

## مقایسه کارایی الگوریتم‌های یادگیری ماشین برای ایجاد یک مدل پیش‌بینی کننده مرگ و میر بیماران بستری مبتلاء به کووید-۱۹

مصطفی شنبه‌زاده<sup>۱</sup> (Ph.D)، علی ولی‌نژادی<sup>۲</sup> (Ph.D)، رامین افراه<sup>۳</sup> (M.Sc)، هادی کاظمی آرپناهی<sup>۴</sup> (Ph.D)، اعظم اروجی<sup>۵</sup> (Ph.D)، محمدرضا کفاشیان<sup>۶</sup> (Ph.D)

۱- گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی ایلام، ایلام، ایران

۲- گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی سمنان، سمنان، ایران

۳- گروه مهندسی پزشکی، دانشکده فناوری‌های نوین، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران

۴- گروه فناوری اطلاعات سلامت، دانشگاه علوم پزشکی آبادان، آبادان، ایران

۵- گروه فناوری‌های نوین، دانشکده پزشکی، دانشگاه علوم پزشکی خراسان شمالی، بجنورد، ایران

۶- گروه فیزیولوژی، دانشکده پزشکی، دانشگاه علوم پزشکی ایلام، ایلام، ایران

تاریخ دریافت: ۱۳۹۹/۱۲/۱۱ تاریخ پذیرش: ۱۴۰۰/۴/۱۲

نویسنده مسئول، تلفن: ۰۹۱۳۸۲۰۰۰۲۷ h.kazemi@abadanums.ac.ir

### چکیده

هدف: شیوع سریع ویروس SARS-CoV-2 در سراسر دنیا، سیستم‌های مراقبت سلامت را با چالش‌های جدی و غیر منتظره در پیش‌بینی رفتار و پیامدهای بیماری روبرو کرده است. برای غلبه بر این چالش‌ها و ابهامات، هدف مطالعه حاضر ایجاد و اعتبارسنجی چند مدل پیش‌بینی مبتنی بر تکنیک‌های یادگیری ماشین به منظور تعیین ریسک مرگ بیماران بستری شده مبتلاء به کووید-۱۹ و انتخاب بهترین مدل می‌باشد.

مواد و روش‌ها: داده‌های ۱۲۲۴ بیمار بستری ثبت شده با تشخیص قطعی کووید-۱۹ از پایگاه داده سامانه ثبت بیماران کووید-۱۹ شهر ایلام استخراج شدند. سپس پارامترهای تاثیرگذار در وقوع مرگ بیماران کووید-۱۹ شناسایی و به عنوان ورودی الگوریتم‌های یادگیری ماشین منتخب شامل کای- نزدیک‌ترین همسایه (k-NN)، ماشین بردار پشتیبان (SVM)، رگرسیون لجیستیک (LR) و جنگل تصادفی (RF) استفاده شدند. در نهایت عملکرد مدل‌های طراحی شده بر اساس معیارهای ارزیابی بر گرفته از ماتریکس آشفتگی (Confusion Matrix) مورد مقایسه قرار گرفت و مناسب‌ترین مدل پیش‌بینی کننده شناسایی گردید. یافته‌ها: پارامتر به عنوان متغیرهای تاثیرگذار در مرگ و میر کووید-۱۹ شناسایی شدند. پس از اندازه‌گیری و مقایسه عملکرد الگوریتم‌های یادگیری ماشین، الگوریتم k-NN با دقت ۹۴/۲۱٪، صحت ۹۳/۷۴٪، فراخوانی ۱۰۰٪، معیار اف ۹۳/۲٪ و سطح زیر نمودار ROC ۹۲٪/۲۳٪ عملکرد بهتری را به دست آورد.

نتیجه‌گیری: الگوریتم KNN قادر به پیش‌بینی خطر مرگ و میر بیماران کووید-۱۹ با یک سطح مناسب از صحت و اطمینان به منظور پیش‌بینی موثر افراد پرخطر و انتخاب مداخله مناسب توسط متخصصین پزشکی است.

واژه‌های کلیدی: کووید-۱۹، کرونا ویروس، هوش مصنوعی، یادگیری ماشین، داده کاوی

### مقدمه

بیمار متفاوت است [۲،۱]. همه‌گیری این بیماری در سال ۲۰۲۰ به یک ابر مشکل مهم جهانی تبدیل شد و بسیاری از نظام‌های مراقبتی و بهداشت عمومی را در سراسر جهان تا مرز فروپاشی کامل پیش برد [۳].

در این میان ارتقاء پتانسیل بهداشت عمومی در مبارزه با ویروس کووید-۱۹ نیازمند توجه به این مقوله است که عامل ایجادکننده مشکلات و مخاطرات تهدیدکننده وضعیت بهداشت عمومی و نیز راه‌حل‌های برطرف‌کننده این مشکلات باید از یک سیستم انطباقی فناورانه و پیچیده پیروی کند [۵،۴]. از این

بیماری جدید کروناویروس ۲۰۱۹ (COVID-19) یک بیماری ویروسی عفونی مسری و منتشره می‌باشد که عامل بیماری‌زایی آن ویروسی به نام کروناویروس سندرم حاد تنفسی ۲ (SARS-CoV-2) است. تظاهرات بالینی بیماری از موارد بدون علامت تا ظهور علائم خفیف مشابه سرماخوردگی و عفونت‌های حاد دستگاه تنفسی فوقانی تا آسیب‌های جدی به دستگاه تنفسی، ایجاد ذات‌الریه، نارسایی اندام‌ها و حتی مرگ

به عنوان سازمان مراقبتی مرکزی و اصلی به منظور تشخیص و درمان بیماری کووید-۱۹ تابعه دانشگاه علوم پزشکی ایلام انجام گردید. هدف مطالعه حاضر ایجاد یک مدل پیش‌گویانه برای پیش‌آگهی در خصوص مرگ و میر بیماران مبتلاء و بستری شده در بیمارستان از طریق تکنیک‌های یادگیری ماشین و شناسایی بهترین مدل بود. مطالعه حاضر در پنج بخش اصلی شامل توصیف مجموعه داده، پیش‌پردازش، انتخاب ویژگی، ایجاد مدل و ارزیابی آن انجام شد. لازم به ذکر است که مطالعه حاضر با کد اخلاق IR.ABADANUMS.REC.1399.128 به عنوان یک طرح تحقیقاتی مشترک بین دانشگاه علوم پزشکی آبادان و ایلام مورد تایید قرار گرفت.

داده‌های مجموع تعداد ۱۳۰۰ بیمار قطعی مبتلاء به بیماری کووید-۱۹ اثبات شده از طریق آزمون Rt-PCR که در بیمارستان مصطفی خمینی شهر ایلام بستری شده‌اند و در سیستم ثبت بیماری‌های کووید-۱۹ (Ilam CoV Reg) این بیمارستان بین ۲۷ اسفند ۱۳۹۸ لغایت ۲۰ آذر ۱۳۹۹ ثبت شدند، به عنوان جامعه پژوهش انتخاب گردید. در مجموع در این مدت حدود ۱۵۵۸۹ مراجعه‌کننده مشکوک به کووید-۱۹ به بخش‌های سرپایی و اورژانس این بیمارستان مراجعه کردند. از این تعداد، ۴۸۵۹ بیمار تحت آزمایش Rt-PCR قرار گرفتند و آزمایش ۲۷۸۶ مورد از آن‌ها مثبت اعلام شد. نهایتاً ۱۳۰۰ بیمار به دلیل وجود علائم شدیدتر بیماری در بیمارستان بستری شدند. داده‌های افراد بستری شده در سیستم ثبت بیماری کووید-۱۹ در این بیمارستان در قالب کلاس‌های داده‌ای پایه / عمومی، علائم و نشانه‌ها، بیماری‌های زمینه‌ای، نتایج آزمایشگاهی و پیامد درمان (کد صفر زنده و یک فوتی) در قالب ۵۴ متغیر بالینی در Ilam CoV Reg ثبت و به عنوان ورودی تکنیک‌های یادگیری ماشین مورد استفاده قرار گرفتند. استخراج اطلاعات از پایگاه داده این بیماری به واسطه مراجعه حضوری پژوهشگران و ثبت در قالب فایل برنامه کاربردی Excel انجام پذیرفت. در مواقع بروز ابهام در خصوص داده‌ها، داده‌های ناقص، تناقضات و عدم توافق از پزشک مسئول کیفی ثبت داده‌ها در سیستم ثبت استفاده شد.

داده‌های بیماران مبتلاء قطعی بستری شده که بالاتر از ۱۸ سال بودند به عنوان ورودی به مطالعه انتخاب شدند. از این میان رکوردهایی با بیش از ۷۰ درصد از دست رفتگی از مطالعه خارج شدند. برای سایر رکوردهای میسینگ سعی شد تا از مقادیر میانگین و مد برای پرکردن موارد خالی استفاده شود.

به منظور ارتقاء عملکرد پیش‌بینی مدل‌های داده کوی از روش هم‌بستگی پیرسن برای متغیرهای کمی، هم‌بستگی اسپیرمن برای انواع رتبه‌ای و آزمون فی برای متغیرهای اسمی

رو استفاده از فنون و مدل‌های پیش‌بینی‌کننده به منظور بررسی روندها و رویدادهای مرتبط با شیوع بیماری، پیش‌آگهی ابتلاء و پیش‌بینی خطرات بدحال شدن و مرگ و میرهای ناشی از آن در مراحل اولیه؛ در شناسایی سریع، جداسازی و قرنطینه موارد مبتلاء و اعمال نظارت‌های دقیق و موثر و نیز پیشگیری از همه‌گیری این بیماری موثر است. به علاوه از طریق بهبود اولویت‌بندی بیماران و ارائه درمان‌های سفارشی و مبتنی بر نیاز بیمار در کاهش ناتوانی‌ها و مرگ‌های ناشی از آن ضروری است [۶،۷].

