک‌ترش در همان کلاس پیدا شود. در مرحله دوم به ازای تمام خصوصیات هر همسایه j در k همسایه نزدیک‌تر، فاصله نمونه i تا j محاسبه می‌شود. سپس در مرحله سوم، مقداری بین ۰ تا ۱ به‌عنوان شکاف در نظر گرفته می‌شود که در فاصله i تا j فرض شده و با مقادیر خصوصیات i جمع می‌شود. در آخر مقادیر جدید به دست آمده به عنوان مقادیر خصوصیات نمونه مصنوعی جدید در نظر گرفته می‌شوند.

برای انتخاب بهترین ویژگی‌ها برای استفاده در الگوریتم‌های یادگیری ماشین و بهبود کیفیت آن‌ها از روش طبقه‌بندی درختان مازاد (Extra Tree) استفاده شد. الگوریتم درختان مازاد، مانند الگوریتم جنگل‌های تصادفی، درخت‌های تصمیم‌گیری زیادی ایجاد می‌کند، اما نمونه‌گیری برای هر درخت، تصادفی و بدون جایگزینی است (۱۵).

روش‌های یادگیری ماشین:

برای اجرای مدل اصلی با به کارگیری تعدادی از الگوریتم‌های یادگیری ماشین با ناظر و اجرای آن بر روی داده‌هایی که در بخش قبلی بررسی شدند، بروز لخته در خون بند ناف نوزاد پیش از تولد پیش‌بینی شد.

به منظور ارزیابی دقت و کارآیی الگوریتم‌ها در پیش‌بینی، از جمله شاخص‌هایی که مورد بررسی قرار گرفته‌اند شامل دقت (Accuracy) صحت (Precision)، بازخوانی (Recall) و امتیاز F1 (F1 score) هستند که محاسبه هر کدام به شرح زیر است:

tp : الگوریتم نمونه را در دسته مثبت طبقه‌بندی کرده و نمونه هم مثبت است

tn : الگوریتم نمونه را در دسته منفی طبقه‌بندی کرده و نمونه هم منفی است

fp : الگوریتم نمونه را در دسته مثبت طبقه‌بندی کرده اما نمونه منفی است

fn : الگوریتم نمونه را در دسته منفی طبقه‌بندی کرده اما نمونه مثبت است

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

برای تنظیم فراپارامتر (Hyperparameter) در الگوریتم‌های یادگیری ماشینی از روش جستجوی شبکه‌ای (Grid Search) استفاده شد که شامل جستجو در محدوده‌ای از مقادیر فراپارامتر برای یافتن ترکیب بهینه فراپارامترها است که بهترین عملکرد را در یک مجموعه داده معین ایجاد می‌کند (۱۶). فراپارامترهای کلیدی و اندازه‌های آزمایش بر اساس تأثیر قابل توجه آن‌ها بر عملکرد و کارآیی محاسباتی مدل انتخاب شدند، که امکان ایجاد فضای جستجوی قابل مدیریت را فراهم می‌آورد و در عین حال نتایج بهینه‌سازی قوی را تضمین می‌کند. هم‌چنین لازم به ذکر است که پس از فراخوانی داده‌ها و اجرای روش SMOTE ، آن‌ها به دو دسته داده‌های تمرینی (۸۰٪) و داده‌های تست (۲۰٪) تقسیم شدند.

الگوریتم‌های مورداستفاده در این پژوهش شامل الگوریتم‌های زیر است:

