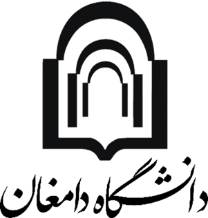
به نام خدا

تحليل مقايسه‌اي الگوريتم‌هاي كا نزديك‌ترين همسايه ، ژنتيك، ماشين بردار پشتيبان، درخت تصميم و حافظه كوتاه‌مدت بلند در يادگيري ماشين.

کوروش جمال‌پور، امیرمهدی حسینی



چكيده

 در یادگیری ماشین یک فناوری پر رونق دوران جدید است که کامپیوترها را قادر می‌سازد به طور خودکار از داده‌های پیشین خوانده و تفسیر کنند. این فناوری از الگوریتم‌های متعدد برای ساخت مدل‌ها با طبیعت ریاضی استفاده می‌کند و سپس با استفاده از داده‌های گذشته و دانش، پیش‌بینی‌ها برای داده‌های جدید انجام می‌دهد. اخیراً، ً این فناوری برای شناسایی متن، تشخیص گفتارهای نفرت‌انگیز، سیستم‌های پیشنهادی، تشخیص چهره و موارد دیگر به کار گرفته است. در این مقاله، به طور مفصل بررسی شده‌اند. تمامی جنبه‌های مربوط به پنج الگوریتم یادگیری ماشین به نام K-Nearest Neighbor، Algorithm Genetic(GA)، Support Vector Machine(SVM)، Tree Decision(TD)، Long Short Term Memory(LSTM) که یک پیش‌نیاز برای ورود به حوزه یادگیری ماشین است. این مقاله نوری افکنده بر نتایج و استنتاج‌های جدید مرتبط با این الگوریتم‌ها از طریق تحقیق و بررسی مقالات اخیر که تحقیقات کمی و کیفی را در مسئله زمان واقعی، به‌ویژه تجزیه‌وتحلیل پیش‌بینی در زمینه‌های چندرشته‌ای انجام داده‌اند. این مقاله همچنین درباره منشأ وضعیتی این الگوریتم‌ها صحبت می‌کند که اگرچه در مقالات قبلی به‌ندرت موردبحث قرار گرفته است، اما نکته‌ای برجسته برای علاقه‌مندان و غیرحرفه‌ای‌های ML است. برای توضیح و درک دقت، قدرت و قابلیت اعتماد الگوریتم‌ها، آن‌ها به طور جامع از نظر کیفی و کمی مورد بررسی و تحقیق قرار گرفته‌اند که در آن شبکه LSTM و الگوریتم SVM رفتار برتری نسبت به سایرین را نشان داده‌اند.

كلمات كليدي

يادگيری ماشين، Kنزديک‌ترين همسايه، الگوريتم ژنتيک، ماشين بردار پشتيبان، درخت تصميم، الگوريتم حافظه کوتاه-مدت بلند.

# -1 مقدمه

## يادگيری ماشين، ترکيبی از مفاهيم آماری و دانش علمی رایانه هاست. این اصطلاح توسط آرتور ساموئل در سال ۱۹۵۹ ابداع شد و اکنون به عنوان زیرمجموعه ای از هوش مصنوعی شناخته می‌شود.

## الگوریتم‌های یادگیری ماشین امکان پردازش و طبقه‌بندی خودکار داده‌های جدید بر اساس اطلاعات قدیمی را برای پردازنده ها یا رایانه‌ها فراهم می‌کنند. بدون برنامه‌نویسی جامع، رایانه‌ها می‌توانند پیش‌بینی کنند و تصمیم بگیرند، زیرا از مدل‌های ریاضی استفاده می‌کنند که توسط این الگوریتم‌های یادگیری ماشین با کمک داده‌های آموزشی (که مجموعه داده‌های نمونه موجود است) ساخته شده‌اند.

## برای بیان یک بیانیه مسئله پیچیده، اگر بخواهیم نوعی پیش‌بینی انجام دهیم، الزام نیست کد کامل مسئله را طراحی و نوشته شود، به جای آن فقط با ارائه اطلاعات موجود به الگوریتم، می‌توان توسط رایانه یک مدل ریاضی یا منطقی ساخت تا نتیجه را پیش‌بینی کند.به طور کلی، یادگیری ماشین به سه دسته عمده تقسیم می‌شود: یادگیری بدون نظارت، یادگیری نظارت شده و یادگیری تقویتی.

## نظریه یادگیری نوع نظارت شده بر کلمه "نظارت" تمرکز دارد، جایی که هدف آن نقشه‌برداری داده‌های مرتبط با ورودی به داده‌های مرتبط با خروجی است.این روش بدون شک به مقدار قابل توجهی از کاربری انسانی برای ساخت مدل نیاز دارد، اما در نهایت منجر به اجرای سریعتر یک کار پیچیده می‌شود. یادگیری ماشین نظارت شده یک دسته گسترده از یادگیری ماشین است و به طبقه‌بندی بیشتر به الگوریتم‌های رگرسیون و طبقه‌بندی تقسیم می‌شود. یادگیری بدون نظارت امکان می‌دهد تا ماشین بدون هیچ نظارتی یاد بگیرد.

## در یادگیری بدون نظارت، یک مجموعه داده غیرجدا شده و بدون برچسب به ماشین ارائه شده و الگوریتم باید بر روی داده‌های بدون هیچ نظارتی عمل کند. این نظریه به هدف دارد عناصر داده ورودی را که الگوهای مشابهی نشان می‌دهند دوباره گروه‌بندی کند.

## در این نظریه امکان پیش‌بینی هیچ نتایجی وجود ندارد و ماشین تلاش می کند تا بر اساس حجم عظیمی از داده‌ها درک‌های مهمی ارائه دهد.این دوباره به زیرشاخه‌های خوشه‌بندی و انجمن تقسیم می‌شود.

## یادگیری تقویتی، این تئوری به عنوان یک مکانیزم مبتنی بر بازخورد وجود دارد، جایی که فرد یادگیرنده برای هر حرکت صحیح پاداش می‌گیرد و برای عمل نادرست مجازات می‌شود. با این انگیزه‌ها، یادگیرندگان می‌توانند سیستم را تصحیح کرده و عملکرد آن را افزایش دهند.

## در این نوع یادگیری، فرد اساسا با محیط ترکیب می‌شود و سعی می‌کند بیشتر درباره آن کشف کند. همانطور که قبلا اشاره شد، دو دسته زیر یادگیری نظارت شده وجود دارد: رگرسیون و طبقه‌بندی.الگوریتم‌های متعلق به زیردسته رگرسیون مفید هستند زمانی که متغیر ورودی به نحوی با متغیر خروجی مرتبط است و الزام است متغیرهایی از طبیعت پیوسته مانند سهام یا برخی از روندهای جمعیتی پیش‌بینی شود. در حالی که الگوریتم‌های طبقه‌بندی مفید هستند زمانی که نتیجه از نوع زمینه ای است مانند “دایره یا مثلث، درست یا غلط، راست یا چپ، بله یا خیر” و غیره.

**2- K نزديک‌ترين همسايه**

## K-NN یکی از الگوریتم‌های حیاتی و موثر در تفکیک داده‌ها است، قادر است تا انتخاب اصلی برای پیاده‌سازی باشد، به ویژه زمانی که داده‌های موجود نسبتاً مبهم باشند.

## این الگوریتم توسط اوولین فيکس و جوزف هاجز در سال ۱۹۵۱ برای بررسی جداکننده ارايه شد، زمانی که تصمیم‌گیری درباره چگونگی چگالی‌های احتمالاتی با استفاده از تخمین پارامتری نسبتاً چالش برانگیز بود. در سال ۱۹۶۷، چند ویژگی مرتبط با این الگوریتم محاسبه شد؛ به عنوان مثال هنگامی که ‌'k' برابر ۱ است و 'n' به بی نهایت نزدیک می‌شود، محدودیت خطای یا اشتباه طبقه‌بندی K-NN بالاتر از دوبرابر نرخ خطای بیز است.

## پس از بررسی این ویژگی‌ها و خصوصیت‌های خاص، تحقیق و آزمایش از طریق دوره‌های طولانی برای شمارش روش‌های جدید ردیابی، بهبودها برای نرخ خطای بیز، روش‌هایی که فقط بر اساس فاصله اعتماد می‌کنند، روش‌های محاسبات نرم و رویکردهای دیگر انجام شد. الگوریتم K-NN در زیرنوعی از روش‌های یادگیری نظارت شده قرار دارد و یکی از آسان‌ترین الگوریتم‌های استفاده شده در یادگیری ماشین است.اگرچه مناسب برای طبقه‌بندی و هم‌زمان‌سازی هر دواستفاده می‌شود، اما اصولاً برای طبقه‌بندی اشیاء استفاده می‌شود. این الگوریتم بسیار کارآمد است و برای اختصاص هر مقدار گم‌شده و بازنمونه‌برداری داده‌ها استفاده می‌شود. برای مجموعه داده داده شده، این الگوریتم پیش‌بینی ارتباط بین داده‌های پنهان و داده‌های موجود را انجام می‌دهد و بر اساس آن پیش‌بینی، داده‌های جدید را به دسته‌بندی موجود نزدیکی بیشتر با آن مطابقت دارند. بنابراین، داده‌های تازه می‌توانند توسط الگوریتم K-NN بطور قطعی دسته‌بندی شوند. این الگوریتم نقطه یا شکل داده‌های جدید را بر اساس ترتیب داده‌های همسایه‌اش مرتب می‌کند.

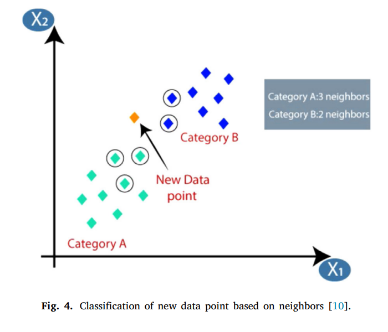
## K-NN همچنین می‌تواند به عنوان الگوریتم یادگیری تنبل معرفی شود، زیرا مجموعه داده ابتدایی تنها در ابتدا ذخیره می‌شود، اما فرآیند یادگیری مجموعه داده‌های آموزش تازه درصورت نیاز به طبقه‌بندی یا پیش‌بینی داده‌های جدید انجام نمی‌شود.

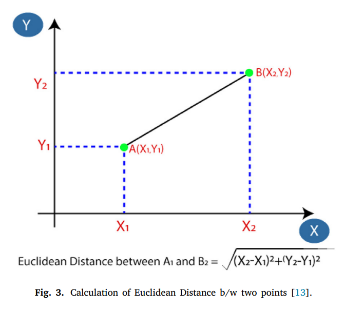
## همچنین این بی‌پارامتری طبیعی است، یعنی در K-NN هیچ روش یا شکل پیش‌تعیین شده‌ای برای رابطه بین ورودی و خروجی وجود ندارد. در شکل ۲، دو حالت وجود دارد، تومور خفیف یا بدخیم یک نقطه داده جداگانه برای مشخص‌کردن بیشتر یا بدخیم انتخاب شده‌است. در این حالت، الگوریتم K-NN می‌تواند به آسانی به تحلیل‌گران در روند طبقه‌بندی نقطه جدید مختلف از مجموعه داده کمک کند، بر اساس شباهت یا شاخص مشابهت نقطه با هر دو مورد موجود. الگوریتم K-NN میتواند زمانی استفاده شود که مجموعه داده مورد نظر دارای برچسب و بدون نویز باشد.