یادگیری ماشین (Machine learning) از طریق کاوش داده‌های بزرگ (Big data mining) از جمله مهم‌ترین فناوری‌های نوین و پیشرفته امروزی است که نقش بسزایی در کنترل و مدیریت همه‌گیری کووید-۱۹ و نیز پاندمی‌های گذشته داشته است [۹،۸]. تاکنون پژوهش‌های متعددی در زمینه کاربرد یادگیری ماشین در مدیریت و کنترل بیماری کووید-۱۹ با هدف نظارت و پیش‌بینی روندهای بیماری و ارزیابی‌های همه‌گیری، ارائه مدل‌های تشخیصی، تعیین بهترین مدل درمانی و مراقبتی و پیش‌بینی بهینه بیماری و عوارض آن از جمله بدحال شدن یا فوت بیمار انجام پذیرفته است [۱۰-۱۶]. این الزام با توجه به ماهیت پیچیده، حساس، و مبهم این بیماری و غیرقابل پیش‌بینی بودن رفتار آن، نبود درمان قطعی، طول دوره کمون و بهبودی بالا، ارزیابی‌های پاراکلینیکی نسبتاً دشوار و وجود تشخیص افتراقی با سایر بیماری‌ها، نقش مهمی در بهبود کیفیت تشخیص و ارائه مراقبت‌های سفارشی و بیمار محور دارد [۱۷-۱۹].

نتایج مطالعه‌ی مروری توسط Albahri و همکاران در سال ۲۰۲۰ نشان داد که استفاده از الگوریتم‌های یادگیری ماشین به منظور ارائه مدل‌های تشخیصی و پیش‌بینی‌کننده با هدف شناسایی به موقع، اثربخش و اقتصادی بیماری کووید-۱۹، در کاهش شیوع بیماری و عوارض مخرب آن بسیار موثر است [۲۰]. بنابراین با توجه به شیوع بالای بیماری و مرگ‌های ناشی از آن در کشورمان و وجود برخی محدودیت‌ها و کمبودهای منابع مراقبتی [۲۱،۲۲]، از این‌رو هدف از انجام پژوهش حاضر پیشنهاد یک مدل موثر و کارآمد فناورانه بر اساس یادگیری ماشین به منظور پیش‌بینی مرگ و میر بیماران بستری مبتلاء به کووید-۱۹ بر اساس مقایسه عملکرد الگوریتم‌های منتخب یادگیری ماشین و شناسایی مهم‌ترین متغیرهای بالینی تاثیرگذار بود.

## مواد و روش‌ها

مطالعه گذشته‌نگر حاضر در قالب یک بررسی توصیفی- کاربردی در سال ۱۳۹۹ در بیمارستان مصطفی خمینی شهر ایلام

جدول ۱. نحوه محاسبه معیارهای عملکردی شش گانه مورد استفاده

معیارهای ارزیابانه	فرمول
صحت	$\frac{TP + TN}{TP + FP + TN + FN}$
دقت	$\frac{TP}{TP + FP}$
فراخوانی	$\frac{TP}{TP + FN}$
ضریب اف	$2 * \left( \frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \right)$

جدول ۲. ماتریس آشفتگی

خروجی	مقدار پیش‌بینی		
	مرگ (+)	زنده ماندن (-)	
مقدار واقعی	مرگ (+)	TP	FN
	زنده ماندن (-)	FP	TN

True positive (TP): تعداد افرادی فوتی که مدل به درستی آنها را تشخیص داده است.

False positive (FP): تعداد افرادی که زنده هستند اما مدل به اشتباه آنها را فوتی تشخیص داده است.

True negative (TN): تعداد افرادی که زنده هستند و مدل به درستی آنها را زنده تشخیص داده است.

False negative (FN): تعداد افرادی که فوتی هستند، اما مدل آنها را به اشتباه زنده تشخیص داده است (۲۴).

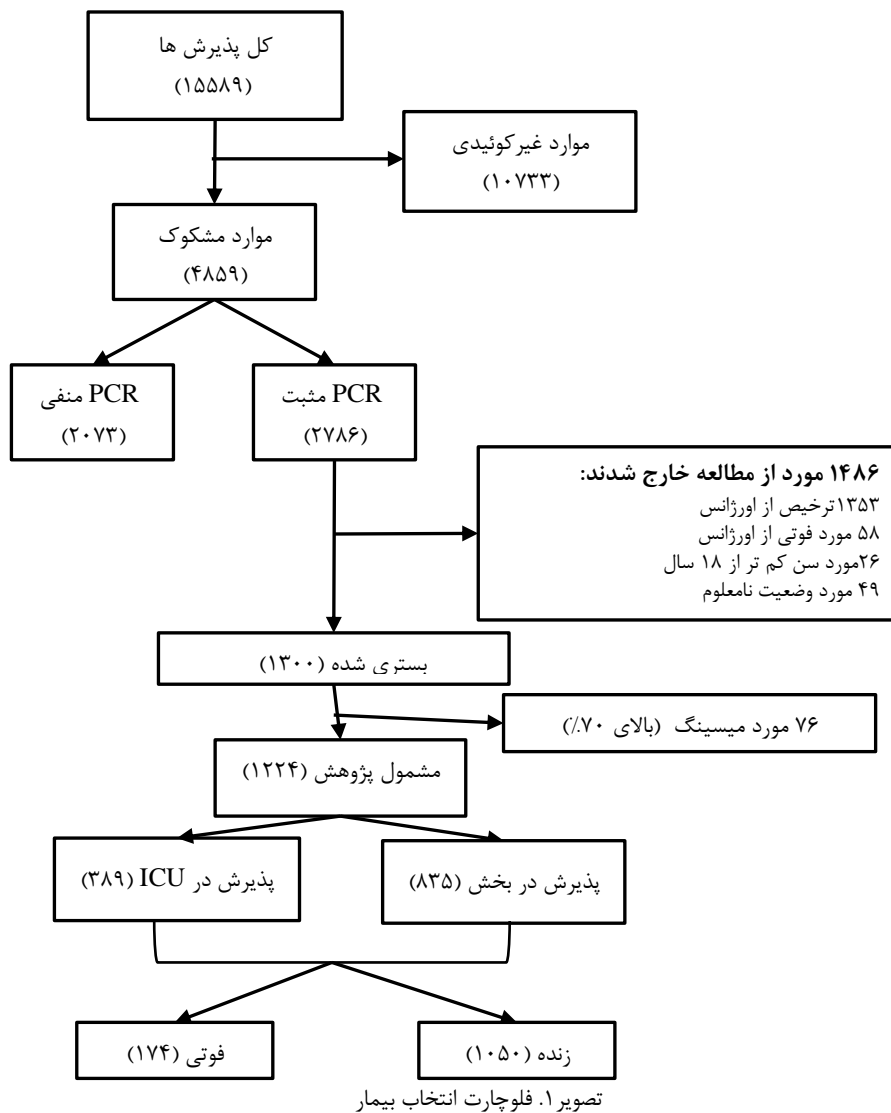
## نتایج

در بازه زمانی حدود ۱۰ ماه از شروع همه‌گیری بیماری و گذران کردن تعداد سه موج از شیوع سنگین بیماری که باعث بیماری و مرگ بسیاری شد، در مجموع ۱۵۵۸۹ بیمار به بخش‌های سرپایی و اورژانس بیمارستان مصطفی خمینی شهر ایلام به منظور ارزیابی و مراقبت بیماری کووید-۱۹ مراجعه کردند. از این تعداد ۱۳۰۰ بیمار قطعی در بیمارستان بستری و اطلاعات آن‌ها در Ilam CoV Reg ثبت شد. معیارهای ورود به مطالعه شامل موارد قطعی ابتلاء به کووید-۱۹ بر اساس نتایج آزمایش Rt-PCR در بازه زمانی بین ۲۷ اسفند ۱۳۹۸ لغایت ۲۰ آذر ۱۳۹۹ بود که در بیمارستان خدمات بستری را دریافت کردند. در این میان نمونه‌هایی با وضعیت درمانی ناشناخته، سن کم‌تر از ۱۸ سال، ترخیص / فوت شده از واحدهای سرپایی و فیلم‌های اطلاعاتی با از دست رفتگی بالای اطلاعات از مطالعه حذف شدند. پس از اعمال معیارهای خروج مطالعه و حذف رکوردهای اطلاعاتی با بیش از ۷۰ درصد مفقودی، در نهایت اطلاعات ۱۲۲۴ مورد بستری شده وارد مطالعه شد (شکل ۱).

به منظور تعیین روابط بین هر یک از متغیرهای موثر در بروز مرگ و میر ناشی از کووید-۱۹ و انتخاب ویژگی‌ها با توجه به حالات دو گانه کلاس خروجی (۰ برای زنده و ۱ برای فوتی) استفاده شد. در این پژوهش سطح معناداری  $P\text{-value} < 0.05$  در نظر گرفته شد و عواملی که رابطه معنادار با متغیر خروجی داشتند به عنوان فاکتورهای خطر موثر در پیش‌بینی شناسایی و وارد مدل نهایی شدند. پس از انتخاب متغیرهای مهم‌تر و با توجه به واحد متغیرهای مورد بررسی، به منظور قابل مقایسه کردن قدرت رابطه هر کدام از متغیرها با مرگ و میر از ضریب هم‌بستگی استاندارد شده استفاده شد. از نرم‌افزار آماری SPSS نسخه ۲۳ جهت تجزیه و تحلیل داده‌ها استفاده شد.