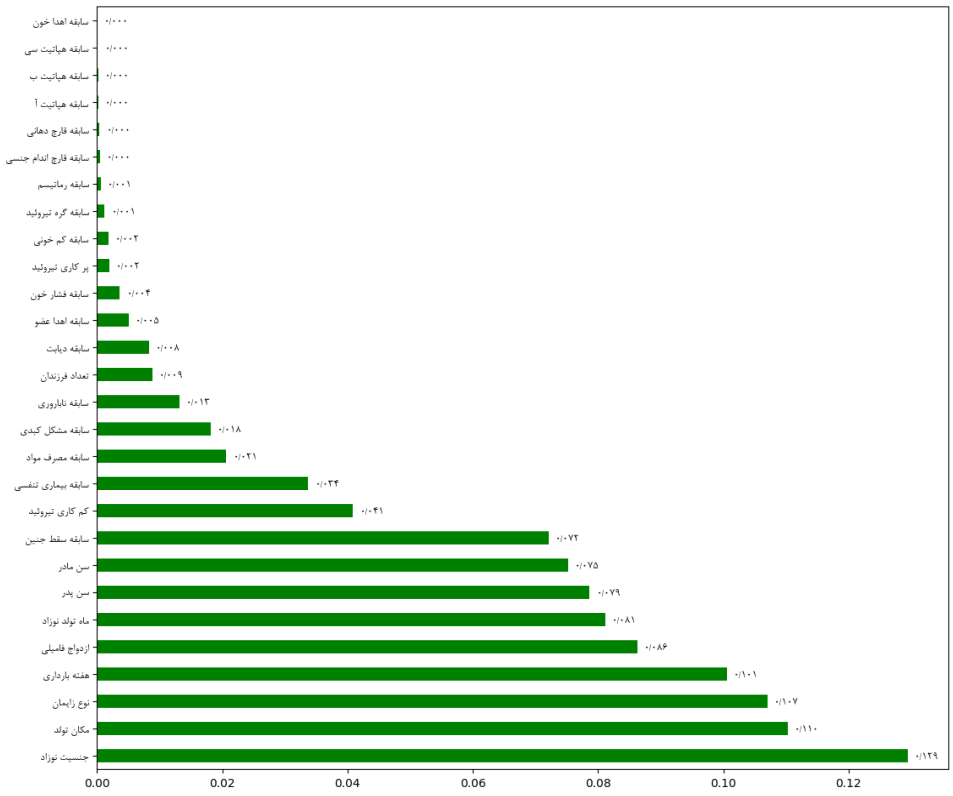
- درخت تصمیم(Decision Tree): درخت تصمیم اغلب برای مسائل دسته‌بندی استفاده می‌شود و به خوبی می‌تواند داده‌های پیچیده را تفکیک کند.
- دسته‌بندی بیز ساده(Naïve Baysian): این مدل بر اساس تئوری احتمال بیز عمل می‌کند و برای داده‌های مستقل از هم مناسب است.
- k–نزدیک‌ترین همسایگی(K-nearest Neighbors): این الگوریتم بر اساس فاصله نزدیک‌ترین همسایه‌ها کار می‌کند و برای مسائل تشخیص الگو مناسب است.
- ماشین بردار پشتیبان(Support Vector Machines): ماشین‌بردار پشتیبان برای مسائل دسته‌بندی خطی و غیرخطی کاربرد دارد.
- جنگل تصادفی(Random Forest):جنگل تصادفی از ترکیب چندین درخت تصمیم برای بهبود دقت و کاهش بیش‌برازش استفاده می‌کند.
- طبقه‌بندی رأی‌گیری اکثریت (Majority-Voting Classifier): این مدل از ترکیب چند مدل مختلف برای بهبود دقت نهایی استفاده می‌کند.
- پرسپترون چندلایه (Multilayer Perceptron): این مدل یکی از انواع شبکه‌های عصبی است که برای مسائل پیچیده و غیرخطی مناسب است (۱۸، ۱۷).

الگوریتم‌های استفاده شده در این پژوهش به دلایل مختلفی از جمله دقت بالا، قابلیت توجیه‌پذیری مدل، قدرت تعمیم‌پذیری، کارآیی در پردازش داده‌های بزرگ و قابلیت انطباق با ساختار داده‌های ورودی انتخاب شده‌اند.

یافته‌ها

پس از متوازن‌سازی داده‌ها با روش SMOTE مجموعه داده‌ها برای انتخاب بهترین ویژگی‌ها آماده شدند. پس از اجرای الگوریتم درختان مازاد می‌توان دید که ویژگی "جنسیت نوزاد" با مقدار وزن ۰/۱۲ بیشترین تأثیر را در مدل دارد. پس از آن، ویژگی‌های "مکان تولد" و "نوع زایمان" به ترتیب با وزن‌های ۰/۱۱ و ۰/۱۰ قرار دارند. ویژگی‌های دیگری مانند "هفته بارداری" و "ازدواج فامیلی" نیز تأثیر قابل توجهی دارند. در مقابل، ویژگی‌هایی مانند "سابقه اهدا خون" و "سابقه هپاتیت C " کمترین اهمیت را در مدل دارند(شکل ۲). شکل ۲ نمایی از اهمیت تمام ویژگی‌ها را نشان می‌دهد که محور افقی میزان اهمیت هر ویژگی و محور عمودی نام ویژگی‌ها را نشان می‌دهد. ویژگی‌ها بر اساس اهمیت و وزن داده شده توسط الگوریتم به صورت نزولی مرتب شده‌اند. بر این اساس ۱۸ ویژگی با بیشترین اهمیت (جنسیت نوزاد تا سابقه فشارخون) برای اجرای الگوریتم‌های بعدی انتخاب می‌شوند.