## 1-2- عملکرد الگوريتم K نزديک ترين همسايه

## حرف ’K’ موجود در K-NNبه تعداد همسایه‌ها (داده‌هایی که نزدیک‌ترین به نقطه داده جدید هستند) اشاره دارد. تعیین یک مقدار مناسب برای K فرآیند اصلی این الگوریتم است. برای دقت بیشتر، حیاتی است که فرد مقدار صحیح K را انتخاب کند، و این فرآیند به تنظیم پارامتر معروف است. مقدار بسیار پایینی برای K مانند ۱ یا ۲ می‌تواند به نتایج نویزی منجر شود، در حالی که مقدار بسیار بالا در برخی موارد ممکن است ابهام ایجاد کند، بسته به مجموعه داده مقدار ثابتی برای K وجود ندارد، با این حال، یکی از مقادیر استانداردی که K غالباً آن را به خود می‌گیرد عدد ‘۵‘ است، یعنی برای فرآیند اکثریت‌گیری، ۵ همسایه نزدیک‌تر به نقطه داده جدید در نظر گرفته می‌شوند.

## برای جلوگیری از اشتباهات و ابهامات میان دو کلاس مجموعه داده، به طور کلی، مقدار نادری از K مناسب می‌باشد. یک دیگر از محاسبه‌ی مبتنی بر فرمول برای K می‌تواند از این فرمول انجام شود: و n تعداد کل نقاط داده را نمایش می‌دهد. دنبال شده توسط آن، فاصله از نوع اقلیدسی نقاط موجود در مجموعه داده تا نقطه داده جدید محاسبه می‌شود. برای انجام این کار حیاتی است که مجموعه داده به شکل گرافیکی نمایش داده شود. فاصله اقلیدسی به شکلی که در شکل ۳ نشان داده شده است محاسبه می‌شود. پس از محاسبه‌ی ارزش‌های فواصل اقلیدسی تمام نقاط از نقطه داده جدید، باید دقت شود به کدام کلاس اکثر از همسایه‌های نزدیکشان تعلق دارند به عنوان مثال، در ۵=K و سپس پس از محاسبه دقیق، آن کلاس را به نقطه داده تعلق داده شده برای طبقه‌بندی، الصاق کرد. مثل شکل، ۴ می‌توان نتیجه گرفت که نقطه به کلاس A تعلق دارد، زیرا دارای ۳ همسایه نزدیک (اکثریت) از آن دسته است.

****



## 2-2- مقايسه ی الگوريتم های يادگيری ماشين رگرسيون لجستيک، بيز ساده و KNN برای تشخيص کلاهبرداری کارت اعتباری - برنامه ی کاربردی اخير

**1-2-2- زمينه ی کار اخير**

کارت‌های اعتباری به دلیل پیشرفت بی‌وقفه فناوری اینترنت امروزه به عنوان یک روش گسترده برای پرداخت‌ها پذیرفته شده‌اند. با این وصف، تقلب‌های بانکی امروزه نیز بیشتر شنیده می‌شوند نسبت به قبل، که به طور دائمی بر بخش‌های مختلفی از جامعه تأثیر گذاشته است، از افراد تا مؤسسات. با هر ویژگی امنیتی پیشرفته، فریب‌گران راه‌های جدیدی برای نزدیک شدن به قربانیان پیدا می‌کنند.

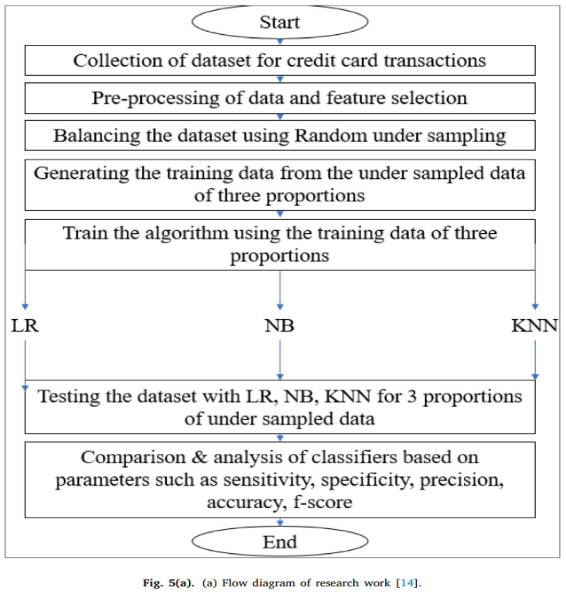
یکی از نقاط ضعف موجود در اطلاعات کارت اعتباری، انحراف داده است که باعث پیش‌بینی ناکارآمد کلاهبرداری‌های آتی می‌شود. این تحقیق انجام شده توسط فياض ایتو و همکاران (2020) از سه تقارن پایگاه داده برای هدف مطالعه استفاده می‌کند و علاوه بر این، یک روش زیرنمونه‌برداری بر روی پایه تصادفی برای مجموعه داده‌های انحرافی انتخاب شده است. تحقیق آزمایشی انجام شده توسط فياض ایتو و همکاران (2020) شامل سه الگوریتم، نزدیک‌ترین همسایه، نویو بیز و رگرسیون لجستیک است. معیارهای ارزیابی که توسط آن‌ها برای اندازه‌گیری مورد بررسی قرار گرفته اند شامل دقت، خصوصیت، حساسیت، اندازه‌گیری اف، مساحت زیر منحنی و دقت است. نتیجه نهایی تحقیق نشان داده است که رگرسیون لجستیک نتایج قابل اطمینان‌تری نسبت به دو الگوریتم دیگر استفاده شده را ارائه داده است.

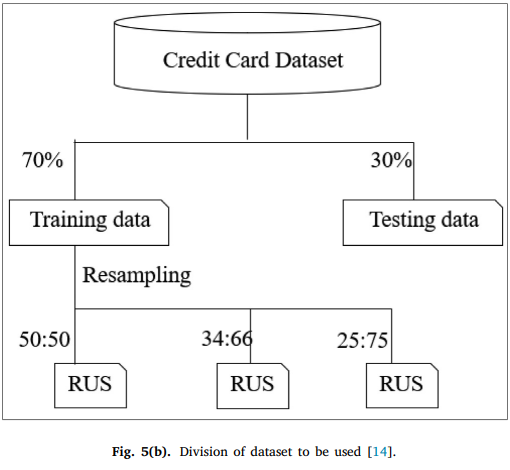
**2-2-2- توضيحات و نتايج**

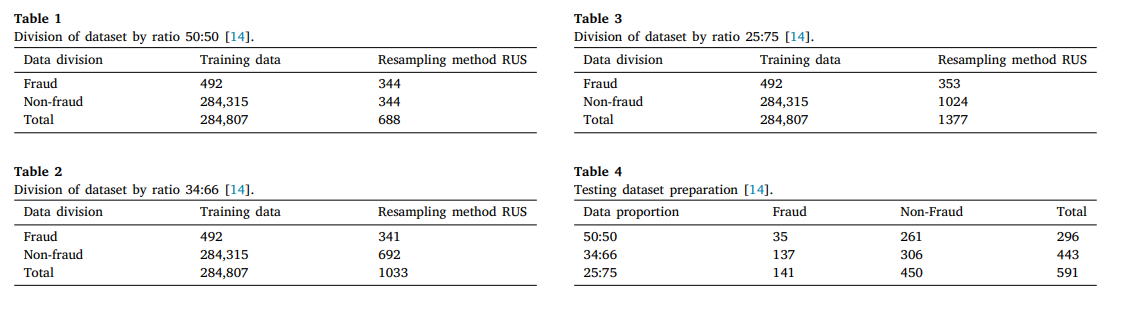
جریان روشی که برای این تحقیق پیروی شده است، در شکل ۵ (a) نشان داده شده است. تقسیم مجموعه داده به دو بخش، نسبت‌های استفاده شده برای هر دو، ۵۰:۵۰، ۳۴:۶۶ و ۲۵:۷۵ می‌باشد، جایی که توزیع از داده‌های کلاهبرداری به داده‌های غیر-کلاهبرداری است.

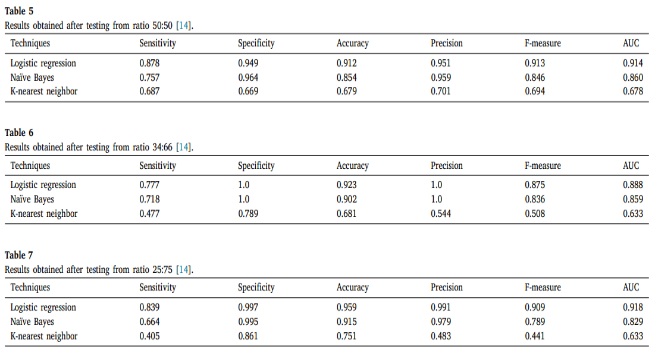
تقسیم‌بندی را می‌توان در شکل ۵ (b) دید. علاوه بیشتر، جداول ۱، ۲ و ۳ بیشتر درباره تقسیم‌بندی مجموعه داده توضیح می‌دهند (برای اطلاعات بیشتر به جدول ۴ مراجعه کنید). از جداول ۵، ۶ و ۷ مشاهده می‌شود که الگوریتم رگرسیون لجستیک نتایج دقیق‌تر و قابل اعتماد‌تری نسبت به الگوریتم‌های نویو بیز و K-NN ارائه داده است.

الگوریتم K-NN همانطور که از شکل‌های بالا مشخص است، بدترین عملکرد را از بین تمام الگوریتم‌ها نشان داده است، این به خاطر مجموعه داده‌آموزشی نمونه‌ای کوچک است، زیرا شباهت بالایی بین داده‌های کلاهبرداری و غیر-کلاهبرداری وجود دارد و بنابراین الگوریتم قادر به طبقه‌بندی به صورت کارآمد بین این دو دسته نبوده است.









**2-3- مزايای الگوریتم K-NN**

الگوریتم K-NN یک الگوریتم آسان برای حل مسائل است. این الگوریتم مقاوم و تحمل‌پذیر نسبت به نویز موجود در مجموعه داده استفاده شده برای آموزش می‌باشد. الگوریتم K-NN سریع، آسان برای تفسیر و موثر است حتی اگر مجموعه داده به اندازه کافی بزرگ باشد.

**4-2- معایب الگوریتم K-NN**

تصمیم‌گیری برای انتخاب مقدار مناسب برای K پیچیدگی است، زیرا گاهی نتایج را به شدت تغییر می‌دهد. زیرا نیاز است که فاصله نوع اقلیدسی بین هر نقطه داده‌ای متعلق به مجموعه داده استفاده شده برای آموزش محاسبه شود، که منجر به هزینه‌ بالای محاسبه شده می‌شود.

**3- الگوریتم ژنتیک**

## در دهه 1950، ریاضیدان انگلیسی به نام الن تورینگ یک دستگاه معرفی کرد که قرار بود نظریه‌ها یا اصول تکاملی را شبیه‌سازی کند. شبیه‌سازی‌های وابسته به تکامل کامپیوتری به‌وسیله نیلز آل باریسلی در سال 1954 آغاز شد، که از دستگاه‌ها و کامپیوترهای موجود در دانشگاه پرینستون در مؤسسه مطالعات پیشرفته استفاده می‌کرد. اما، در میان مخاطبان به خوبی شناخته نشد. پس از آن، در سال 1957، متخصص ژنتیک کمی الکس فریزر اهل استرالیا، مجموعه‌ای از مقالات مرتبط با شبیه‌سازی انتخاب مصنوعی ارگانیسم‌ها را کار کرد و منتشر کرد. پس از آن، شبیه‌سازی‌های کامپیوتری مرتبط با تکامل توسط بیولوژیست‌های مختلف در دهه 1960 به وجود آمدند و تکنیک‌ها در متون فریزر و بورنل و کرسبی منتشر شد و تمامی جنبه‌های اصلی الگوریتم‌های ژنتیک پوشش داده شد.