پس از انجام بررسی‌های مروری و ارزیابی مطالعات انجام شده در زمینه کاربرد یادگیری ماشین برای پیش‌بینی وضعیت‌های مرتبط با بیماری کووید-۱۹ و نیز با در نظر داشتن نوع و کیفیت داده‌های مجموعه داده انتخابی، در نهایت الگوریتم‌های کی-نزدیک‌ترین همسایه (KNN)، ماشین بردار پشتیبان (SVM)، رگرسیون لجیستیک (LR) و جنگل تصادفی (RF) به عنوان الگوریتم‌های انتخابی برای مطالعه حاضر در نظر گرفته شدند. هم‌چنین، از نرم‌افزار Weka v3.9.2 برای تحلیل و محاسبه منحنی‌ها، معیارها و ماتریس آشفتگی استفاده شد.

متغیرهای خروجی در پژوهش حاضر شامل حالت دوتایی: زنده ماندن بیماران بستری مبتلا و ۱ فوت بیماران. برای ارزیابی عملکرد الگوریتم‌های پیش‌بینی‌کننده از روش fold cross validation 10 و شاخص‌های صحت (accuracy)، دقت (precision)، فراخوانی (recall)، ضریب اف (f-measure) و سطح زیر منحنی راک (Receiver Operating Characteristic (ROC)) استفاده شد. این معیارها با استفاده از مولفه‌های ماتریس آشفتگی (confusion matrix) تعریف و محاسبه شدند. منظور از معیار صحت نسبت تعداد افراد زنده فوت شده است که زنده یا فوت شدن آن‌ها به درستی تشخیص داده شده است. دقت یعنی تعداد افرادی که فوت شده‌اند و مدل به درستی آن‌ها تشخیص داده است. فراخوانی به معنی نسبتی از افراد فوت‌شده هستند که مدل به درستی آن‌ها را فوتی تشخیص داده است. بنابراین هرچه این مقدار بزرگ‌تر باشد، نشان‌دهنده دقت در تشخیص افراد فوت شده است [۲۳] (جدول ۱ و ۲).



مرگ و میر ناشی از بیماری کووید-۱۹ با توجه به کلاس‌های دوگانه (۱ برای فوتی و صفر برای زنده) از طریق ضریب همبستگی و استفاده از نرم‌افزار آماری SPSS در جدول ۴ نشان داده شده است. برای این منظور ویژگی‌هایی با مقدار P-value کم‌تر از ۰/۰۵ به عنوان متغیرهای پیشنهادی برای پیش‌بینی مرگ و میرهای ناشی از کووید-۱۹ شناسایی و وارد مدل نهایی شدند.

بر اساس جدول ۴، ۱۷ متغیر در پیش‌بینی مرگ و میر ناشی از کووید-۱۹ مهم‌تر بودند و به عنوان متغیرهای ورودی جهت طراحی مدل پیش‌بینی مرگ و میر انتخاب شدند. در ادامه نسبت شانس هر کدام از متغیرهای نهایی شده و ضریب استاندارد آن‌ها از طریق تحلیل همبستگی استاندارد شده به منظور قابل مقایسه کردن قدرت رابطه هر کدام از متغیرهای ۱۷ گانه با بروز مرگ و میر به علت کووید-۱۹ در بیماران بستری نمایش داده شده است.

از مجموع ۱۳۰۰ رکورد اطلاعاتی که معیارهای ضروری ورود به مطالعه را مهیا ساختند، تعداد ۷۶ رکورد اطلاعاتی (۵/۸۵ درصد) که بیش از ۷۰ درصد داده آن‌ها از دست رفته بود از مطالعه حذف شد. در نهایت تعداد ۱۲۲۴ مورد به عنوان رکورد نهایی وارد مطالعه شد. برای سایر موارد فیلدهای خالی سعی شد تا از طریق جایگزینی آن‌ها با معیارهای میانگین یا میانه، مشکلات ناشی از وجود فیلدهای خالی تا حدودی برطرف شود. از میان جامعه انتخابی، ۶۹۸ مورد (۵۷/۰۲ درصد) مذکر و ۵۲۶ بیمار (۴۲/۹۸ درصد) مونث و میانگین سن جامعه پژوهش ۵۷/۲۵ سال بود. از طرفی ۱۰۵۰ رکورد برای زنده و ۱۷۴ رکورد برای فوتی‌ها بود. در جدول ۳ اطلاعات آمار توصیفی متغیرهای پژوهش وارد شده است (جدول ۳).

موارد بیماری در سامانه Ilam CoV reg در قالب ۵۴ متغیر بالینی ثبت شدند که به منظور ارتقاء کیفیت مدل‌سازی سعی شد تا از طریق تحلیل‌های آماری متغیرهای کم اهمیت از مطالعه حذف شوند. نتایج شناسایی مهم‌ترین معیارهای پیش‌بینی‌کننده

معیارهای مهم‌تر برای پیش‌بینی مرگ‌ومیرها مورد توجه قرار گرفته شد. از این‌رو نتایج طبقه‌بندی حاصل از عملکرد الگوریتم‌ها بر روی ویژگی‌های مستخرج نشان داد که الگوریتم KNN با مقدار سطح زیر نمودار رآک ۹۲/۲۳ درصد و مقدار فراخوانی ۱۰۰ درصد بهترین عملکرد را به نمایش گذاشت و به عنوان بهترین مدل جهت طبقه‌بندی خطرات مرگ و میر بیماران معرفی گردید. برای اطمینان از این انتخاب، تعداد همسایگی موثر (تعداد همسایگی با عملکرد برتر) با اعمال روش ۱۰ Fold و بر اساس معیار فراخوانی، برای الگوریتم KNN محاسبه شد (تصویر ۳).

بر اساس نتایج حاصل الگوریتم KNN با تعداد همسایگی "یک"، بر اساس معیار فراخوانی بهترین عملکرد را داشت و این در حالی است که به تدریج با افزایش تعداد همسایگی‌ها از کیفیت عملکرد آن کاسته شد. در نهایت تعداد همسایگی "یک" را به عنوان موثرترین تعداد همسایگی با اعمال روش ۱۰ Fold ماتریس آشفتگی جهت طبقه‌بندی در نظر گرفته شد.

بر اساس نتایج حاصل از جدول ۵، متغیرهای بیماری‌های قلبی عروقی (۰/۳۷۸)، پذیرش در ICU (۰/۳۷۸)، سن (۰/۳۸۹) و لوله‌گذاری تنفسی (۰/۳۹۷) بیش‌ترین تاثیر را در پیش‌بینی صحیح مرگ و میر داشتند.

در بخش زیر نتایج حاصل از طبقه‌بندی ۱۲۲۴ مورد به منظور پیش‌بینی مرگ‌ومیرهای بیماران بستری مبتلاء به کووید-۱۹ به تفکیک برای الگوریتم‌های KNN، SVM، LR و RF با اعمال روش ارزیابی 10 Fold ارائه شده است (تصویر ۲). نتایج طبقه‌بندی الگوریتم‌های یادگیری ماشین در پیش‌بینی مرگ و میرهای ناشی از کووید-۱۹ در جدول شماره ۶ به صورت مقایسه‌ای ارائه شده است.

نتایج حاصل از طبقه‌بندی مدل‌های یادگیری منتخب حاکی از برتری الگوریتم KNN در پیش‌بینی مرگ و میرهای ناشی از کووید-۱۹ بود. به دلیل هم‌بستگی معکوس معیار فراخوانی با تعداد تشخیص‌های منفی کاذب، در نتیجه با افزایش آن تعداد تشخیص‌های منفی کاذب نیز کاهش خواهد یافت. بنابراین دو معیار ارزیابانه سطح زیر نمودار رآک و فراخوانی به عنوان