نتایج نشان می‌دهد که جنگل تصادفی بهترین دقت (۰/۸۴)، بالاترین صحت (۰/۹۵) و امتیاز F1 (۰/۸۴) را به اجرا گذاشته است اما بازخوانی متوسط (۰/۷۳) می‌تواند نشان‌دهنده این باشد که این مدل در شناسایی همه موارد مثبت موثر نیست. مدل k–نزدیک‌ترین همسایه با دقت (۰/۸۳)، صحت (۰/۸۰)، بازخوانی (۰/۸۹) و امتیاز F1 (۰/۸۴) کارآیی بالایی در همه شاخص‌ها را از خود نشان داده است و بیانگر آن است که برای پیش‌بینی لخته‌شدن خون بند ناف مناسب است. رأی‌گیری اکثریت و درخت تصمیم عملکرد نسبتاً خوبی را در همه شاخص‌ها از خود نشان داده است و مشخص می‌کند که می‌توان از آن‌ها نیز



شکل ۲: اهمیت ویژگی‌های مجموعه داده به همراه وزن آن‌ها

جدول ۱: مقایسه نتایج ارزیابی مدل‌های مختلف بر اساس شاخص‌های عملکردی دقت، صحت، بازخوانی و امتیاز F1

نام مدل	امتیاز F1	بازخوانی	صحت	دقت
درخت تصمیم	۰/۸۰	۰/۸۰	۰/۸۱	۰/۸۰
بیزین ساده	۰/۶۸	۰/۷۹	۰/۶۰	۰/۶۳
K–نزدیک‌ترین همسایه	۰/۸۴	۰/۸۹	۰/۸۰	۰/۸۳
ماشین‌بردار پشتیبان	۰/۶۳	۰/۶۰	۰/۶۶	۰/۶۵
جنگل تصادفی	۰/۸۲	۰/۷۳	۰/۹۵	۰/۸۴
رأی‌گیری اکثریت	۰/۸۱	۰/۸۱	۰/۸۴	۰/۸۱
پرسپترون چند لایه	۰/۷۴۱	۰/۷۱	۰/۷۷	۰/۷۴

به عنوان الگوریتم‌های قابل اطمینان برای پیش‌بینی لخته استفاده کرد. پرسپترون چندلایه نتایج نسبتاً متوسطی را در مقایسه با دیگر روش‌ها نتیجه داده است. دسته‌بندی بیز ساده با وجود بازخوانی نسبتا خوب (۰/۷۹) عملکرد ضعیفی در دیگر شاخص‌ها دارد. ماشین بردار پشتیبان اما عملکرد ضعیفی نسبت به سایر مدل‌ها دارد که می‌توان گفت استفاده از این روش برای پیش‌بینی لخته مناسب نیست. بعد از اجرای هر الگوریتم و عیب‌یابی، کارآیی الگوریتم‌ها با هم مقایسه شدند (جدول ۱).

بحث

در این مطالعه مدل یادگیری ماشینی برای پیش‌بینی لخته‌شدن خون بند ناف پیش از تولد نوزاد با استفاده از داده‌های بالینی و آزمایشگاهی طبقه‌بندی شد. شناسایی بند ناف‌هایی که در معرض خطر بروز لخته‌شدن هستند، تأثیر به سزایی در کیفیت خون جمع‌آوری شده به منظور ذخیره‌سازی در بانک‌های بند ناف و هم‌چنین کاهش هزینه جمع‌آوری و آزمایش در این بانک‌ها دارد. استفاده از یادگیری ماشین در پیش‌بینی وضعیت لخته‌شدن خون بند ناف پیش از تولد نوزاد، می‌تواند به این فرآیند کمک کند و امکان مداخله به‌موقع و پیشگیری از عوارض آن را فراهم

نماید.