## علاوه بر این، مجموعه‌ای از مقالات منتشر شده توسط هانس-یواخیم برمرمن حاوی تنوع گسترده‌ای از راه حل‌ها برای مسائل مربوط به انتخاب، جهش و بازترکیبی که به بهینه‌سازی وابسته است، بود. همچنین، جنبه‌های مربوط به الگوریتم‌های ژنتیک مدرن نیز توسط برمرمن در کار تحقیقی خود پوشش داده شد. تا دهه 1970 تکنیک تکامل مصنوعی تا آنجا که باید شناخته نشده بود تا زمانی که اینگو ریچنبرگ و هانسپل شوفل پژوهش‌هایشان را در دهه 1960 و 1970 ارائه کردند و ریچنبرگ و گروهشان به طور صحیح راه حل‌هایی برای مواقع پیچیده مهندسی از طریق اصول ژنتیک و تکامل فراهم کرده بودند. روشی جایگزین برای مسائل تکاملی توسط لارنس جی. فوگل ارائه شد، اصولاً برای تولید هوش مصنوعی. در ابتدا، مفهوم برنامه‌ریزی تکاملی از ماشین‌های حالت متناهی برای پیش‌بینی محیطی استفاده می‌شد و تکنیک‌های انتخاب و تغییر برای طراحی پیش‌بینی استفاده می‌شدند.

## در نهایت، در اوایل دهه 1970 جان هالند بود که قادر بود الگوریتم‌های ژنتیک را از طریق کتابش تطبیق در سیستم‌های طبیعی و مصنوعی منتشر کند که جریان کار او با تحلیل سلولی خود، که به صورت شخصی توسط او و دانشجویانش انجام شد، شروع شد. مطالعات و تحقیقات مربوط به الگوریتم‌های ژنتیک اصولاً تئوری ای بودند تا اواسط دهه 1980 که در پیتسبورگ، پنسیلوانیا اولین کنفرانس بین‌المللی در مورد الگوریتم‌های ژنتیک برگزار شد.

## براساس مفاهیم بیولوژیکی مهم انتخاب طبیعی و وراثت، الگوریتم‌های ژنتیک بنیان‌های، الگوریتم‌های جستجو و بهینه‌سازی هستند. آن‌ها را می‌توان به عنوان یک دسته بازیافت شده از یک دامنه نسبتاً گسترده محاسبه به نام محاسبات تکاملی خواند. الگوریتم ژنتیک عمدتاً یک الگوریتم بهینه‌سازی مبتنی بر احتمال است. مشابه ژنتیک در زیست‌شناسی، در اینجا، چندین راه حل که بدست آمده، جهش و بازترکیبی را تجربه می‌کنند که در نهایت منجر به زاییدن جدیدترین ترکیبی می‌شوند، تازه‌نوزادان، که توسط تکرار این فرآیند برای چند نسل بعد به دست می‌آیند.

## هر فرزند از میزان تناسبی تعیین شده برخوردار است که توسط ارزیابی تابع هدف آن اندازه‌گیری می‌شود و در نهایت، افراد پرتراکم‌تر احتمال بیشتری برای تولید نسلی با تناسب بیشتر دارند. این تکنیک تضمین می‌کند که موجودیت‌های یا راه حل‌های در نسل‌های پی رو به صورت صحیح‌تر درست می‌شوند تا نسل نهایی به دست آید. تولید فرزندان براساس اصل زیر انجام می‌شود:

## 1. کاراکترها یا اجسام برای منابع تلاش می‌کنند و سپس تولید می‌شوند.

## 2. کاراکترها با امتیاز تناسب بالاتر نسل می‌کنند تا فرزندان تولید کنند.

## 3. بهترین ژن‌ها از کروموزوم‌های پدری به نسل‌های پی رو منتقل می‌شوند.بنابراین، با پیشرفت هر نسل جدید، آن‌ها بهتر و مناسب‌تر برای محیط حاکم می‌شوند.

**1-3- فضای جستجو**

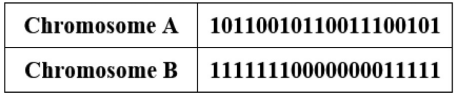
تمام جمعيت در منطقه خاصی که فضای جستجو ناميده می‌شود محدود شده است. هر موجود حاضر در اينجا دارای يک کليد يا راه حل برای مسئله داده شده است. هر کروموزوم به عنوان يک بردار با طول منظوری رمزگذاری می‌شود. پس از انتخاب و ايجاد نسل اوليه، الگوريتم ژنتيک منجر به تکامل گروه می‌شود، با استفاده از فرايندهای انتخاب، گذردهی و جهش.

**2-3- انتخاب**

در این فرایند، اساساًکروموزوم‌هایی که امتیاز تناسب بالاتری دارند، جستجو می‌شوند و اجازه داده می‌شود نسل‌های پی روی بهتر و رقابتی‌تر را تولید کنند تا ژن‌های بهتر و جذاب‌تری را منتقل کنند.

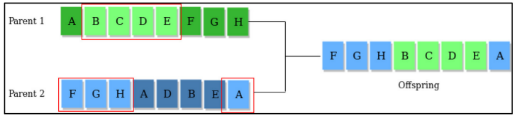
**3-3- رمزگذاری**

قالبی که یک کروموزوم به آن دست یافته است، دارای داده‌های مربوط به خروجی یا راه‌حلی است که نمایانگر آنها است. یکی از روش‌های متداول برای رمزگذاری، استفاده از یک رشته دودویی است که در شکل ۶ نشان داده شده است. هر کروموزوم می‌تواند از این فرمت رمزگذاری شود. هر بیت حاضر در رشته، شامل بخشی از راه‌حل خروجی است.



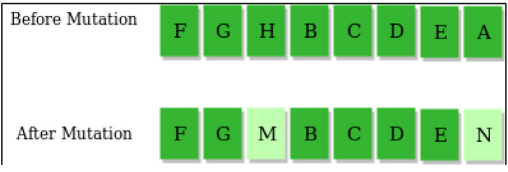
**4-3- تلاقی**

در گذردهی، دو کروموزوم پدری از طریق فرایند انتخاب، انتخاب شده و نقطه تصادفی برای گذردهی ژن‌ها مشخص می‌شود. بعد از انجام گذردهی، جدیدین نوزاد به وجود می‌آیند.



**-5-3جهش**

برای جلوگیری از همگرایی زودرس جمعیت، ژن‌های تصادفی وارد نوزاد‌های تازه تولید شده می‌شوند تا انواع موجود در جمعیت تشویق شوند.



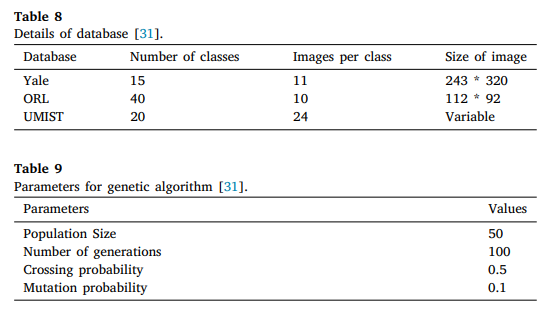
**6-3- تشخيص چهره بر اساس بهينه سازی الگوريتم ژنتيک- برنامه کاربردی اخير**

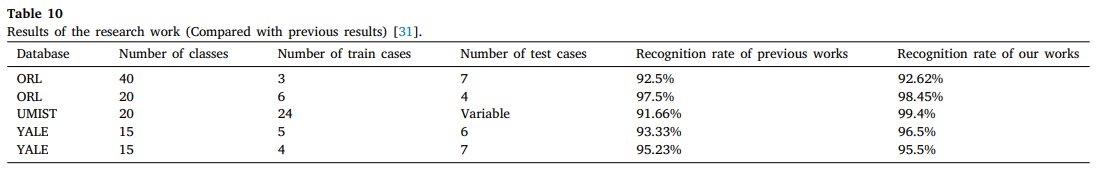
**1-6-3- پس زمینه کار اخیر**

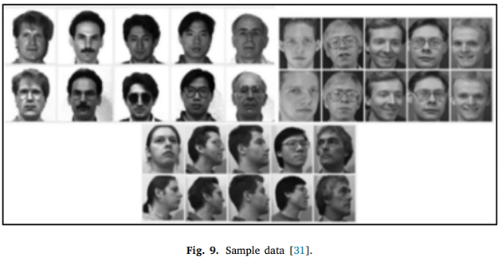
"موراد موسی و همکاران (۲۰۱۸) ، بر روی روش تشخيص چهره بر اساس تحليل مولفه اصلی معروف و روش تبديل کوزين متمرکز انجام دادند. برای توسعه يک عمليات تشخيص چهره پايدار و قابل اطمينان، لازم است ابتدا به انتخاب ويژگی ها توجه کرد، که مسئول لغو سر و صداهای غير ضروری، داده های اضافی و ويژگی های متعدد ديگر نامربوط هستند. با اين حال، توسعه الگوريتم ژنتيک، که الگوريتمی نسبتاً جديدتر برای انتخاب ويژگی ها است، می تواند برای رفع اين مسئله استفاده شود. برای استفاده از الگوريتم ژنتيک به منظور حل يک مسئله، لازم است راه حل های موثر را در زنجيره های بيت قابل اندازه گير کد کرد تا شامل کروموزوم های آمده از نقاط خاص شوند. هدف نهايی استفاده از اپراتورهای ژنتيک و توسعه تمييز معقولی ميان کروموزوم ها است. موراد موسی و همکاران (۲۰۱۸) ، يک سيستم تشخيص چهره را با استفاده از الگوريتم ژنتيک در کنار ترکيبی از تحليل M تبديل تبديل کوزين متمرکز - تحليل مولفه اصلی PCA-DCT طراحی کردند که برای کاهش بعد و انتخاب ويژگی بر روی يک مجموعه تصاوير چهره انسان به کار برده شد. نتايج ارائه شده توسط موراد موسی و همکاران (۲۰۱۸) ، کارايی اين روش را نسبت به کارهای قبلی نشان می دهد."

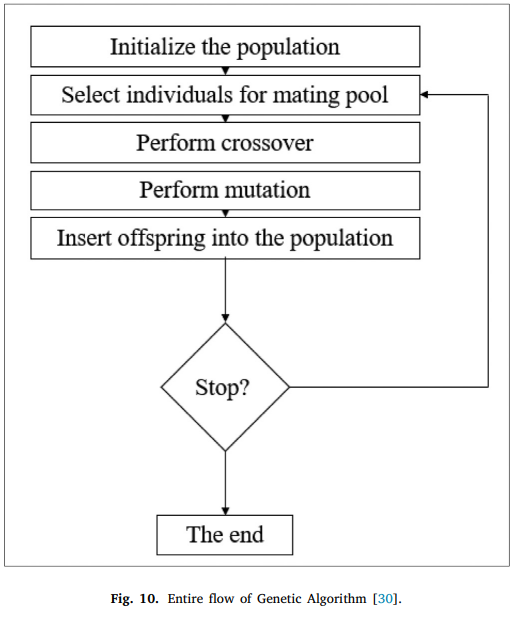
**2-6-3- شرح و نتایج**

موراد موسی و همکاران (۲۰۱۸) ، از سه طرح استاندارد به نام موسسه علم و فناوری دانشگاه منچستر، UMIST آزمایشگاه تحقیقاتی اولیوتی ORL و ییل که در جدول ۸ زیر ارائه شده است، به همراه چهره‌های نماینده (شکل ۹) که برای تست استفاده شدند، استفاده کردند. پایگاه داده‌ها به طور تصادفی برای آموزش یا برای مجموعه‌های تست استفاده شدند و تمام ترتیبات مصلحی برای پژوهش استفاده شدند. نتایج میانگین و مشاهدات ارائه شده است. برای این پژوهش آزمایشی از نسخه MATLAB a۲۰۱۵R استفاده شد و ژن‌های قفل شده به ۳۰ گرفته شد. متغیرهای مختلف دیگری که برای پژوهش مورد نظر بررسی شده اند، در جدول ۹ نشان داده شده اند. نتایج ارائه شده توسط موراد موسی و همکاران (۲۰۱۸) در جدول ۱۰ ارائه شده اند و رویکرد تعقیب شده توسط آنها کمک کرده است تا این سیستم شناسایی چهره به نرخ تشخیص ۹۹٪ برسد. آشکار است که این روش نوین منجر به بهبود ۸٪ نسبت به کارهای قبلی شده است. بنابراین، این رویکرد مبتنی بر الگوریتم ژنتیک باعث موفقیت در بهبود کارایی و سرعت این سیستم تشخیص چهره شده و در انتخاب مناسب ضرایب مورد نیاز کمک کرده است.