جدول ۳. آمار توصیفی متغیرهای کمی و کیفی پژوهش گروه خونی

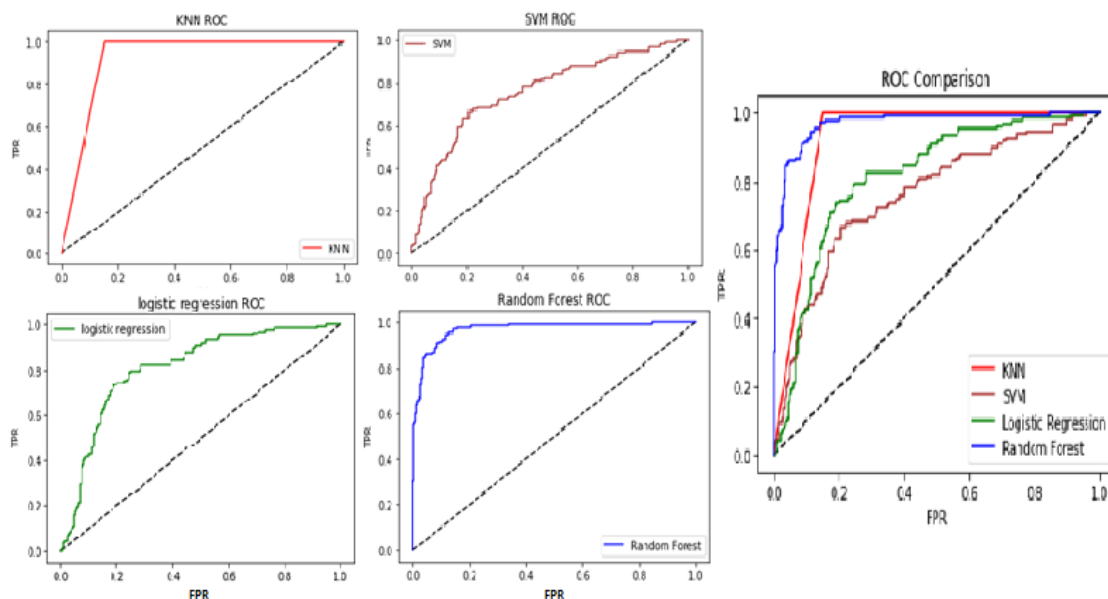
متغیر (کمی)	دامنه	میانگین (انحراف معیار)	متغیر (کیفی)	ارزش‌ها	فراوانی
سن	۱۰-۱۸	۵۷/۲۵ (۱۷/۸)	جنس	مذکر، مونث	۶۹۸، ۵۲۶
قد	۱/۲-۶۶۱/۸۸	۱/۰۲ (۱/۳۶)	حالت تهوع / استفراغ	بلی، خیر	۴۳۴، ۷۹۰
وزن	۱۰۹-۴۶	۶۵/۲۳ (۴/۰۶)	سرفه خشک	بلی، خیر	۵۸۰، ۶۴۴
گروه خونی	۳۶/۴۳-۴/۲	۳۸/۰۴ (۱/۰۲)	تب / لرز	بلی، خیر	۴۵۲، ۷۷۲
کراتینین خون	۰/۱۷-۱/۹	۱/۳۹ (۱/۴)	تنگی نفس	بلی، خیر	۳۹۶، ۸۲۸
شمارش گلبول‌های قرمز	۱/۱۳-۶۱	۴/۶۵ (۱/۶۳)	گلودرد / احتقان گلو	بلی، خیر	۶۹۲، ۵۳۲
شمارش گلبول‌های سفید	۶۳۰۰-۱۳۰۰	۸۱۸۲/۳۴ (۴/۸۹)	کاهش حس بویایی	بلی، خیر	۲۸۳، ۹۴۱
زمان نسبی ترومبوپلاستین	۱۲۰-۱	۲۷/۹ (۱۲/۰۲)	کاهش حس چشایی	بلی، خیر	۷۱۴، ۵۱۰
هموگلوبین	۵/۴۶-۱	۱۳/۳۳ (۲/۳۳)	سر درد	بلی، خیر	۶۳۹، ۵۸۵
شمارش لخته	۶۹۱۰۰۰-۱۰۸۰۰۰	۲۴۵۴۹۳/۶۶ (۸۸۳۸۰/۱)	صرع	بلی، خیر	۱۱۱۲، ۱۱۲
تعداد مطلق لنفوسیت‌ها	۵۹-۲	۲۳/۷۴ (۱/۱۸)	آبریزش بینی	بلی، خیر	۸۵۶، ۳۶۸
تعداد مطلق نوتروفیل‌ها	۹۸-۸	۷۴/۵۲ (۱۲/۳)	درد عضلانی	بلی، خیر	۹۰۴، ۳۲۰
کلسیم	۰/۱۲-۹/۷	۹/۷۲ (۰/۸۲)	احساس ضعف و خستگی	بلی، خیر	۲۰۲، ۱۰۲۲
فسفر	۲/۱۰-۱۳/۷	۳/۴۵ (۰/۶۳)	سابقه فشار خون	بلی، خیر	۲۳۰، ۹۹۴
منیزیم	۱/۳-۱۴/۷	۲/۱۴ (۰/۳۱)	سابقه بیماری‌های قلبی - عروقی	بلی، خیر	۸۹، ۱۱۳۵
سدیم	۳/۱۵۷-۹	۱۳۷/۹ (۵/۳۳)	سابقه دیابت	بلی، خیر	۲۰۳، ۱۰۲۱
پتاسیم	۲/۴۰-۵	۴ (۱/۵۸)	سابقه بدخیمی	بلی، خیر	۲۶، ۱۱۹۸
نیترژن اوره خون	۰/۲۵۱-۵	۴۲/۵۲ (۳/۱۷)	سابقه بیماری‌های کلیوی	بلی، خیر	۱۰۶، ۱۱۱۸
بیلی روبین کامل	۰/۶-۱/۱	۰/۶ (۰/۵۷)	سابقه بیماری‌های تنفسی	بلی، خیر	۴۵۲، ۷۷۲
آسپاراتات آمینوترانسفراز	۳/۹۲۴-۸	۴۴/۴۵ (۵۳/۵)	سابقه ذات‌الریه	بلی، خیر	۱۰۸، ۱۱۱۶
آلانین آمینوترانسفراز	۶۷۲-۲	۳۸/۲۹ (۴/۱۶)	پروتئین واکنشی سی	مثبت، منفی	۸۸۰، ۳۴۴
آلبومین	۰/۸-۲/۲	۴/۰۷ (۰/۶۶)	مصرف الکل	بلی، خیر	۳۹، ۱۱۸۵
گلوکز	۹۴۴-۱۸	۱۳۶/۰۹ (۷۴/۲)	مصرف دخانیات	بلی، خیر	۸۹، ۱۱۳۵
لاکتات دی‌هیدروژناز	۴/۶۹۷۳-۶	۵۵۵/۶۸ (۳۳۹/۰)	ICU پذیرش	بلی، خیر	۳۸۹، ۸۳۵
هماتوکریت	۳/۶-۷۳/۹	۲۸/۴ (۴/۵۶)	لوله‌گذاری تنفسی	بلی، خیر	۱۸۸، ۱۰۳۶
زمان پروترومبین	۰/۴۶-۹/۸	۱۲/۸۲ (۱/۹)			
آلکالین فسفاتاز	۹/۲۸۴۶-۶	۲۱۳/۱۲ (۱۳۹/۲)			
میزان رسوب گلبول قرمز	۲۵۸-۲	۴۰/۶۵ (۲۸/۲)			

جدول ۴. ضریب همبستگی و مقدار احتمال متغیرهای پژوهش

متغیر	ضریب همبستگی	مقدار احتمال	متغیر	ضریب همبستگی	مقدار احتمال	متغیر	ضریب همبستگی	مقدار احتمال
سن	۰/۲۸۱۹	*۰/۰۰۱	میزان رسوب گلیبول قرمز	-۰/۰۲۶۸	۰/۰۷۸۲	کاهش حس بویایی	۰/۰۲۵۶	۰/۰۶۱۲
جنس	۰/۰۴۲۳	۰/۰۸۵۴	بیلی روبین کامل	۰/۳۱۵	۰/۰۵۹۶	آبریزش بینی	۰/۰۰۳۶	۰/۰۶۶۸
قد	۰/۰۰۷۹	۰/۰۶۷۹	آسپارات آمینوترانسفراز	۰/۱۴۲	*۰/۰۲۶	کاهش حس چشایی	۰/۰۲۵۶	۰/۰۸۱۴
نیترژن اوره خون	۰/۲۸۹	*۰/۰۰۵	آلانین آمینوترانسفراز	۰/۱۸۹۲	*۰/۰۰۹	سردرد	۰/۰۱۴۵	۰/۰۷۴۴
وزن	۰/۰۳۸۸	۰/۰۸۹۷	تب / لرز	۰/۰۶۱۲	*۰/۰۰۱۵	گلودرد/احتقان گلو	۰/۰۳۲۲	۰/۰۸۶۶
گروه خونی	۰/۰۰۵۵	۰/۰۶۸۷	پروتئین واکنش سی	۰/۰۷۸۴۶	*۰/۰۰۴	درد عضلانی	۰/۰۱۷۲	۰/۰۷۶۹
شمارش گلبولهای قرمز	-۰/۰۱۴۷	۰/۰۹۸۵	آلبومین	-۰/۰۲۵۷	۰/۰۶۳۲	احساس ضعف/خستگی	۰/۰۳۱۹	۰/۰۸۴۷
شمارش گلبول سفید	۰/۶۰۷	۰/۰۷۰۴	گلوکز	-۰/۰۳۷۱	۰/۰۶۸۴	سابقه ذات الریه	۰/۰۹۱۴	۰/۰۵۹۰
هماتوکریت	-۰/۰۲۱۲	۰/۰۸۴۶	لاکتات دی هیدروژناز	۰/۰۲۳۷	۰/۰۵۹۶	سابقه بیماری های قلبی - عروقی	۰/۲۱۴۰	*۰/۰۳۴
هموگلوبین	-۰/۰۲۴۲	۰/۰۶۹۶	زمان نسبی ترومبوپلاستین	-۰/۰۳۱۳	۰/۰۵۲۱	سابقه فشارخون	۰/۱۰۱۸	*۰/۰۲۸
شمارش لخته خون	-۰/۰۳۱۳	۰/۰۵۵۸	زمان پروترومبین	-۰/۰۴۲۱	۰/۰۷۴۹	سابقه دیابت	۰/۱۶۳۵	*۰/۰۳۹
تعداد مطلق لنفوسیت	۰/۸۱۵	*۰/۰۲۵	آلکالین فسفاتاز	۰/۰۲۱۱	۰/۰۷۲۰	سابقه سرطان / بدخیمی	۰/۱۱۱۴	*۰/۰۰۲
تعداد مطلق نوتروفیل	۰/۷۸۳	*۰/۰۱۲	کراتینین خون	۰/۱۲۶	۰/۰۶۳۰	سابقه بیماری های کلیوی	-۰/۰۳۰۵	۰/۰۵۸۲
کلسیم	۰/۰۳۸۳	۰/۰۸۸۵	سرفه خشک	۰/۱۶۰۲	*۰/۰۲۰۱	سابقه بیماری های تنفسی	۰/۱۲۱۲	*۰/۰۱۷
فسفر	-۰/۰۲۸۵	۰/۰۶۶۳	صرع	۰/۰۷۰۲	۰/۰۵۹۶	مصرف دخانیات	-۰/۰۲۲۲	۰/۰۸۲۲
منیزیم	۰/۰۱۳۴	۰/۰۷۲۶	حالت تهوع/ استفراغ	-۰/۰۵۹۸	۰/۰۷۱۰	مصرف الکل	-۰/۰۲۵۳	۰/۰۷۰۱
سدیم	۰/۰۲۳۶	۰/۰۷۹۹	سردرد	-۰/۰۱۱۷	۰/۰۸۲۲	پذیرش ICU	۰/۲۴۲۰	*۰/۰۰۹
پتاسیم	-۰/۰۵۲۲	۰/۰۵۶۸	تنگی نفس	۰/۰۸۰۵	*۰/۰۱۰	لوله گذاری تنفسی	۰/۲۸۰۲	*۰/۰۰۳