در همین راستا ابتدا داده‌هایی که در بانک خون بند ناف رویان جمع‌آوری شده بود، پیش‌پردازش شدند و با توجه به این که داده‌ها بالانس نبودند، با کمک روش SMOTE بالانس شدند و سپس بهترین ویژگی‌ها با کمک طبقه‌بندی درختان مازاد انتخاب شدند تا برای الگوریتم‌های یادگیری ماشین آماده شوند. نتایج این روش نشان دادند که جنسیت نوزاد بیشترین اهمیت را در بروز لخته در خون بند ناف پس از زایمان دارد که حیفض نیز در پژوهش خود به آن اشاره کرده است (۵). هم‌چنین با توجه به ادبیات موضوع دیابت، فشار خون نقش مستقیمی در لخته شدن خون بند ناف نوزاد دارد که این پژوهش با انتخاب آن به عنوان بهترین ویژگی‌ها این موضوع را تصدیق می‌کند (۷، ۴).

هم‌چنین فاکتورهای تاثیرگذار جدید توسط مدل درخت مازاد شناسایی شدند که پیش از این در ادبیات موضوع شناسایی نشده بودند. مکان تولد نوزاد، نوع زایمان، هفته بارداری و ازدواج فامیلی، به ترتیب با ۰/۱۱، ۰/۱، ۰/۱ و ۰/۸۶ وزن از دیگر عوامل مهم در پیش‌بینی لخته‌شدن خون بند ناف هستند.

مدل با استفاده و مقایسه الگوریتم‌های طبقه‌بندی یادگیری با نظارت درخت تصمیم، بیزین ساده، k-نزدیک‌ترین همسایه، ماشین‌بردار پشتیبان، جنگل تصادفی، رأی‌گیری اکثریت و پرسپترون چندلایه اجرا شد.

نتیجه‌گیری

عملکرد بالای دو روش جنگل تصادفی و k-نزدیک‌ترین همسایه با دقت‌های به ترتیب (۰/۸۴) و (۰/۸۳) نشان می‌دهد که می‌توان با کمک الگوریتم‌های یادگیری ماشین با دقت بالایی لخته‌شدن خون بند ناف نوزاد را پیش از زایمان پیش‌بینی کرد و به کمک آن از نمونه‌برداری نمونه‌های دارای لخته به‌منظور کاهش هزینه و مشکلات ذخیره‌سازی آن‌ها جلوگیری کرد. عملکرد نسبتاً متوسط الگوریتم پرسپترون چندلایه در مقایسه با روش‌های اشاره شده نشان می‌دهد که در جایی که روش‌های کلاسیک یادگیری ماشین عملکرد بالایی از خود نشان می‌دهند، نیازی به استفاده از روش‌های شبکه‌های عصبی عمیق نیست. هم‌چنین عملکرد نسبتاً پایین دو روش بیزین ساده و ماشین‌بردار پشتیبان با دقت‌های به ترتیب ۰/۶۳ و ۰/۶۵، نقش انتخاب روش مناسب برای پیش‌بینی لخته‌شدن و اهمیت مقایسه روش‌های مختلف را نشان می‌دهد.

پیشنهاد می‌شود در پژوهش‌های آتی، از روش‌های دیگری برای انتخاب ویژگی استفاده شود. هم‌چنین سابقه بیماری‌های بیشتری برای تکمیل داده‌ها در نظر گرفته شود. روش‌های پیچیده‌تر تنظیم هایپرپارامترها نیز بررسی شده تا نتایج با دقت پیش‌بینی بالاتری ارائه شوند.

حمایت مالی

مطالعه فوق بدون حمایت مالی ارگان و نهاد خاصی انجام شده است.

ملاحظات اخلاقی

این پروژه از کمیته اخلاق در پژوهش دانشگاه خوارزمی با کد اخلاق IR.KHU.REC.1402.068 در تاریخ ۱۴۰۲/۸/۲ مجوز گرفته است.

عدم تعارض منافع

نویسندگان اظهار کردند در انتشار این اثر منافع تجاری نداشتند و در مقابل ارائه اثر وجهی دریافت نکرده‌اند.