7-3- مزایای الگوریتم ژنتیک

لگوریتم ژنتیک عملکرد بسیار پایداری در مقایسه با خروجی‌های محلی بیشینه یا کمینه فراهم می‌کند. آن‌ها ارتقاء داده‌های بزرگ فضای حالت را فراهم می‌کنند. در مقایسه با سیستم‌های هوش مصنوعی سنتی، آن‌ها نسبت به ورودی‌ها، ورودی‌های متغیر و سر و صدا ضعیف نمی‌شوند. الگوریتم‌های ژنتیک توابع متمایز و ناپایدار را بهبود می‌بخشند. این الگوریتم به داده‌ها یا اطلاعات تقلیدی نیاز ندارد. از نظر گسترده‌تر و بهینه‌تر، چندبرابر با روش‌های ابتدایی است.

8-3- معایب الگوریتم ژنتیک

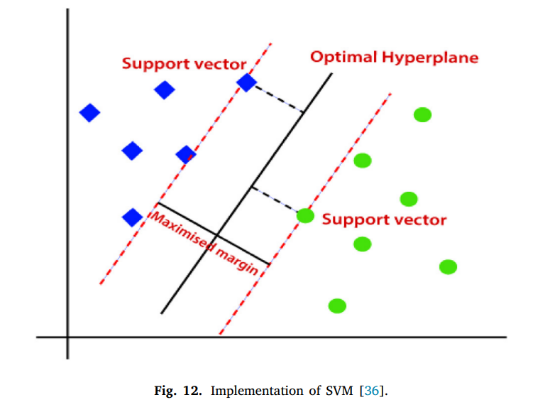
یکی از معایب محتمل الگوریتم ژنتیک این است که اغلب می‌تواند منجر به همگرایی زودرس جمعیت شود، به دلیل یکنواختی ژن‌ها. این امر هر گونه تحقیق مفیدی را باز می‌دارد. با اینکه این الگوریتم به اندازه زیادی به اطلاعات در مورد بیان مسئله نیاز ندارد، اما طراحی یک تابع هدف و دستیابی به عملیات چالشی است. اعمال الگوریتم ژنتیک زمان‌بر است.

**4- ماشین بردار پشتیبان**

مسائل جداسازی و طبقه بندی تصویر، هايپرمتن و متن توسط الگوریتم ماشین بردار پشتیبان (SVM) پوشش داده می‌شوند. SVM الگوریتمی پیشرفته است که در تشخیص متن دست‌نوشته و مرتب‌سازی پروتئین‌ها در آزمایشگاه‌های زیست‌شناسی نیز استفاده می‌شود. این الگوریتم در حوزه‌های مختلفی از جمله خودروهای خودران، چت‌بات‌ها، و تشخیص چهره نیز کاربرد دارد.

SVM منطبق بر یک هایپرپلاین مناسب معروف به هایپرپلاین، که فضای بعدی-n را به کلاس‌های مختلف تقسیم می‌کند و نقاط مختلف را در رده‌های مناسب می‌گذارد. Support Vectors نام گرفته به نقاط بردار استوانه‌ای کمک می‌کنند تا یک هایپرپلاین مناسب ایجاد شود.

الگوریتم SVM برای مسائل رگرسیون و طبقه‌بندی طراحی شده است و از آن برای تشخیص چهره‌ها، دسته‌بندی تصاویر، و دسته‌بندی متن‌ها استفاده می‌شود. الگوریتم SVM می‌تواند مفید باشد برای تشخیص مواردی مانند آیا یک تصویر سگ است یا گربه با ویژگی‌های مشابه. به کل، SVM یک الگوریتم پرکاربرد و تاثیرگذار در حل مسائل جداسازی و طبقه‌بندی می‌باشد.



**1-4- انواع بردار پشتیبان**

**1-4-4- نوع خطی ماشين‌های‌ بردار پشتيبان**

لگوریتم SVM نوع خطی مفید است در مواردی که داده‌ها باید به صورت خطی جدا شوند، به این معنا که مجموعه داده می‌تواند به دو کلاس تقسیم شود که توسط یک خط مستقیم جدا شوند.

**2-1-4- نوع غيرخطی ماشين های بردار پشتيبان**

الگوریتم SVM نوع غیرخطی مفید است در مواردی که داده‌ها به صورت غیرخطی جدا شوند، به این معنا که مجموعه داده نمی‌تواند با استفاده از یک خط مستقیم به کلاس‌ها تقسیم شود.

**2-4- هايپرپلاين و بردارهای پشتيبان در الگوريتمSVM**

**1-2-4- هایپرپلاین**

هایپرپلاین در الگوریتم SVM به عنوان بهترین مرز تصمیم تعریف می‌شود که بین مرزهای تصمیم ممکن، به روش دقیقی کلاس‌ها را در فضای بعدی-n دسته‌بندی می‌کند. ویژگی‌های مجموعه داده تعیین کننده ابعاد هایپرپلاین هستند؛ به این معنا که اگر مجموعه داده دو ویژگی داشته باشد، هایپرپلاین یک بعدی است و اگر سه ویژگی داشته باشد، هایپرپلاین دو بعدی است. هایپرپلاینی که حاشیه‌های بیشینه دارد، به این معنا است که فاصله بیشینه بین دو نقطه داده را کمینه می‌کند و در نتیجه اولویت دارد.

**2-2-4- بردارهای پشتیبان**

بردارهای پشتیبان نقاط داده‌ای هستند که نزدیک‌ترین به موقعیت هایپرپلاین قرار دارند و تحت تأثیر قرار گرفتن آنها، موقعیت هایپرپلاین تغییر می‌کند. به دلیل حمایت و تاثیر قوی‌شان بر روی موقعیت هایپرپلاین، این نقاط داده به عنوان بردارهای پشتیبان شناخته می‌شوند.

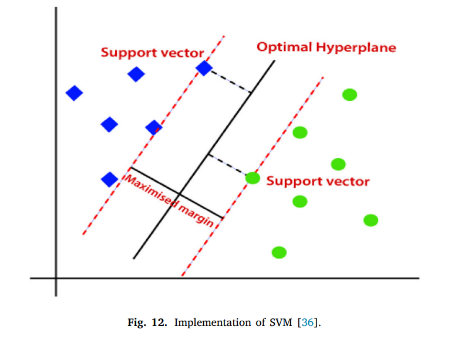
**3-4- عملکرد الگوريتم SVM**

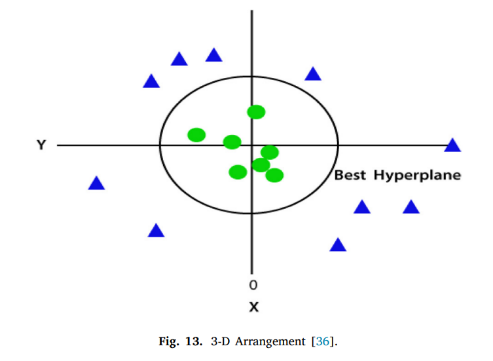
**1-3-4- SVM خطی**

مدل کارایی الگوریتم SVM می‌تواند با یک مثال توضیح داده شود. فرض کنید یک مجموعه داده دارای دو شئ مختلف (قرمز و زرد) و دو ویژگی، مانند ۱X و ۲X است. یک الگوریتم دسته‌بندی که بتواند جفت مختصات و ۱X و ۲X را به درستی در یکی از دو رنگ قرمز یا زرد جدا کند، مطلوب است. از آنجا که این یک فضای دو بعدی است و دارای دو ویژگی است، بنابراین راحت تر است که این دو دسته را فقط با یک خط مستقیم تفکیک کرد، اما بسیاری از خطوط مستقیم امکان پذیر هستند. نقش این الگوریتم در اینجا معرفی می‌شود که خط تصمیم مناسب‌ترین را از بین تمام خطوط یا مرزهای تصمیم انتخاب می‌کند؛ این مرز تصمیم بهترین مرز تصمیم نام دارد. الگوریتم SVM با استفاده از بردارهای پشتیبان، نزدیک‌ترین نقاط از مرز تصمیم در دو کلاس را مشخص می‌کند و فاصله بین هایپرپلاین و بردارها را بیشینه می‌کند تا یک راه حل بهینه را تأمین کند.

**SVM 2-3-4 غيرخطی**

در نظر داشته باشید که داده‌ها به صورت غیرخطی ترتیب داده شده‌اند. در اینجا نمی‌توان به سادگی یک خط مستقیم رسم کرد. بنابراین، برای جدا کردن این نقاط داده، نیاز به یک بعد دیگر وجود دارد. برای داده‌های به صورت خطی، تنها دو بعد (x و y) استفاده شده است، اما برای داده‌های به صورت غیرخطی، یک بعد سوم اضافه می‌شود. از آنجا که کار در یک فضای سه بعدی انجام می‌شود، این در واقع یک صفحه است که موازی با محور z قرار دارد. با تبدیل آن به یک فضای دو بعدی با z = 1، یک دایره با شعاع 1 واحد به دست می‌آید که در شکل 13 (شکل 12 را ببینید) نشان داده شده است.





**4-4- استفاده از الگوریتم SVM برای تشخیص سرطان پستان-کاربرد اخیر**

**1-4-4- پیش‌زمینه کاراخیر**

جینی ای. ام. سایدی-گیبونز و همکاران (2019) تحقیقات دقیقی در مورد الگوریتم‌های خاص یادگیری ماشین که می‌توانند برای پیش‌بینی سرطان، به ویژه سرطان پستان، استفاده شوند، انجام دادند. الگوریتم‌های یادگیری ماشین بسیار کارآمد هستند و می‌توانند در علوم پزشکی برای تشخیص زودرس یا پیش‌داوری بسیاری از بیماری‌های فاتال مفید باشند.

در تحقیقات آزمایشی آن‌ها، طرح‌های پیش‌بینی مبتنی بر الگوریتم‌های مختلف برای تشخیص سرطان بر اساس موادی که از جرم پستان استخراج شده، انجام شد. الگوریتم‌های استفاده شده در کار تحقیقاتی آن‌ها شامل شبکه‌های عصبی مصنوعی تک لایه، الگوریتم ماشین بردار پشتیبان با هسته تابع پایه گرد، و مدل تخطی عمومی (GLM) بود. تقریباً 456 نمونه از جرم‌های پستان برای ارزیابی و 227 نمونه برای اعتبارسنجی استفاده شدند. قبل از آزمایش الگوریتم‌ها و مدل‌ها در مجموعه اعتبارسنجی برای تشخیص بیماری، آن‌ها با استفاده از نمونه‌های ارزیابی آموزش دیده شدند.

به منظور مقایسه عملکردهای مدل‌های الگوریتمی موردنظر، معیارهای ارزیابی که توسط جینی ای. ام. سایدی-گیبونز و همکاران (2019) استفاده شد، حساسیت، خصوصیت و دقت بودند. پس از تحقیقات انجام شده توسط آن‌ها، مشخص شد که الگوریتم SVM بیشینه مساحت زیر منحنی و دقت را نسبت به دو الگوریتم دیگر فراهم کرد.