جدول ۵. نسبت شانس و ضریب استاندارد شده شاخص های مهم پیش بینی کننده مرگ و میر

متغیر	P-value	نسبت شانس	نسبت شانس با اطمینان ۹۵٪	ضریب استاندارد شده	متغیر	P-value	نسبت شانس	نسبت شانس با اطمینان ۹۵٪	ضریب استاندارد شده
سن	<۰/۰۱	۱/۲۴۸	۱/۱-۱۱۵/۶۶۵	۰/۳۸۹	تنگی نفس	۰/۰۱	۱/۹۸۷	۱/۲-۷۸۵/۳۴۱	۰/۲۲۹
تعداد مطلق لنفوسیت	~۰/۰۲	۱/۸۴۱	۱/۲-۷۵۲/۲۸۷	۰/۲۷۶	بیماری های قلبی-عروقی	~۰/۰۲	۱/۱۲۶	۰/۱-۸۹۱/۴۷۸	۰/۳۷۸
تعداد مطلق نوتروفیل	~۰/۰۱	۱/۴۷۸	۱/۱-۲۸۵/۹۴۱	۰/۳۲۵	سابقه فشارخون	~۰/۰۲	۱/۲۴۱	۱/۱-۰۴۱/۵۴۸	۰/۲۲۵
پروتئین واکنش سی	<۰/۰۱	۱/۶۵۴	۱/۱-۲۲۰/۹۱۳	۰/۱۱۰	سابقه دیابت	~۰/۰۲	۱/۲۸۱	۱/۱-۱۱۰/۴۵۷	۰/۲۰۸
نیترژن اوره خون	<۰/۰۱	۱/۶۶۰	۱/۲-۴۴۱/۱۱۲	۰/۲۸۶	سرطان / بدخیمی	<۰/۰۱	۱/۷۸	۱/۲-۵۲۳/۱۱۰	۰/۲۱۰
آسپارات آمینوترانسفراز	~۰/۰۲	۱/۲۰۱	۱/۱-۱۷۰/۴۵۷	۰/۱۴۸	بیماری های تنفسی	~۰/۰۱	۱/۵۴۴	۱/۱-۳۱۵/۸۰۱	۰/۳۲۲
آلانین آمینوترانسفراز	<۰/۰۱	۱/۴۶۶	۱/۱-۲۴۸/۷۳۶	۰/۱۳۶	پذیرش ICU	<۰/۰۱	۱/۴۶	۱/۱-۲۲۰/۷۱۱	۰/۳۷۸
سرفه خشک	۰/۰۲	۱/۲۴۱	۱/۱-۱۲۳/۳۲۴	۰/۲۴۵	لوله گذاری تنفسی	<۰/۰۱	۱/۸۹۲	۱/۱-۶۵۲/۹۲۵	۰/۳۹۷
تب / لرز	<۰/۰۱	۱/۱۱۷	۰/۸۸۹-۳/۲۴۷	۰/۳۳۹					



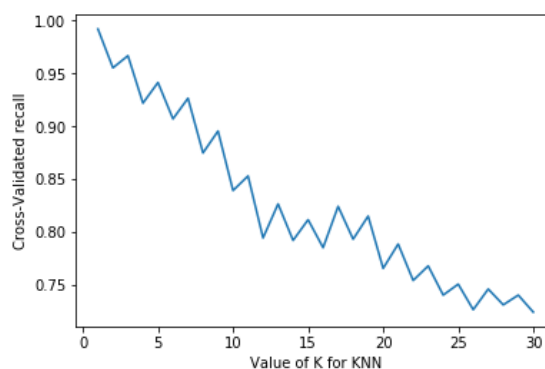
شکل ۲. نمودار ROC الگوریتم های انتخابی در پیش بینی مرگ و میر

جدول ۶. مقایسه الگوریتم های منتخب در طبقه بندی بر روی ویژگی های مستخرج

KNN	SVM	Logistic Regression	Random Forest	الگوریتم	معیار ارزیابانه
۹۴٪/۲۱	۹۶٪/۲۵	۹۴٪/۰۴	۹۱٪/۰۲		دقت
۹۳٪/۷۴	۹۱٪/۱۹	۸۹٪/۹۴	۹۳٪/۲۰		صحت
۱۰۰٪	۹۰٪/۳۸	۸۶٪/۰۶	۹۶٪/۴۰		فراخوانی
۹۳٪/۲	۸۴٪/۲۱	۷۷٪/۱۳	۸۳٪/۵		معیار اف
۹۲٪/۲۳	۸۰٪/۲۷	۸۴٪/۲۴	۸۸٪/۲۸		سطح زیر نمودار راک

کارایی و عملکرد روش های مختلف یادگیری ماشین در پیش بینی مرگ و میر بیماران بستری مبتلاء به کووید-۱۹ پرداخته شد. تکنیک های مورد استفاده در پژوهش حاضر شامل روش های  $k$ -NN، SVM، LR و RF بود که از طریق تعداد ۱۷ متغیر منتخب بالینی و آزمایشگاهی مربوط به ۱۲۲۴ داده بیمار بستری شده مورد آموزش (ایجاد مدل) قرار گرفتند. معیار ارزیابی یا اعتبارسنجی الگوریتم های پیاده سازی شده بر اساس ماتریس آشفتگی بود. در نهایت پس از پیاده سازی مدل های مختلف، الگوریتم  $k$ -NN با صحت طبقه بندی ۹۳/۷۴ درصد عملکرد بهتری را نسبت به سایر الگوریتم ها به دست آورد.

با توجه به حجم بودن برخی از پایگاه های داده ای به خصوص بیماران کووید-۱۹ و فراوانی متغیرهای آنها، بنابراین کاهش متغیرها به عنوان یکی از پیش نیازهای مهم در بهینه سازی ورودی ها در یادگیری ماشین است که نقش بسزایی در ارتقاء عملکرد دسته بندی و پیش بینی ها خواهد داشت [۲۵]. در مطالعات گذشته مهم ترین متغیرهای بالینی تاثیرگذار در پیش بینی مرگ و میرهای ناشی از کووید-۱۹ به عنوان یکی از



شکل ۳. نتیجه ی عملکرد الگوریتم KNN با اعمال 10 Fold بر اساس معیار فراخوانی

## بحث و نتیجه گیری

پژوهش حاضر به صورت گذشته نگر و از طریق شناسایی متغیرهای بالینی مهم در تعیین خطر وقوع مرگ و میر در بیماران بستری مبتلاء به کووید-۱۹ در قالب تکنیک های آماری سعی در طبقه بندی متغیرهای مهم پیش بینی کننده داشت. پس از شناسایی و اولویت بندی مهم ترین متغیرهای بالینی و استفاده از آنها به عنوان ورودی به مدل های داده کاوی، سپس به مقایسه

بیمار، بستری در ICU، لوله‌گذاری تنفسی و سابقه بیماری‌های قلبی-عروقی به عنوان پراهمیت‌ترین متغیرهای پیش‌کننده مرگ‌ومیرهای ناشی از کووید-۱۹ شناسایی شدند. از طرفی متغیرهای قد، گروه خونی، سردرد و آبریزش بینی به عنوان کم‌اهمیت‌ترین متغیرها شناسایی و به همراه ۳۴ متغیر دیگر از مطالعه حذف شدند.

در مجموع از تحلیل نتایج حاصل از انتخاب متغیرها در مطالعه حاضر با نتایج مطالعات گذشته (انتخاب ویژگی در یادگیری ماشین)، درجه بالایی از مطابقت و هم‌خوانی در شناسایی و اولویت‌بندی متغیرهای مهم پیش‌بینی‌کننده مرگ و میر در بیماران بستری مبتلاء به کووید-۱۹ دیده شد. به علاوه بسیاری از مطالعات از دید بالینی (مطالعات مبتنی بر کارآزمایی‌های بالینی و آزمایشگاهی) به فرآیند دسته‌بندی و انتخاب مهم‌ترین متغیرهای پیش‌بینی‌کننده مرگ‌ومیر در بیماران بستری مبادرت کرده‌اند. در مطالعات مورد بررسی [۳۷-۴۱]، ۱۰ متغیر برتر به عنوان موثرترین متغیرها به منظور پیش‌بینی مرگ‌ومیر ناشی از کووید-۱۹ شامل سن بالا، افزایش طول مدت اقامت، لوله‌گذاری تنفسی، تب بالا، کاهش اشباع اکسیژن خون، افزایش اینترلوکین-۶، فشارخون بالا، افزایش گلبول‌های سفید (لکوسیتوزیس)، افزایش نیتروژن اوره خون، بیماری‌های قلبی عروقی و تنفسی (آسم و انسداد مزمن راه‌های هوایی) بود. این نتایج با دسته‌بندی و انتخاب متغیرهای پراهمیت در پژوهش حاضر نیز هماهنگی دارد.