نقش نویسندگان

امیر حسین اسماعیل‌پور: تحلیل و بررسی داده‌ها، اجرای روش‌ها، نوشتن مقاله

دکتر مریم عاملی: نظارت بر اجرای روش‌ها، روش‌شناسی و تفسیر نتایج، ویرایش مقاله

دکتر اشکان مزدگیر: طراحی مطالعه، روش‌شناسی، ویرایش مقاله

دکتر اُرد احمدی: بررسی روش‌ها و نتایج

دکتر مرتضی ضرابی: فراهم آوردن داده‌های مورد نیاز، طراحی مطالعه

References:

1- Pezeshki SMS, Ghasemzadeh M, Hosseini E. Cord blood stem cells: an overview of biology and current applications. Sci J Iran Blood Transfus Organ 2023; 20(4): 335-45. [Article in Farsi]
2- Niazi V, heydari Keshel S, Shahbazi M. Advances and challenges in storage, transplantation, expansion and homing of Umbilical Cord Blood Hematopoietic Stem Cells (UCB-HSCs). Sci J Iran Blood Transfus Organ 2020; 17(3): 226-41. [Article in Farsi]
3- Zhu D, Barabadi M, McDonald C, Kusuma G, Inocencio IM, Lim R. Implications of maternal-fetal health on perinatal stem cell banking. Gene Ther 2024; 31(3): 65-73.
4- Fritz MA, Christopher CR. Umbilical vein thrombosis and maternal diabetes mellitus. J Reprod Med 1981; 26(6): 320-4.
5- Heifetz SA. Thrombosis of the Umbilical Cord: Analysis of 52 Cases and Literature Review. Pediatr Pathol 1988; 8(1): 37-54.
6- Schröcksnadel H, Holböck E, Mitterschiffthaler G, Tötsch M, Dapunt O. Thrombotic occlusion of an umbilical vein varix causing fetal death. Archives of Gynecology and Obstetrics. 1991;248(4):213-5.
7- Lox CD, Word RA, Jeter M, Corrigan JJ. Cord blood coagulation changes in maternal hypertensign. Pediatr Res 1984; 18(4): 317.
8- Funk A, Buechel J, Huhn EA, Mueller D, Granado C, Tsakiris D, *et al*. Antenatal predictors of stem cell content for successful umbilical cord blood donation. Arch Gynecol Obstet 2021; 304(2): 377-84.
9- Haghbayan MH, Karimi B, Mozdgir A, Abbaspanah B. Increasing the Efficacy of Umbilical Cord Blood Banking Using Machine Learning Algorithms: A Case Study from Royan Cord Blood Bank. Scientia Iranica 2023; 1-26. [Article in press]
10- Hare J, DeLeon PG, Pool K, Reioux D, Fontenot M, Champlin RE, *et al*. Optimal umbilical cord blood collection, processing and cryopreservation methods for sustained public cord blood banking. Cytotherapy 2021; 23(11): 1029-35.
11- Jamshidi R, Rajabpour Sanati S, Zarrabi M. A New Method to Predict the Quality of Umbilical Cord Blood Units based on Maternal and Neonatal Factors and Collection Techniques. Journal of Applied Research on Industrial Engineering 2023; 10(2): 218-37.
12- Raschka S, Patterson J, Nolet C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information 2020; 11(4): 93.
13- Barbieri MC, Grisci BI, Dorn M. Analysis and comparison of feature selection methods towards performance and stability. Expert Systems with Applications 2024; 249: 123667.
14- Kaur H, Pannu HS, Malhi AK. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Comput Surv 2019; 52(4): 1-36.
15- Alfian G, Syafrudin M, Fahrurrozi I, Fitriyani NL, Atmaji FT, Widodo T, *et al*. Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. Computers 2022; 11(9): 136.
16- Mezzatesta S, Torino C, Meo PD, Fiumara G, Vilasi A. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Comput Methods Programs Biomed 2019; 177: 9-15.
17- Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, *et al*. Causal machine learning for predicting treatment outcomes. Nat Med 2024; 30(4): 958-68.
18- Asnicar F, Thomas AM, Passerini A, Waldron L, Segata N. Machine learning for microbiologists. Nat Rev Microbiol 2024; 22(4): 191-205.