**2-4-4- توضیحات و نتایج**

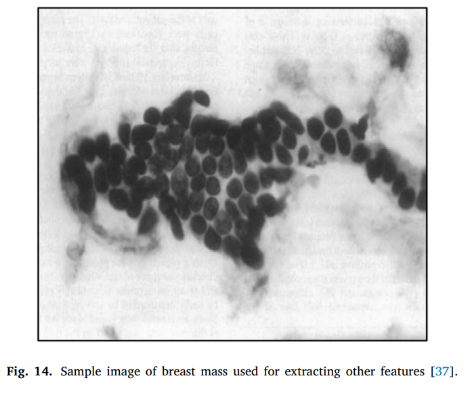
مجموعه داده‌ای که برای تحقیقات آن‌ها استفاده شد، مجموعه داده سرطان پستان ویسکانسین بود که از طریق مخزن یادگیری ماشین دانشگاه کالیفرنیا آیرون که به صورت رایگان در دسترس است، قابل دسترسی است.

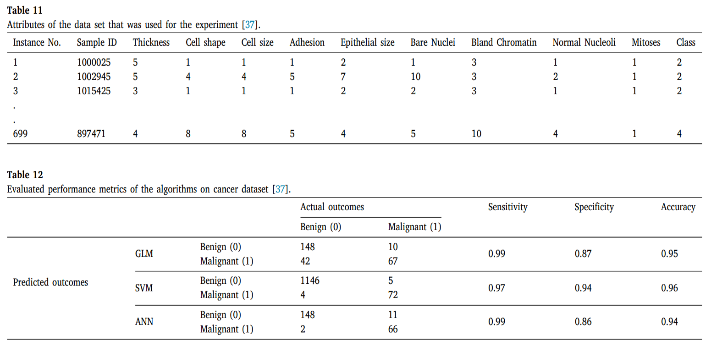
ویژگی‌های جرم‌های پستان که در این تحقیق به آن‌ها تمرکز داده شده‌است، هسته‌های سلولی هستند که تحت انجام سوزن-نمونه‌برداری دقیق (FNA)، یک روش تشخیصی پزشکی استاندارد در زمینه انکولوژی تجربی شدند. یک شکل نمونه از جرم پستان در شکل 14 نشان داده شده است. تعداد نمونه‌های استفاده شده برای این تحقیق در واقع تعداد نمونه‌هاست که به یک شناسه یکتا (ID) اختصاص یافته‌اند و به جز آن، ویژگی‌های مختلف دیگر متصل شده‌اند. ستون کلاس نمایش داده شده در شکل حاصل، تشخیص است که بیماری سرطانی بدخیم یا مهلک است که بستگی به نمونه‌برداری ضخیم سوزن FNA که سرطانی بودن آن یا خیر بود، دارد.

همان طور که می‌توان در جدول 11 دید، 241 نمونه به خوبیم و 458 نمونه هم پیوندیم بوده‌اند. نمونه‌های هم پیوندیم دارای کلاس دو، در حالی که نمونه‌های خوب‌خیم دارای کلاس چهار هستند.

در این تجربه نه ویژگی برای آزمون وجود دارد که در جدول 11 نشان داده شده‌است، که هر ویژگی بر اساس یک مقیاس از 1 تا 10 ارزیابی شده است. هرچه مقدار به 10 نزدیک‌تر باشد، ویژگی طبیعتاً بدخیم است و هرچه مقدار به 1 نزدیک‌تر باشد، ویژگی طبیعتاً مهلک است.

با وجود اینکه تمام الگوریتم‌ها یک سبک کاری بسیار متنوع دارند، آن‌ها سطحی مناسب از دقت، حساسیت و خصوصیت را در عملکرد خود نشان داده‌اند و الگوریتم SVM بهترین عملکرد را با دقت 96.0 در جدول 12 نشان داده شده‌است.





**5-4- مزایای الگوریتم SVM**

الگوریتم SVM بیشترین تطابق را در مواردی ارائه می‌دهد که تقسیم واضحی بین کلاس‌ها وجود دارد. SVM عملکرد بهتری در فضاهای با ابعاد بالاتر نشان می‌دهد و در مواقعی که تعداد فضاهای بعدی از مقدار نمونه‌های موجود بیشتر باشد، به طور مؤثر عمل می‌کند. از نظر حافظه، SVM نسبت به سایر ویژگی‌ها کاربردی واقعی است.

**6-4- معایب الگوریتم SVM**

الگوریتم SVM مناسب برای مجموعه داده‌های بزرگ نیست. در شرایطی که مجموعه داده‌ها دارای حجم زیادی از سر و صدا هستند، عملکرد مناسبی ندارد که این مسئله خیلی اغلب در عمل رخ می‌دهد. SVM در مواردی که مقدار عددی ویژگی‌های هر نقطه داده بیشتر از نمونه‌های داده آموزشی باشد، عملکرد ضعیفی دارد. الگوریتم SVM برای طبقه‌بندی صورت گرفته توسط آن توجیه احتمالی نمی‌دهد.

**5- درخت تصمیم**

الگوریتم درخت تصمیم (DT) که به دسته الگوریتم‌های یادگیری نظارت شده تعلق دارد، اغلب برای حل مسائل طبقه‌بندی استفاده می‌شود، اما همچنین می‌توان از آن برای هر دو حالت طبقه‌بندی و پیش‌بینی استفاده کرد. این الگوریتم شامل گره‌های داخلی که ساختار شاخه‌ها را نشان می‌دهند، مجموعه داده که نتیجه‌ای ارائه شده توسط الگوریتم را نشان می‌دهد، و هر گره برگ که یک نتیجه را نمایندگی می‌کند، است.

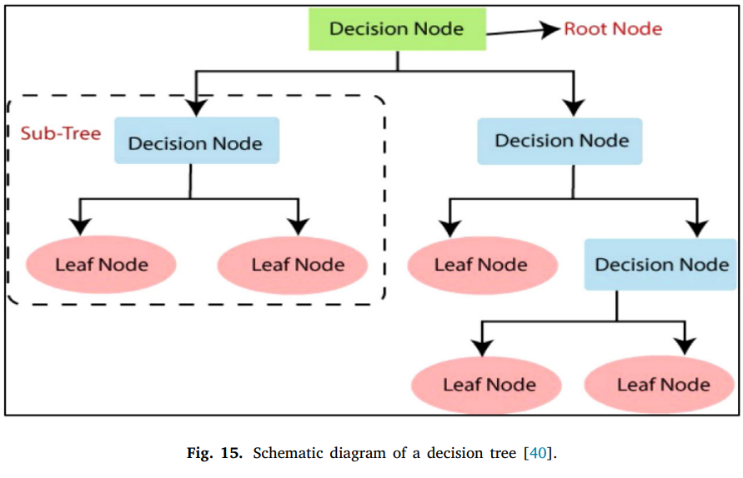
دو نوع گره وجود دارد: اول، گره تصمیم که برای تصمیم‌گیری استفاده می‌شود و شاخه‌های مختلفی دارد؛ و دوم، گره برگ که خروجی گره‌های تصمیم است و دیگر شاخه‌هایی ندارد. این الگوریتم نام خود را به دلیل شباهتی که به یک درخت دارد بدست آورده است.

گره ریشه نقطه شروعی است که به شاخه‌های مختلفی توسعه می‌یابد و یک ساختار شبیه به یک درخت را شکل می‌دهند. درخت تصمیم به اختصار درخت را بر اساس پاسخ به سوالات تقسیم می‌کند، به سوالی از نوع بله یا خیر.

**1-5-** **طبقه‌بندی درخت‌های تصمیم**

**1-1-5- درخت تصمیم دارای متغیر خوشه‌ای**

درخت تصمیم دارای متغیر خوشه ای به عنوان هدف. مثال: - جمله مشکل با داشتن متغیر هدف به عنوان “آیا با پرتاب سکه، شیر ظاهر می شود یا خیر“ (مشاهده شکل ۱۵).



**2-1-5- درخت تصميم دارای متغير ثابت**

درخت تصمیم دارای متغیر ثابت به عنوان هدف. مثال: - آیا شخص می‌تواند یک وام را پس بدهد یا نه. در صورتی که بانک‌ها اطلاعات درآمد را نداشته باشند، که یک متغیر مهم در این مورد است، آنگاه می‌توان یک درخت تصمیم برای پیش‌بینی درآمد ماهانه یک فرد بر اساس عوامل مختلفی مانند دارایی‌ها، استاندارد زندگی، شغل و غیره برپا کرد. در اینجا مقادیری که پیش‌بینی می‌شوند برای متغیرهای پیوسته نوعی است.

**2-5- اصطلاحات درخت تصميم**

گره ریشه: بخش ابتدایی از درخت تصمیم که از آن شروع به تقسیم کل داده می‌شود و وارد مجموعه‌های مختلف ممکن می‌شود که همگن هستند.

گره برگ: گره نهایی به عقب که دیگر تقسیم درختی امکان پذیر نیست. تقسیم: شامل فرآیند تقسیم گره اصلی به زیرگره‌ها بر اساس محدودیت‌های ارائه شده می‌شود.

زیردرخت: تقسیم یک سلسله‌مراتب به یک زیردرخت یا شاخه منجر می‌شود. قالب‌بندی: شامل حذف شاخه‌های بیش از حد از درخت تصمیم به منظور به‌دست‌آوردن نتایج بهینه است. در واقع، اندازه درخت را بدون تأثیر بر دقت کاهش می‌دهد. این از دو نوع، قابلیت هزینه و قبول خطا تقسیم‌بندی است. گره والدین و کودک: این گره پایه نامیده می‌شود که همچنین گره والدین نیز نامیده می‌شود، درحالی که گره‌های باقی‌مانده به سادگی گره‌های کودک نامیده می‌شوند.

**3-5- اندازه گيری انتخاب ويژگی ها**

اندازه‌گیری انتخاب ویژگی (ASM) شامل جمع‌آوری ویژگی بهینه مربوط به گره منبع و همچنین زیرگره‌ها است. دو عملکرد اصلی برای ASM عبارت‌اند از:

**1-3-5- بهره اطلاعات**

بهره اطلاعات همانطور که از نام پیداست، میزان اطلاعاتی که توسط یک ویژگی در مورد کلاس ارائه می‌شود را محاسبه می‌کند. گره تقسیم می‌شود و درخت بر اساس ارزش‌های بهره اطلاعاتی ساخته می‌شود.

**2-3-5- شاخص Gini**

شاخص Gini میزان خلوص یا اصالت را که در ایجاد یک الگوریتم درخت تصمیم استفاده می‌شود اندازه‌گیری می‌کند. ویژگی‌های کوچکتر از شاخص Gini بهتر از ویژگی‌هایی با شاخص Gini بزرگتر توسط الگوریتم درخت تصمیم در هنگام تصمیم‌گیری ترجیح داده می‌شوند.

**4-5- مراحل ساخت درخت تصمیم**

گره ریشه، به نام "X" که شامل کل مجموعه داده است، به عنوان نقطه شروع درخت در نظر گرفته می‌شود. با استفاده از ASM بهترین ویژگی مطابق از مجموعه داده را جستجو نمایید. "X" را به زیربخش‌هایی با مقادیر با کیفیت بهتر تقسیم می‌کند. فقط با استفاده از ویژگی ایده‌آل، گره‌های درخت تصمیم را توسعه دهید. به تکرار گره‌های منحصر به فرد درخت تصمیم، با استفاده از زیرمجموعه‌های موجود مجموعه داده ایجاد شده در "۳"، توسعه دهید. این فرایند را ادامه دهید تا به نقطه ای برسید که دیگر امکان نداشته باشید به زیربخش‌ها برسید. این گره نتیجه‌ی نهایی نهایی به عنوان گره برگ شناخته می‌شود.