در زمینه کاربرد یادگیری ماشین در پیش‌بینی و تشخیص مرگ و میر ناشی از بیماری کووید-۱۹ تاکنون پژوهش‌های متعددی انجام پذیرفته شده است که در برخی از آن‌ها مقایسه الگوریتم‌های مختلف به منظور شناسایی کارآمدترین الگوریتم به منظور ارائه یک مدل تشخیصی بهینه در دستور کار بوده است. Yadaw و همکاران (۲۰۲۰) پس از ارزیابی عملکرد چهار الگوریتم یادگیری ماشین بر روی تعداد ۳۸۴۱ داده بیمار بستری مبتلاء به کووید-۱۹ به منظور پیش‌بینی احتمال مرگ‌ومیرها، الگوریتم XGBoost با ۹۱ درصد AUC بهترین عملکرد را به دست آورد [۱۳]. در مطالعه Yan بر روی داده‌های ۴۸۵ نفر از بیماران، الگوریتم XGBoost با صحت پیش‌بینی ۹۰ درصد کارآمدترین الگوریتم در پیش‌بینی بدحال شدن بیماران بستری مبتلاء به کووید-۱۹ معرفی شده است [۱۴]. پژوهش Gao و همکاران (۲۰۲۰) به صورت گذشته‌نگر بر روی داده‌های ۲۵۲۰ بیمار قطعی کووید-۱۹، الگوریتم شبکه عصبی با ۹۷/۶۰ درصد AUC کارآمدتر از سه الگوریتم دیگر بود [۳۵]. ارزیابی‌های Vaid و همکاران نیز در سال ۲۰۲۰ بر روی داده‌های مربوط به ۴۰۲۹ نفر از بیماران کووید-۱۹ نشان داد

پیش‌نیازهای راه‌اندازی مدل‌های پیش‌بینی‌کننده مبتنی بر یادگیری ماشین در پنج بخش به شرح زیر شناسایی و بهینه‌سازی شد. ۱- اطلاعات پایه / عمومی شامل سن (بالا) [۲۶-۳۰]، جنس (مذکر) [۳۲،۳۱،۱۰]، شاخص توده بدنی (بالا) [۳۰،۲۹،۲۷،۱۶]، علائم و نشانه‌ها مانند تب [۳۲،۳۱،۲۹،۲۸،۷]، سرفه خشک [۳۳،۳۱،۲۹]، تنگی نفس [۳۴،۳۳،۲۹،۲۶،۱۰]، درد / سفتی سینه [۳۱،۲۸،۱۳]، کاهش سطح هوشیاری و شوک [۳۴،۳۱،۲۹،۱۶،۱۳]، ۳- اطلاعات آزمایشگاهی از قبیل تعداد نوتروفیل / لنفوسیت (کاهش) [۳۵،۳۴،۳۲،۲۹،۱۶،۱۳]، اشباع اکسیژن خون (کاهش) [۳۶،۳۵،۳۲،۲۶،۱۲،۷]، پروتئین فعال C (افزایش)، آسپاراتات / آلانین آمینوترانسفراز (افزایش)، تروپونین قلب (افزایش) [۳۴،۳۲،۳۰،۷]، لاکتات دی هیدروژناز (افزایش) [۳۰-۳۳]، نیتروژن اوره خون (بالا) [۳۴،۳۲،۳۱،۲۷،۲۶]، پروکلسیتونین (بالا) [۲۸،۷] و کراتینین سرم (بالا) [۳۲،۱۰،۷]، ۴- گروه‌های پرخطر مانند مصرف سیگار [۲۹،۲۷]، سابقه مسافرت اخیر [۳۰،۲۹،۱۰]، تنفس مصنوعی (لوله‌گذاری) [۳۴-۳۲،۲۸،۱۳]، سابقه فشارخون [۳۴-۳۲،۲۸،۱۳]، سابقه بیماری‌های قلبی-عروقی [۳۴،۳۲،۳۱،۲۸،۱۳]، سابقه دیابت [۲۹،۲۸،۱۳]، سابقه بدخیمی [۳۴،۳۱،۲۸،۱۳]، سابقه نقص ایمنی [۳۰-۲۸،۱۳] و سابقه بیماری‌های تنفسی [۳۱،۲۸،۲۶] و ۵- اطلاعات تصویربرداری مانند ضایعه ریوی (وجود) [۳۵،۳۳،۳۰،۲۹،۲۷]، گسترش ضایعه (دوطرفه و منتشره) [۲۶،۳۰،۲۸-۳۵]، نوع و بافت‌شناسی ضایعه [۳۰،۲۹،۲۷،۲۶]، از طرفی در این مطالعات موارد نژاد / ملیت [۳۲،۱۰]، نوع تماس / در معرض بودن [۲۸،۱۳،۱۰]، آبریزش بینی [۳۱،۱۰،۷]، سردرد [۳۵،۳۴]، علائم گوارشی [۳۵،۳۱،۲۹،۱۳]، خستگی و درد عضلانی [۳۱،۱۶،۱۰]، فاکتورهای خونی شامل هماتوکریت [۳۱،۱۶،۷]، هموگلوبین [۳۱،۱۶،۱۰،۷] و شمارش گلبول‌های قرمز [۱۶،۷]، از اهمیت پایینی برخوردار بودند.

مشابهاً در مطالعه حاضر پس از انجام تحلیل انتخاب ویژگی به منظور وزن‌دهی به متغیرهای بالینی از بین مجموع ۵۴ متغیر پیش‌بینی‌کننده اولیه، ۱۷ متغیر شامل سن، تعداد مطلق لنفوسیت، تعداد مطلق نوتروفیل، پروتئین واکنش سی، نیتروژن اوره خون، آسپاراتات آمینوترانسفراز، آلانین آمینوترانسفراز، سرفه خشک، تب / لرز، تنگی نفس، بیماری‌های قلبی-عروقی، سابقه فشارخون، سابقه دیابت، سرطان / بدخیمی، بیماری‌های تنفسی، پذیرش ICU و لوله‌گذاری تنفسی دارای سطح P-value کم‌تر از ۰/۰۵ بود و به عنوان متغیرهای نهایی شناسایی شد. پس از ارزیابی نسبت شانس و ضریب استاندارد شده، متغیرهای سن



اطلاعاتی و استفاده از نظرات متخصصین بالینی، خلاءها کیفی فیلدهای اطلاعاتی به حداقل رسد. علاوه بر این به منظور برطرف کردن مشکل فیلدهای خالی از حذف رکوردها با بیش از ۷۰ درصد فیلد خالی و نیز جایگزینی آن‌ها با شاخص‌های میانگین و میانه استفاده شد. از طرفی با توجه به نامتوازن بودن کلاس‌های فوتی و زنده، سعی شد تا از روش نمونه‌برداری پیشرفته (SMOTE) [۴۳] به منظور کاهش بایاس از طریق متعادل‌سازی کلاس‌های دوگانه استفاده شود. از دیگر محدودیت‌ها، استفاده از تنها چهار روش داده کاوی و نیز عدم استفاده از داده‌های تصویربرداری در مجموعه داده انتخابی به منظور پیش‌بینی مرگ و میر بود.

پیشنهاد می‌گردد به منظور بهبود کیفیت مدل و کاهش بیش‌داوری در پیش‌بینی، پژوهش‌های بیشتری پس از انجام اعتبارسنجی‌های دقیق‌تر، با استفاده از الگوریتم‌های داده کاوی متنوع‌تر در پایگاه‌های داده بزرگ‌تر، با در نظر داشتن متغیرهای تشخیصی بیشتر و با کیفیت‌تر، برگرفته از چند مرکز مراقبتی، و با توجه به ماهیت آینده‌نگر بودن جمع‌آوری داده‌ها، انجام پذیرد.

از آنجایی که شدت علائم و وخامت ناخوشی در بیماران مبتلاء به کووید-۱۹ بسیار متغیر است، بنابراین ارائه مدل‌هایی جهت حمایت از تصمیم و پیش‌بینی رویدادها و پیامدهای ناشی از بیماری به‌خصوص تعیین خطر مرگ و میر و بدحال شدن بیماران با استفاده از تکنیک‌های یادگیری ماشین به منظور استفاده بهینه از منابع محدود بیمارستانی، سطح‌بندی بیماران، حمایت از تصمیمات بالینی و انتخاب بهترین مداخله درمانی و حمایتی، حیاتی است. نتایج حاصل از ارزیابی عملکرد روش‌های یادگیری ماشین منتخب در پژوهش حاضر نشان داد الگوریتم KNN قابلیت بالاتری در پیش‌بینی زود هنگام موارد در معرض خطر بدحال شدن و فوت ناشی از کووید-۱۹ داشت. با این وجود پیشنهاد می‌گردد به منظور بهبود کیفیت مدل و کاهش بیش‌داوری در پیش‌بینی، پژوهش‌های بیشتری پس از انجام اعتبارسنجی‌های دقیق‌تر، با استفاده از الگوریتم‌های داده کاوی متنوع‌تر در پایگاه‌های داده بزرگ‌تر، با در نظر داشتن متغیرهای تشخیصی بیشتر و با کیفیت‌تر، برگرفته از چند مرکز مراقبتی، و با توجه به ماهیت آینده‌نگر بودن جمع‌آوری داده‌ها، انجام پذیرد.

### تشکر و قدردانی

مقاله حاضر برگرفته از یک طرح پژوهشی مصوب و مشترک بین دانشگاه علوم پزشکی آبادان و ایلام با کد IR.ABADANUMS.REC.1399.128 می‌باشد. از مدیران

که الگوریتم پرسپترون چندلایه (MLP) با  $AUC=82/20$  درصد به نسبت الگوریتم رگرسیون لاسو (LASSO) در پیش‌بینی بدحال شدن بیماران و مرگ آن‌ها از قابلیت بالاتری برخوردار است [۳۶]. در مطالعه دیگری که توسط An (۲۰۲۰) از طریق آزمایش تکنیک‌های داده کاوی مختلف بر روی یک مجموعه داده شامل ۱۰۲۳۷ بیمار در نهایت SVM با حساسیت، اختصاصیت و منحنی مشخصه عملکرد ROC به ترتیب  $7/90$  درصد،  $91/4$  درصد و  $3/96$  درصد بهترین عملکرد را در پیش‌بینی مرگ و میرهای ناشی از کووید-۱۹ داشت. در مطالعه Chen (۲۰۲۰) از طریق آموزش مدل‌های مختلف یادگیری ماشین بر روی پایگاه داده مشتمل بر داده‌های ۱۲۷۲ فرد مبتلاء به کووید-۱۹ و بستری شده در بیمارستان، الگوریتم SVM با صحت ۹۳ درصد به عنوان کارآمدترین الگوریتم برای پیش‌بینی بستری شدن در ICU و بدحال شدن بیمار معرفی شد [۴۲]. مشابهاً در بررسی Agieb (۲۰۲۰) بر روی ۱۰۸۷ داده بیماران قطعی کووید-۱۹، الگوریتم SVM با صحت ۱۰۰ درصد، بهترین عملکرد را به نسبت دو الگوریتم دیگر در پیش‌بینی بدحال شدن و مرگ بیماران کسب کرد [۳۶]. در پژوهش‌های جداگانه توسط Booth [۲۷] و Parchure [۱۱] بر روی داده‌های ۶۸۰۰ و ۵۴۸۰ بیمار بستری شده مبتلاء به کووید-۱۹، الگوریتم RF به ترتیب با سطح زیر نمودار ۹۳ درصد و ۸۵ درصد، حساسیت ۹۱ درصد و  $87/8$  درصد و اختصاصیت ۹۱ درصد و  $91/4$  درصد به منظور پیش‌بینی بقاء بیماران بهترین کارایی و دقت را داشتند. در پژوهش حاضر نتایج حاصل از ارزیابی عملکرد الگوریتم‌های منتخب پس از انجام انتخاب ویژگی و شناسایی متغیرهای پیش‌بینی‌کننده مهم (۱۷ متغیر) نشان داد که روش KNN با دقت  $94/21$  درصد، صحت  $93/74$  درصد، فراخوانی ۱۰۰ درصد، معیار اف  $93/2$  درصد و سطح زیر نمودار رآک  $93/21$  درصد بهترین عملکرد را در طبقه‌بندی بیماران بر اساس خطرات مرگ و میر آن‌ها دارد.

در پایان، هر چند مطالعه حاضر عملکرد بهینه‌ای در پیش‌بینی مرگ و میر بیماران بستری مبتلاء به کووید-۱۹ داشت، اما یک سری محدودیت‌ها و چالش‌های بالقوه داشت. مهم‌ترین محدودیت پژوهش حاضر، گذشته‌نگر بودن عناصر داده و تک مرکزی بودن پایگاه داده انتخابی بود که می‌توانست در کیفیت، جامعیت و تعمیم‌پذیری داده‌ها اثرگذار باشد. در این میان وجود برخی فیلدهای اطلاعاتی خالی، غیریکپارچه، ناقص، دارای خطا و موارد غیرطبیعی می‌توانست در فرآیند آموزش روش‌های داده کاوی تاثیر منفی گذارد. بنابراین در مرحله نخست تلاش شد تا از طریق واریاسیون دستی رکوردهای

- [12] Ryan L, Lam C, Mataraso S, Allen A, Green-Saxena A, Pellegrini E, et al. Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: a retrospective study. *Ann Med Surg (Lond)* 2020; 59: 207-216. <https://doi.org/10.1016/j.amsu.2020.09.044> PMID:33042536 PMCID:PMC7532803
- [13] Yadaw AS, Li YC, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit Health* 2020; 2: e516-e525. [https://doi.org/10.1016/S2589-7500\(20\)30217-X](https://doi.org/10.1016/S2589-7500(20)30217-X)
- [14] Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020; 2: 283-288. <https://doi.org/10.1038/s42256-020-0180-7>
- [15] Zhao Z, Chen A, Hou W, Graham JM, Li H, Richman PS, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19. *PloS One* 2020; 15: e0236618. <https://doi.org/10.1371/journal.pone.0236618> PMID:32730358 PMCID:PMC7392248
- [16] Zhou Y, He Y, Yang H, Yu H, Wang T, Chen Z, et al. Exploiting an early warning Nomogram for predicting the risk of ICU admission in patients with COVID-19: a multi-center study in China. *Scand J Trauma Resusc Emerg Med* 2020; 28: 1-13. <https://doi.org/10.1186/s13049-020-00795-w> PMID:33109234 PMCID:PMC7590555
- [17] Kaiafas KN. Clinical Decision Support Tools and the COVID-19 Pandemic. *J Christ Nurs* 2020; 37: 192. <https://doi.org/10.1097/CNJ.0000000000000737> PMID:32516262
- [18] McRae MP, Simmons GW, Christodoulides NJ, Lu Z, Kang SK, Fenyo D, et al. Clinical decision support tool and rapid point-of-care platform for determining disease severity in patients with COVID-19. *Lab Chip* 2020; 20: 2075-2085. <https://doi.org/10.1039/D0LC00373E> PMID:32490853 PMCID:PMC7360344
- [19] Bredmose PP, Diczbalis M, Butterfield E, Habig K, Pearce A, Osbakk SA, et al. Decision support tool and suggestions for the development of guidelines for the helicopter transport of patients with COVID-19. *Scand J Trauma Resusc Emerg Med* 2020; 28: 43. <https://doi.org/10.1186/s13049-020-00736-7> PMID:32450877 PMCID:PMC7247287
- [20] Albahri A, Hamid RA. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. *J Med Syst* 2020; 44: 1-11. <https://doi.org/10.1007/s10916-020-01582-x> PMID:32451808 PMCID:PMC7247866
- [21] Mohammadi F, Kouzehgari S. Predicting the prevalence of COVID-19 and its mortality rate in Iran using lyapunov exponent. *J Inflamm Dis* 2020; 24: 108-123. (Persian). <https://doi.org/10.32598/JQUMS.24.2.2415.1>
- [22] Mohammadzadeh N, Shahriary M, Shirmohammadlou N, Lohrasbi V. A glance at the prevalence of coronavirus disease 19 (COVID-19) in Iran: Strengths and weaknesses. *Infect Control Hosp Epidemiol* 2020; 41: 1479-1482. <https://doi.org/10.1017/ice.2020.193> PMID:32362292 PMCID:PMC7237388
- [23] Ramsay MA, Usman M, Lagow E, Mendoza M, Untalan E, De Vol E. The accuracy, precision and reliability of measuring ventilatory rate and detecting ventilatory pause by rainbow acoustic monitoring and capnometry. *Anesth Analg* 2013; 117: 69-75. <https://doi.org/10.1213/ANE.0b013e318290c798> PMID:23632055
- [24] Evans JR, Fisher RP. Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Appl Cogn Psychol* 2011; 25: 501-508. <https://doi.org/10.1002/acp.1722>
- [25] Brownlee J. Data preparation for machine learning: data cleaning, feature selection, and data transforms in python: Machine Learning Mastery; 2020.
- [26] Allenbach Y, Saadoun D, Maalouf G, Vieira M, Hellio A, Boddaert J, et al. Development of a multivariate prediction model of intensive care unit transfer or death: A French prospective cohort study of hospitalized COVID-19 patients. *PloS One* 2020; 15: e0240711.

محترم معاونت تحقیقات و فناوری این دانشگاه‌ها و هم‌چنین مسئولین بیمارستان مصطفی خمینی شهر ایلام که در انجام این پژوهش تیم تحقیق را یاری رساندند، کمال تشکر و قدردانی را داریم.

## مشارکت و نقش نویسندگان

مصطفی شنبه‌زاده و محمدرضا کفایشیان: ایده و طراحی مطالعه، هادی کاظمی آرپناهی: جمع‌آوری داده‌ها، رامین افراه و اعظم اروجی: آنالیز و تفسیر نتایج، هادی کاظمی آرپناهی و علی ولی نژادی: نگارش نسخه اول مقاله. همه نویسندگان نتایج را بررسی نموده و نسخه نهایی مقاله را تایید نمودند.