**5-5- پيش بينی بيماری کبد با استفاده از تکنيک های مختلف درخت تصميم-برنامه نويسی اخير**

**1-5-5- زمینه کاراخیر**

بیماری‌های مرتبط با کبد یکی از بیماری‌های مهلکی است که می‌تواند بر جان انسان‌ها تأثیر گذار باشد. کشف هر گونه تکنولوژی که بتواند این گونه بیماری‌ها را در مراحل ابتدایی پیش بینی کند، برای نجات جان انسان‌ها بسیار مفید است. نازمون نهار و همکاران (2018)، این تحقیق را در این زمینه انجام داده اند با اختلاف و مقایسه انواع مختلفی از الگوریتم‌های درخت تصمیم برای کمک به پیش بینی بیماری کبد در مراحل ابتدایی. الگوریتم‌های درخت تصمیم در بسیاری از زمینه‌ها، به ویژه در حوزه علوم پزشکی، مورد استفاده قرار می‌گیرند.

نازمون نهار و همکاران (2018)، از مجموعه داده‌ای که شامل ویژگی‌هایی از قبیل بیلی روبین مستقیم، بیلی روبین کل، جنسیت، عامل سن، پروتئین‌های کلی و غیره استفاده کردند. تکنیک‌های درخت تصمیمی که در این تحقیق آزمایشی مورد آزمایش قرار گرفتند، شامل درخت مدل لجستیک (LMT)، درخت تصادفی، درخت تصمیم دهانه، جنگل تصادفی، درخت خطای کم کننده و درخت بازتراشی خطا (REPTree) بودند. مطالعه تجربی آنها نشان داد که درخت تصمیم دهانه نتایج مطمئن و دقیق‌تری ارائه کرده است.

**2-5-5- توضیحات و نتایج**

هدف اصلی این تحقیق کشف این است که آیا یک بیمار تحت تأثیر بیماری‌های مرتبط با کبد قرار دارد یا خیر، با استفاده از انواع مختلف الگوریتم‌های درخت تصمیم است. تکنیک‌های مختلف براساس معیارهای ارزیابی مختلفی نظیر دقت، خطا مطلق میانگین، آماره کاپا، زمان اجرا، دقت، بازخاصیت و غیره آزمایش و مقایسه شدند. نازمون نهار و همکاران (2018)، از ابزار استخراج داده قدرتمندی به نام ویکا استفاده کردند تا دقت انواع الگوریتم‌ها را با استفاده از آنها روی مجموعه داده‌های مختلف آزمایش نمایند.

**6-5- مزايای الگوريتم درخت تصميم**

پیچیدگی بسیار پایین این الگوریتم بسیار ساده قابل فهم است و نیازی به دانش ویژه مرتبط با آمار برای تفسیر آن ندارد. مفید برای کاوش داده - همچنین می‌تواند در مراحل کاوش داده مورد استفاده قرار گیرد زیرا الگوریتم درخت تصمیم یکی از سریع‌ترین الگوریتم‌ها در ایجاد یا شناسایی ویژگی‌های جدید است.

به طور مقایسه‌ای نیاز کمتری به مراحل تمیزکاری داده دارد و تحت تأثیر مقادیر و داده‌های گم شده نیست. بدون محدودیت نوع داده - این قادر است به طور انعطاف پذیر با متغیرهای عددی و همچنین متغیرهایی با طبیعت خوشه‌ای کنار بیاید. روش غیر پارامتریک - درخت تصمیم از یک روش غیر پارامتریک استفاده می‌کند، که به معنی عدم وابستگی به هیچ گونه فرضیه‌ای درباره توزیع فضایی است.

**7-5- معایب الگوریتم درخت تصمیم**

بیش‌آموزش یکی از مسائل عملی اصلی بر روی مدل درخت تصمیم است. با این حال، با تنظیم پرونده و محدودیت پارامترهای مدل،مشکلات بیش‌آموزش می‌تواند کاهش یابد. نامناسب برای متغیرهای پیوسته - درخت تصمیم برخی از اطلاعات ارزشمند را از دست می‌دهد در هنگام دسته‌بندی متغیرها در دسته‌های مختلف.

**6- الگوريتم حافظه کوتاه-مدت بلند (LSTM)**

به دلیل پس‌انتشار با یادگیری مداوم واقعی یا زمانی، سیگنال‌های حاوی خطا که به سمت عقب در زمان حرکت می‌کنند، ممکن است ناپدید شوند یا بزرگ شوند؛ جابجایی‌های زمانی سیگنال حاوی خطا به طرز قابل توجهی به اندازه وزن‌ها بستگی دارد.

در صورت بزرگ شدن، وزن‌ها به احتمال زیاد شروع به نوسان کردن می‌کنند و در صورت ناپدید شدن، یا زمان مصرف شده برای یادگیری اتصالات با تاخیرهای زمانی بلند از حد بیرون می‌رود، یا در بدترین حالت به دلایلی کار نمی‌کند. به عنوان درمان، الگوریتم حافظه کوتاه-مدت بلند (LSTM)، نوعی جدید از شبکه‌های عصبی مکرر در سال ۱۹۹۱ به وجود آمد که توسط Sepp Hochreiter و Jurgen Schmidhuber توسعه یافت تا سیستم‌های موجود را پیش رویی کند و مشکلات پس‌انتشار خطا مورد بحث فوق را برطرف کند. نسخه اصلی این الگوریتم حافظه کوتاه-مدت بلند فقط شامل سلول‌ها، دروازه‌های ورودی و خروجی بود.

این الگوریتم قادر است تا در شکاف‌های زمانی بیش از گام‌ها پل روابر بیاند، حتی زمانی که دنباله‌های استفاده شده برای ورود، غیرقابل فشرده‌سازی یا نویزی هستند، در حالی که از از دست دادن توانایی شکاف زمانی کوتاه جلوگیری می‌کند. حافظه کوتاه-مدت بلند، طراحی شده توسط Hochreiter و Schmidhuber یک نوع ویژه از شبکه عصبی مکرر (RNN) است که در برابر وابستگی‌های طولانی مدت به طور پیش‌فرض همراه با آن مجهز است. در الگوریتم LSTM ورود یک گام کنونی خروجی گام قبلی است، و این امر با حل مشکلات وابستگی‌های طولانی مدت RNN که در آن RNN پیش‌بینی دقیقی از اطلاعات اخیر انجام می‌دهد اما قادر به پیش‌بینی داده‌های ذخیره‌ شده در حافظه طولانی مدت نیست، بهبود یافته است.

با افزایش طول شکاف، کارایی RNN کاهش می‌یابد. برخی از کاربردهای اصلی LSTM شامل توضیح تصاویر، تولید چت‌بات‌های خط نویسی برای پاسخ‌گویی به سوالات و موارد مختلف دیگر هستند.

**1-6- ساختار LSTM**

ساختار LSTM که شامل چهار شبکه عصبی و بلوک های حافظه

مختلفی به نام سلول ها است، در زير تصويب شده است. دروازه ها

تغييرات حافظه را بر داده های ذخيره شده در سلول ها انجام می دهند.

دروازه ها سه نوع هستند.

**1-1-6- دروازه فراموشی**

اطلاعاتی که دیگر نیازی به آن‌ها نیست، از سلول با استفاده از دروازه فراموش حذف می‌شوند. ورودی در یک زمان خاص، و خروجی سلول قبلی با استفاده از ماتریس‌های وزن‌دار ضرب می‌شوند و با اضافه شدن باياس، به بیرون می‌روند. برای دریافت یک خروجی دودویی، نتیجه از آنالیز یک عملکرد فعال‌سازی عبور می‌کند. اگر خروجی '۱' باشد، اطلاعات در حالت سلول حفظ می‌شود و اگر خروجی '۰' باشد، پاک می‌شود.

**2-1-6- دروازه ورود**

این دروازه وظیفه افزودن اطلاعات حیاتی به حالت سلول را انجام می‌دهد. اطلاعات از طریق یک تابع سیگموئید پردازش می‌شوند و مقادیری که باید نگه‌داشته شوند، تصفیه می‌شوند. مرحله بعد شامل ایجاد بردار با استفاده از تابع tanh می‌شود که یک خروجی از -۱ تا +۱ را تولید می‌کند که شامل تمام مقادیر ممکن از -1 تا +1 است. در نهایت، مقادیر بردار و نتایج تصفیه شده تابع سیگموئید با هم ضرب شده تا نتایج مفید مشتق شوند.

**3-1-6- دروازه خروجی**

بر اساس داده‌های ذخیره شده در حالت سلول فعلی، خروجی اعلام می‌شود. در ابتدا، با استفاده از تابع tanh یک بردار برای مقادیر سلولی ایجاد می‌شود. مرحله بعد شامل تنظیم اطلاعات با استفاده از تابع سیگموئید و تصفیه مقادیری که باید نگه‌داشته شوند است. در نهایت، حاصلضرب مقادیر تنظیم شده و مقادیر برداری به عنوان خروجی ارسال می‌شود که به عنوان ورودی برای سلول بعدی عمل می‌کند.

**2-6- عملکرد LSTM**

مرحله اول نیاز به تصمیم‌گیری در مورد حذف اطلاعات غیرضروری از حالت سلولی دارد. این تصمیم‌گیری‌ها توسط "لايه دروازه فراموش" که یکی از لایه‌های سیگموئیدی است، حل می‌شوند. در زمان تصمیم‌گیری، x و h در نظر گرفته می‌شوند و نتایج برای تمام اعداد متعلق به سلول C می‌تواند هر عددی در بازه ۰ تا ۱ باشد. در صورتی که خروجی '۱' باشد، این نشان می‌دهد که اطلاعات باید ذخیره شوند، در حالی که '۰' نشان می‌دهد که اطلاعات باید دور انداخته شوند. پس از آن، حالا باید برنامه‌ریزی کرد که چه اطلاعاتی باید در سلول‌ها ذخیره شود. این فرآیند از دو بخش تشکیل شده است. ابتدا، لایه دروازه استفاده شده برای ورود، که همچنین یک لایه پوشش سیگموئیدیست، مقادیری را که باید به روز شوند حل می‌کند. ثانوی، یک بردار جدید به نام t توسط یک لایه تانچ، که برای اضافه شدن در این حالت استفاده می‌شود، تولید می‌شود. در پایان، حالا مهم است که برنامه‌ریزی شود که خروجی، که براساس حالت سلول است، تصمیم‌گرفته شود، با این حال، خروجی یک خروجی تصفیه شده خواهد بود. بنابراین، ابتدا، یک لایه سیگموئیدی بخشی از حالت سلول را که باید به عنوان خروجی ارائه شود انتخاب می‌کند. پس از آن، حالت این سلول از طریق tanh گذر داده می‌شود (برای محدود کردن نتایج از -۱ تا ۱) و سپس می‌توان آن را با ضرب آن به همراه نتیجه لایه دروازه سیگموئیدی، برای به دست آوردن خروجی دقیق مورد نظر، افزایش داد.

**3-6- مدل RNN-LSTM برای پيش بينی نياز به بار برق - کاربرد اخير**

**1-3-6- زمينه کار اخير**

در زمینه تکنولوژی خانه‌های هوشمند، تخمین و پیش‌بینی نیاز به بار برق الکتریکی مسأله‌ای بسیار اهم است، اصلیاً به دلیل اینکه شرکت‌ها و انجمن‌های مربوط به برق و الکتریسیته، می‌توانند برنامه‌ریزی و زمان‌بندی موثرتری برای بارها داشته باشند و میزان تولید اضافی انرژی را کاهش دهند. Salah Bouktif و همکاران (2018)، تحقیقات تجربی را در مورد استفاده از مدل الگوریتم LSTM برای پیش‌بینی بار برق با استفاده از انتخاب ویژگی و الگوریتم ژنتیک انجام دادند.

آن‌ها هدف داشتند که یک مدل مبتنی بر LSTM بسازند تا مدل‌های پیش‌بینی برای برنامه‌ریزی و زمان‌بندی بار را طراحی کنند. بسیاری از الگوریتم‌های غیرخطی و خطی آموزش داده شدند تا مناسب‌ترین یکی به عنوان پایه انتخاب شود، با استفاده از پارامترهای متناسب و در آخر استفاده از الگوریتم ژنتیک برای تعیین تاخیر زمانی بهینه و لایه‌هایی که باید توسط شبکه LSTM استفاده شوند. داده‌های مصرف برق شهری فرانسه برای تحقیق و تحلیل استفاده شدند.

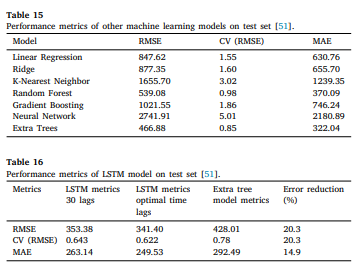
آن‌ها از طریق تحقیقات تجربی خود اثبات کردند که مدل LSTM نتایج بسیار دقیق‌تری ارائه می‌دهد در مقایسه با مدل‌های یادگیری ماشین که با تنظیم پارامترهای فوق‌پارامتر بهینه شده است. نتایج آن‌ها نشان داد که شبکه LSTM با استفاده از ویژگی‌های زمانی محدود شده، تمام ویژگی‌ها ویژگی‌های سری زمانی پیچیده را کسب کرده و با خطای متوسط ریشه مربع و خطای میانگین مطلق کوچک‌تری برای یک فضای شهری بزرگ در مورد پیش‌بینی و پیش‌بینی نشان می‌دهد.

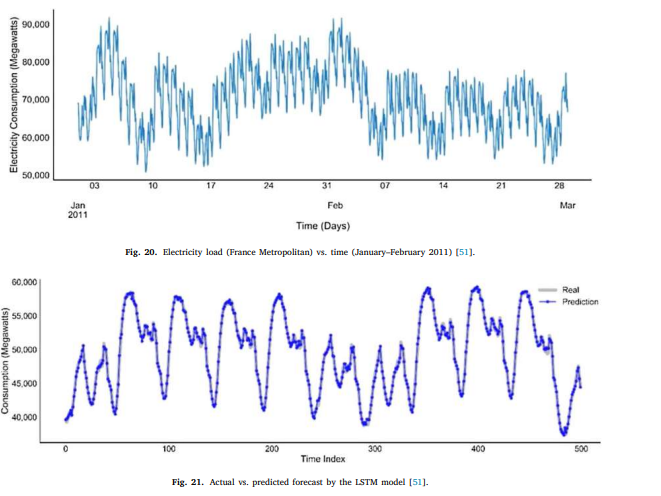
**2-3-6- شرح نتایج**

Salah Bouktif و همکاران (۲۰۱۸)، يک مدل توسعه دادند که از سيستم پوشش دهنده و متنوع، تاخير زمانی منطقی، و لايه های برای مدل LSTM استفاده می کند، و در نهايت، الگوريتم ژنتيک آن ها را قادر به کنترل بيش اندازه گيری می کند و به دست آوردن پيش بينی دقيقتر و قابل اعتماد تر. آن ها مجموعه داده های بزرگی را برای يک فضای شهری که يک بازه زمانی حدود ۹ سال در تعريف ۳۰ دقيقه داشت، جمع آوری کردند، استفاده کردند که با استفاده از آن، يک سيستم شامل RNN-LSTM را آموزش دادند تا ميزان متوسط نياز به بار برق الکتريکی را پيش بينی کنند.

معيارهای ارزيابی که برای تحليل استفاده کردند شامل ضريب واريانس، خطای مربع ميانگين ريشه، و خطای مطلق ميانگين بودند. به عنوان بخشی از تحقيقات تجربی خود، سيستم طراحی شده RNN-LSTM با استاندارد يادگيری ماشين مقايسه شد، و مدل طراحی شده نتايج بهتری را در ميان مدل های غيرخطی و خطی ارائه داد.

نمودارها و نتايج مربوط به الگوريتم LSTM بدين صورت است: نمودار نشان دهنده مصرف برق به مگاوات از ژانويه تا فوريه سال ۲۰۱۱ است. جداول ،۱۵ ۱۶ عملکرد مدل های يادگيری ماشين از جمله مدل LSTM در مجموعه آزمايشی را نشان می دهد، که با ديدن می توان فهمید که مدل LSTM در اين مورد نسبت به ديگر مدل ها برتر است، خطای مربع ميانگين ريشه RMSE و خطای مطلق ميانگين به طور مقايسه ای با مدل LSTM بسيار کوچک تر است. علاوه بر اين، نمودار نشان داده شده در شکل ۲۰ تفاوت بين بار واقعی و پيش بينی شده توسط مدل LSTM را نشان می دهد. تحقيق انجام شده توسط Salah Bouktif و همکاران (۲۰۱۸) نشان می دهد که مدل LSTM نتايج بسيار دقيقی برای پيش بينی بار برق فراهم کرده است.





**4-6- مزايای الگوريتم LSTM**

توانایی پل رفتن از تاخیرهای زمانی بزرگ توسط گسترش خطای ثابت داخل سلول حافظه LSTM ها در مقابل کاهش گرادیان ها مقاوم است. LSTM ها قادر به مدیریت وابستگی‌های موالی طولانی مدت هستند. آن‌ها نیازی به تنظیم دقیق پارامتر ندارند و حافظه‌ای تا زمان بیشتری دارند و در مقیاس پیش‌بینی دقت بالایی ارائه می‌دهند.

**5-6- معايب الگوريتم LSTM**

LSTM ناتوان در حل مشکلات ناپديدشدن گراديان به طور کامل هست زیرا سلول پيچيده تر شده است. LSTM ها زمان و منابع به مقدار زيادی نياز دارند برای آموزش، به عبارت ديگر، نياز به پهنای باند حافظه بسيار بالا دارند. بنابراين، در مورد سخت افزار ناکارامد می باشند. با افزايش تقاضا برای استخراج داده، جستجويی برای مدل های دارای زمان ذخيره سازی بلندتر وجود دارد. شروع وزن بر LSTM ها تصادفی تاثير می گذارد و باعث می شود آن ها شبيه يک شبکه عصبی خوراک به نظر برسند. مشکلات نصب که حتی با الگوريتم قطع شبکه برطرف نمی شوند.

**7- مقایسه کمی الگوریتم ها**

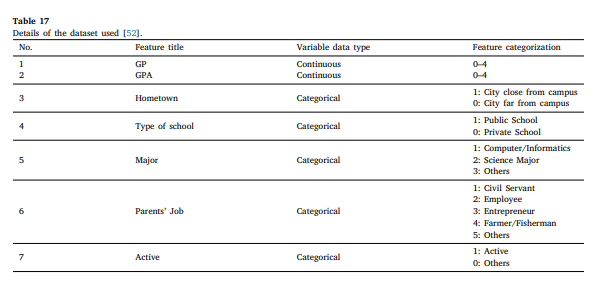
**1-7- مطالعه موردی برای مقايسه الگوريتم های K-NN،**

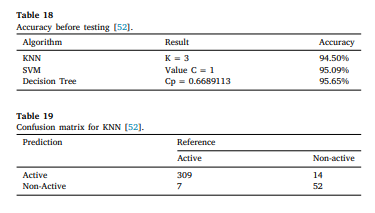
**SVM و درخت تصميم**

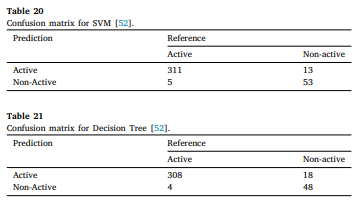
برای انجام تحلیل کمی، مقاله تحقیقی با عنوان “مطالعه مقایسه‌ای الگوریتم‌های KNN، SVM و درخت تصمیم برای پیش‌بینی عملکرد دانش‌آموزان“ نوشته شده توسط Slamet Wiyono و همکاران در سال 2020 مورد بررسی قرار گرفته است. در این مقاله، از یک مجموعه داده در زمان واقعی که شامل ۶ متغیر مختلف بود، برای انجام تحقیق جامع استفاده شده است. تحقیقات آن‌ها ادامه‌ای از کارهای مختلف بوده است که در گذشته برای پیش‌بینی عملکرد دانش‌آموزان بر اساس چندین الگوریتم یادگیری ماشین

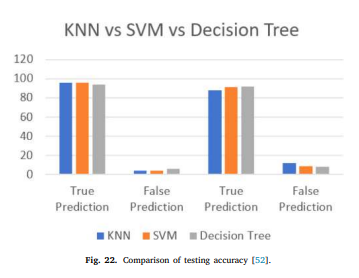
منتشر شده بود. آن‌ها از پلتفورم R Studio برای تحلیل استفاده کرده و نتایج قابل اعتماد و معتبری پس از جمع‌آوری داده‌ها، پیش‌پردازش داده‌ها، ساخت مدل‌های قوی پس از آموزش، اعتبارسنجی و آزمایش مدل‌ها بر اساس الگوریتم‌های مختلف، و در نهایت مقایسه و ارزیابی آن‌ها به نحو کمی ارائه داده‌اند. این مقاله به استفاده از پیش‌پردازش داده برای حذف اشکالات خاص مرتبط با مقادیر نقطه‌داده‌ای گم‌شده یا ویژگی‌های مختلف پرداخته و تحقیقات خود را بر اساس کیفیت و قابلیت اعتماد مدل‌ها تقویت کرده‌اند. در نهایت، سه مدل براساس پارامترهای عملکرد مختلف برای پیش‌بینی عملکرد دانش‌آموزان را مقایسه و ارزیابی نمودند. این مقاله شامل شش متغیر مهم در مجموعه داده بود که به صورت جدول ۱۷ نمایش داده شده است.

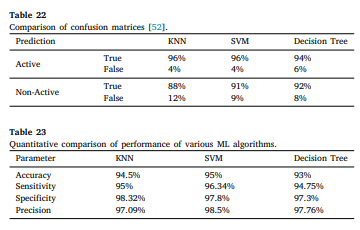
آن‌ها پیش فرض‌های دقیق‌تر مدل را بررسی کرده و مدل‌های خود را بر اساس آن‌ها ارزیابی کرده‌اند. از روش معمول تقسیم داده به داده‌های آموزش و آزمایش استفاده کرده‌اند و در نهایت مدل‌ها را با داده‌های آزمایش ارزیابی و مقایسه کرده‌اند. در نتیجه، مدل‌های آن‌ها با استفاده از رویکرد مشابه با موفقیت اجرا شده و ارزیابی شده‌اند.

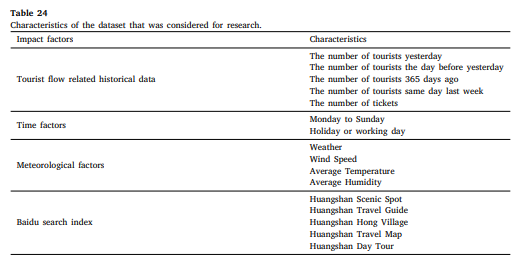




****



****

****

**2-7- مطالعه موردی برای مقايسه الگوريتم ژنتيک و**

**الگوريتم LSTM**

برای مقایسه و تجزیه و تحلیل کمی الگوریتم‌های دوم باقی‌مانده، یعنی الگوریتم ژنتیک و الگوریتم حافظه کوتاه مدت و بلند مدت (LSTM)، مقاله تحقیقی با عنوان “روشی بر اساس ژنتیک، CNN و LSTM برای پیش‌بینی جریان گردشگران روزانه در اماکن دیدنی“ نوشته شده توسط Wenxing Lu و همکاران در سال 2020 مورد بررسی قرار گرفته است.