## منابع

- [1] Tang D, Comish P, Kang R. The hallmarks of COVID-19 disease. *Plos Pathogens* 2020; 16: e1008536. <https://doi.org/10.1371/journal.ppat.1008536> PMID:32442210 PMCID:PMC7244094
- [2] Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ, et al. COVID-19: consider cytokine storm syndromes and immunosuppression. *The Lancet* 2020; 395: 1033-1034. [https://doi.org/10.1016/S0140-6736\(20\)30628-0](https://doi.org/10.1016/S0140-6736(20)30628-0)
- [3] Mirmohammadkhani M, Paknazar F, Rashidy-Pour A. Evaluation of the epidemiological pattern of COVID-19 applying basic reproduction number: an educational review article. *Koomesh* 2020; 22. (Persian). <https://doi.org/10.29252/koomesh.22.3.373>
- [4] Allam Z, Jones DS, editors. On the coronavirus (COVID-19) outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (AI) to benefit urban health monitoring and management. *Healthcare*; 2020: Multidisciplinary Digital Publishing Institute. <https://doi.org/10.3390/healthcare8010046> PMID:32120822 PMCID:PMC7151018
- [5] Chaudari SN, Mene SP, Bora RM, Somavanshi KN. Role of Internet of Things (IOT) In Pandemic Covid-19 Condition. Retrieved September, 2020, 9: 2020.
- [6] Peck KRJCM, Infection. Early diagnosis and rapid isolation: response to COVID-19 outbreak in Korea. *Clin Microbiol Infect* 2020; 26: 805-807. <https://doi.org/10.1016/j.cmi.2020.04.025> PMID:32344168 PMCID:PMC7182747
- [7] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020; 369. <https://doi.org/10.1136/bmj.m1328> PMID:32265220 PMCID:PMC7222643
- [8] Chamola V, Hassija V, Gupta V, Guizani M. A comprehensive review of the COVID-19 pandemic and the role of IoT, Drones, AI, Blockchain, and 5G in managing its impact. *IEEE Access* 2020; 8: 90225-90265. <https://doi.org/10.1109/ACCESS.2020.2992341>
- [9] Montazeri M. Machine learning models for predicting the diagnosis of liver disease. *Koomesh* 2014; 16. (Persian).
- [10] Pan P, Li Y, Xiao Y, Han B, Su L, Su M, et al. Prognostic Assessment of COVID-19 in the Intensive Care Unit by Machine Learning Methods: Model Development and Validation. *J Med Int Res* 2020; 22: e23128. <https://doi.org/10.2196/23128> PMID:33035175 PMCID:PMC7661105
- [11] Parchure P, Joshi H, Dharmarajan K, Freeman R, Reich DL, Mazumdar M, et al. Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19. *BMJ Support Palliat Care* 2020. <https://doi.org/10.1136/bmjspcare-2020-002602> PMID:32963059

- [35] Gao Y, Cai GY, Fang W, Li HY, Wang SY, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun* 2020; 11: 1-10.  
<https://doi.org/10.1038/s41467-020-18684-2>  
PMid:33024092 PMCid:PMC7538910
- [36] Agieb R. Machine learning models for the prediction the necessity of resorting to icu of covid-19 patients. *Int J Adv Trends Com Sci Eng*. 2020; 6980-4.  
<https://doi.org/10.30534/ijatse/2020/15952020>
- [37] Bhargava A, Szpunar SM, Sharma M, Fukushima EA, Hoshi S, Levine M, et al. Clinical features and risk factors for in-hospital mortality from COVID-19 infection at a tertiary care medical center, at the onset of the US COVID-19 pandemic. *J Intensive Care Med* 2021; 36: 711-718.  
<https://doi.org/10.1177/08850666211001799>  
PMid:33759606 PMCid:PMC7992101
- [38] Chilimuri S, Sun H, Alemam A, Mantri N, Shehi E, Tejada J, et al. Predictors of mortality in adults admitted with COVID-19: Retrospective cohort study from New York city. *West J Emerg Med* 2020; 21: 779.  
<https://doi.org/10.5811/westjem.2020.6.47919>  
PMid:32726241 PMCid:PMC7390589
- [39] Lai X, Liu J, Zhang T, Feng L, Jiang P, Kang L, et al. Clinical, laboratory and imaging predictors for critical illness and mortality in patients with COVID-19: protocol for a systematic review and meta-analysis. *BMJ Open* 2020; 10: e039813.  
<https://doi.org/10.1136/bmjopen-2020-039813>  
PMid:33361074 PMCid:PMC7768616
- [40] Moledina SM, Maini AA, Gargan A, Harland W, Jenney H, Phillips G, et al. Clinical characteristics and predictors of mortality in patients with COVID-19 infection outside intensive care. *Int J Gen Med* 2020; 13: 1157.  
<https://doi.org/10.2147/IJGM.S271432>  
PMid:33244256 PMCid:PMC7683500
- [41] Sacak ME, Karacabey S, Sanri E, Omecikoglu S, Ünal E, Onur ÖE, et al. Variables affecting mortality among COVID-19 patients with lung involvement admitted to the emergency department. *Cureus* 2021;13: e12559.
- [42] Chen Y, Linli Z, Lei Y, Yang Y, Liu Z, Xia Y, et al. Risk factors for mortality in critically ill patients with COVID-19 in Huanggang, China: A single-center multivariate pattern analysis. *J Med Virol* 2021; 93: 2046-2055.  
<https://doi.org/10.1002/jmv.26572>  
PMid:32997344 PMCid:PMC7537509
- [43] Han J, Pei J, Kamber M. *Data mining: concepts and techniques*: Elsevier 2011.  
<https://doi.org/10.1371/journal.pone.0240711>  
PMid:33075088 PMCid:PMC7571674
- [27] Booth AL, Abels E, McCaffrey P. Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Modern Pathol* 2021; 34: 522-531.  
<https://doi.org/10.1038/s41379-020-00700-x>  
PMid:33067522 PMCid:PMC7567420
- [28] Hu H, Yao N, Qiu Y. Comparing rapid scoring systems in mortality prediction of critically ill patients with novel coronavirus disease. *Acad Emerg Med* 2020; 27: 461-468.  
<https://doi.org/10.1111/acem.13992>  
PMid:32311790 PMCid:PMC7264631
- [29] Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *Eur Respir J* 2020; 56: 2001104.  
<https://doi.org/10.1183/13993003.01104-2020>  
PMid:32616597 PMCid:PMC7331655
- [30] Zhang Y, Xin Y, Li Q, Ma J, Li S, Lv X, et al. Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *Biomed Eng Online* 2017; 16: 125.  
<https://doi.org/10.1186/s12938-017-0416-x>  
PMid:29096638 PMCid:PMC5668968
- [31] Chin V, Samia NI, Marchant R, Rosen O, Ioannidis JP, Tanner MA, et al. A case study in model failure? COVID-19 daily deaths and ICU bed utilisation predictions in New York State. *Eur J Epidemiol* 2020; 35: 733-742.  
<https://doi.org/10.1007/s10654-020-00669-6>  
PMid:32780189 PMCid:PMC7417851
- [32] Das AK, Mishra S, Gopalan SS. Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ* 2020; 8: e10083.  
<https://doi.org/10.7717/peerj.10083>  
PMid:33062451 PMCid:PMC7528809
- [33] Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Machine Intell* 2020; 1-6.  
<https://doi.org/10.1038/s42256-020-0180-7>
- [34] Assaf D, Gutman Ya, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Intern Emerg Med* 2020; 15: 1435-1443.  
<https://doi.org/10.1007/s11739-020-02475-0>  
PMid:32812204 PMCid:PMC7433773

# Comparison of machine-learning algorithms efficiency to build a predictive model for mortality risk in COVID-19 hospitalized patients

Mostafa Shanbehzadeh (Ph.D)<sup>1</sup>, Ali Valinejadi (Ph.D)<sup>2</sup>, Ramin Afrah (M.Sc)<sup>3</sup>, Hadi Kazemi-Arpanahi (Ph.D)<sup>\*4</sup>, Azam Orooji (PhD)<sup>5</sup>, Mohammadreza Kaffashian (PhD)<sup>6</sup>

1 – Dept. of Health Information Technology, Faculty of Paramedical, Ilam University of Medical Sciences, Ilam, Iran

2 - Dept. of Health Information Technology, School of Allied Medical Sciences, Semnan University of Medical Sciences, Semnan, Iran

3- Dept. of Biomedical Engineering, Faculty of Advanced Technologies, Isfahan University of Medical Sciences, Isfahan, Iran

4- Dept. of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran

5- Dept. of Advanced Technologies, Faculty of Medicine, North Khorasan University of Medical Science (NKUMS), Bojnourd, Iran

6- Dept. of Physiology, Faculty of Medicine, Ilam University of Medical Sciences, Ilam, Iran

\* Corresponding author. +98 9138200027 h.kazemi@abadanums.ac.ir

Received: 1 Mar 2021; Accepted: 3 Jul 2021

**Introduction:** The rapid worldwide outbreak of SARS-CoV-2 has posed serious and unprecedented challenges to healthcare systems in predicting disease behavior and outcomes. To overcome these challenges or ambiguities, this study aimed to create and validate several predictive models using of selected ML algorithms to stratify the mortality risk in COVID-19 hospitalized patients and choice the best performing algorithm.

**Materials and Methods:** Data of 1224 hospitalized patients with COVID-19 diagnosis based on the findings of the confirmed-laboratory test were extracted from the Ilam COVID-19 registry (Ilam CoV reg) database. Then the most important clinical parameters in the COVID-19 mortality were identified and used as inputs of the selected ML algorithms, including K-Nearest Network (KNN), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF). Finally, the performance of the developed models was compared based on different confusion matrix evaluation criteria and the most appropriate predictive model was determined.

**Results:** A total of 17 parameters were identified as the most influential clinical variables in the mortality of COVID-19. By comparing the performance of ML algorithms according to various evaluation criteria, the KNN algorithm with precision of 94.21%, accuracy of 93.74%, recall of 100%, F-measure of 93.2% and ROC of 92.23%, yielded better performance than other developed algorithms.

**Conclusion:** KNN enables a reasonable level of accuracy and certainty in predicting the mortality of patients with COVID-19 and potentially facilitates identifying high risk patients and, inform proper interventions by the clinicians.

**Keywords:** COVID-19, SARS-CoV-2, Artificial Intelligence, Machine Learning, Data Mining