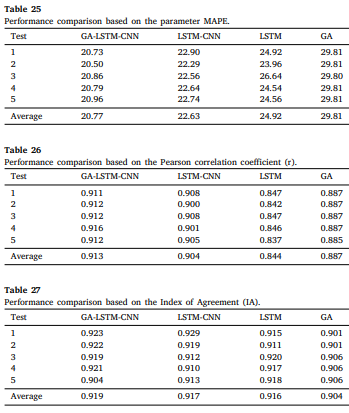
در این مقاله، هدف ایجاد یک مدل برای پیش‌بینی جریان گردشگران در اماکن جذاب و دیدنی بوده تا این اماکن به‌صورت صحیح حفظ و اداره شوند. زیرا هیچ مدلی نمی‌تواند به تنهایی پیش‌بینی دقیقی انجام دهد به دلیل داده‌های بسیار متغیر، نویسندگان این مقاله بر روی یک مدل کار کردند که از شبکه‌های عصبی پیچشی (CNN) همراه با الگوریتم حافظه کوتاه و بلند مدت (LSTM) و در نهایت بهینه‌سازی‌شده توسط الگوریتم ژنتیک برای پیش‌بینی گروه روزانه یک مکان به نام Huangshan در چین استفاده کردند.

به عنوان بخشی از اجرای تحقیقات، آن‌ها در ابتدا نقشه‌های ویژگی پیوسته را از انواع مختلف داده‌ها مانند داده‌های هواشناختی، جستجوی شبکه و غیره تشکیل دادند. در ادامه، استخراج بردار توسط شبکه پیچشی رخ داد و پس از استخراج موفق، بردارهای مشتق شده به شبکه LSTM برای پیش‌بینی داده‌های سری زمانی داده شد.

مجموعه داده قبل از انجام پیش‌بینی، به پیش‌پردازش و نرمال‌سازی می‌گذشت. مدل طراحی شده از نظر عملکردش به صورت کمی مقایسه شده و بدون بهینه‌سازی با الگوریتم ژنتیک و با بهینه‌سازی با الگوریتم ژنتیک با استفاده از پارامترهای عملکرد مشترک، مورد بررسی قرار گرفتند.

بعد از مقایسه منصفانه بین انجام مدل ژنتیک و LSTM GA-LSTM-CNN، LSTM-CNN، که انجام شده توسط نویسندگان، مدل شامل الگوریتم ژنتیک-LSTM-GA(CNN حدود 22.8 جدول 25 مقایسه عملکرد مدل‌های مختلفی را نشان می‌دهد که برای پیش‌بینی جریان گردشگران در روزانه در شهری به نام Huangshan در چین استفاده شده است. همان‌طور که از جدول مشخص است، اگر الگوریتم‌ها برای عملکرد فردی مدنظر قرار گیرند، LSTM حدود 5 جدول 26 نتایجی که برای ضریب همبستگی پیرسون (r) جرد مقدار کوچکی الگوریتم GA نسبت به برتری LSTM ارائه داده شده است.

جدول 27 نتایجی که برای شاخص توافق (IA) جرد عملکرد LSTM به طور روشن از عملکرد الگوریتم ژنتیک فراتر رفته است. بنابراین، براساس نتایج و محاسبه سه پارامتر عملکرد، MAPE r و IA این مطمئن شدن نتیجه می‌دهد که LSTM در عملکرد برتر به الگوریتم ژنتیک برای چنین پیش‌بینی‌های تحلیلی است.



**8- دامنه‌ی آینده**

به دلیل ویژگی‌های انقلابی یادگیری ماشین، دامنه آن روزبه‌روز گسترش می‌یابد. صنعت خودرو یک مثال است که نمایانگر نوآوری‌های عالی به کمک یادگیری ماشین می‌باشد. برندهای معروف اتومبیل‌ها، مانند تسلا، تویوتا، مرسدس بنز، گوگل، نیسان و غیره مقادیر بزرگی پول در این حوزه سرمایه‌گذاری کرده‌اند تا برنامه‌های نوآورانه‌ای با استفاده از یادگیری ماشین و سایر هوش مصنوعی‌ها تدارک ببینند.

خودروی خودران معروفی که توسط تسلا به وجود آمده، با استفاده از سنسورهای اینترنت اشیاء (IoT)، یادگیری ماشین، دوربین‌های با وضوح بالا و غیره ساخته شده است که تنها نیاز به ورود انسان برای برنامه‌ریزی مقصد مورد نظر به سیستم دارد و بقیه کار توسط ماشین انجام می‌شود؛ یعنی انتخاب مسیر مناسب، خالی از ترافیک و تضمین رساندن مسافر به مقصدش به صورت ایمن. رباتیک یک حوزه دیگر است که به طور مداوم در میان دانشمندان، محققان و حتی مردم عادی مورد بحث قرار دارد. یادگیری ماشین و هوش مصنوعی امکان ابداعاتی همچون ربات قابل برنامه‌ریزی اولین بار در سال 1954 با نام Unimate و سپس ایجاد اولین ربات هوش مصنوعی به نام Sophia را امکان‌پذیر کرده است.

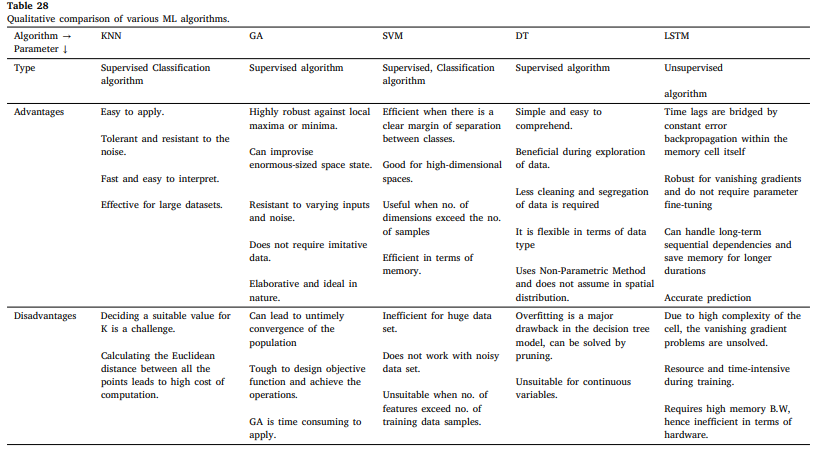
در این حوزه دامنه‌ی روشنی برای تحقیقات وجود دارد و انتظار می‌رود آینده ربات‌های ایجاد شده با استفاده از یادگیری ماشین و هوش مصنوعی و فناوری‌های انقلابی دیگر که قادر به انجام وظایف مشابه انسان‌ها در همه حوزه‌ها شامل پزشکی باشند. یادگیری ماشین هنوز هم باید فراتر از حدود بررسی شود و یکی از حوزه‌هایی که به شدت کمک می‌کند در بررسی یادگیری ماشین، محاسبه کوانتوم است. این محلولیت وقوع مکانیکی شبیه به سوپرپوزیشن و پیچیدگی کوانتوم را تشکیل می‌دهد. به بررسی جزئیات نوآورانه‌ی کاربردهای پنج الگوریتم یادگیری ماشین در مقاله پرداخته شده است.

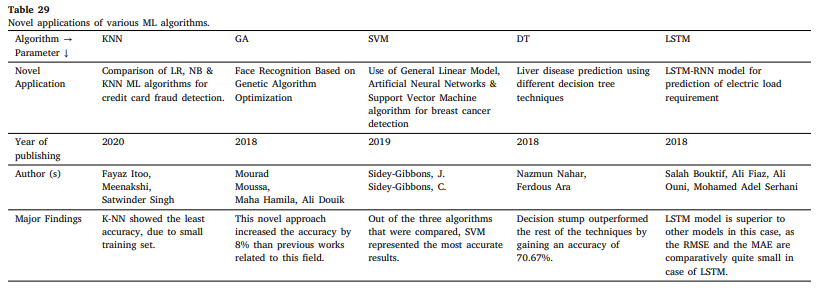
**9- نتیجه گیری**

اين مقاله یک مطالعه مقايسه‌ای از الگوريتم‌های يادگيری ماشین، KNN ژنتيک SVM درخت تصميم و LSTM به همراه برخی از کاربردهای نوآورانه اخيرشان را ارائه می‌دهد که در زمينه تحقيقات آينده دارای دامنه ی عظيمی می باشد. الگوريتم‌ها و مفاهيم مربوطه به جزئيات خیلی زیادی توضیح داده شده اند، از آغاز تا مهمترين کاربردهای جديدشان. اين مقاله نوری بر بسیاری از جنبه‌های حياتی افکار پرتاب می‌کند، مانند زمانی و در چه شرايطی الگوريتم‌ها بر هايجسته نمودند و چگونه در سناریوی امروزی مفيد بودند برای کاربردهای پيش بينی واقعی زمان و کاربردهای ديگر. بينش‌های متعلق به روش‌های پياده سازی این الگوريتم‌ها نيز به طور جزئی مورد بحث قرار گرفته است و نتايج و عملکرد آن ها در کارهای تحقيقی اصيل و نو آور بحث شده است.

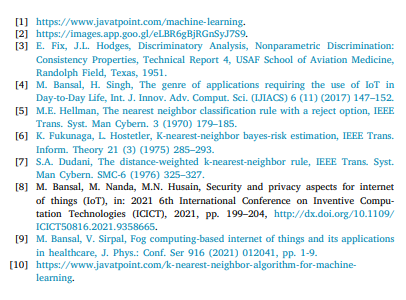
مقايسه‌ی دقيقی از انواع الگوريتم‌های يادگيری ماشين براساس مبنای کيفی و کمی صورت گرفته است و همچنين به صورت جدولی خلاصه شده است. پس از انجام یک بررسی جامع و تحقیق در این حوزه، ما قادر بودیم به نتايج مهمی برسيم که شبکه LSTM و الگوريتم SVM از بهترين نتايج ارائه داده اند زمانی که به پیش‌بینی تحليلی در برنامه‌های واقعی زمانی مرتبط با حوزه‌های چندرشته‌ی مانند پزشکی، تقلب‌های بانکی، شناسايی چهره، پیش‌بینی عملکرد دانش‌آموز، پیش‌بینی مصرف برق و غیره می‌پردازيم.

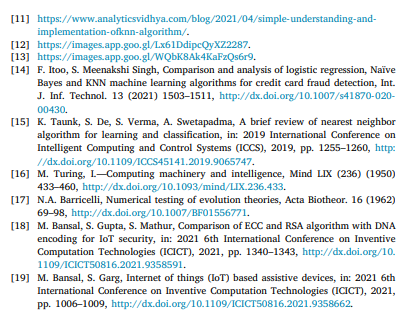
شبکه LSTM شبکه یادگيری عميق با بازخورد است و دارای اين مزيت است که اطلاعات مورد نياز را حفظ می‌کند، که اين امر به آن امکان می‌دهد تا نتايج بسيار دقيقی را ارائه بدهد در زمينه پيش‌بینی. در نهايت، دامنه ی آينده تاکيد می‌کند که تقاضای مورد انتظار و محبوبيت يادگيری ماشين و هوش مصنوعی در آينده، که انتظار می‌رود یا انسان‌ها را در زمينه‌های مختلف پشتیبانی نمايد یا کاملاً آن‌ها را جايگزين کند و تمام روند اتوماتيک کردن را در مقياس و سرعت بزرگتر با کمک تحقيقات پیشرفته و دقيقتری فراهم آورد.





**منابع**



****

