به نام خدا

تحلیل مقایسه‌ای الگوریتم‌های کا نزدیک‌ترین همسایه ، ژنتیک، ماشین بردار پشتیبان، درخت تصمیم و حافظه کوتاه‌مدت بلند در یادگیری ماشین.

کوروش جمال‌پور، امیرمهدی حسینی

دانشگاه دامغان

چکیده

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.**

کلمات کلیدی

یادگیری ماشین،K **نزدیک‌ترین همسایه، الگوریتم ژنتیک، ماشین بردار پشتیبان، درخت تصمیم، الگوریتم حافظه کوتاه-مدت بلند.**

1- مقدمه

یادگیری ماشین، ترکیبی از مفاهیم 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 نزدیک‌ترین همسایه

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 فیکس و جوزف هاجز در سال ۱۹۵۱ برای بررسی جداکننده ارایه 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.

2-1- عملکرد الگوریتم K نزدیک ترین همسایه

حرف 'K' موجود در K-NN به 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 K=۵ 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.

Fig. 4. Classification of new data point based on neighbors [10].

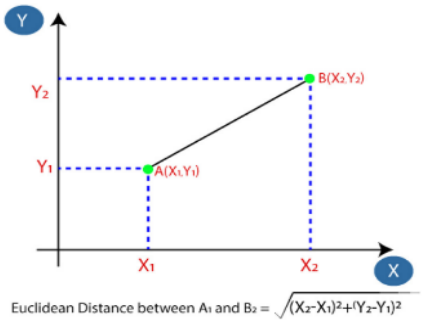


Fig. 3. Calculation of Euclidean Distance b/w two points [13].

2-2- مقایسه ی الگوریتم های یادگیری ماشین رگرسیون لجستیک، بیز ساده و KNN برای تشخیص کلاهبرداری کارت اعتباری - برنامه ی کاربردی اخیر

2-2-1- زمینه ی کار اخیر

کارت‌های اعتباری به دلیل پیشرفت بی‌وقفه فناوری اینترنت امروزه به عنوان یک روش گسترده برای پرداخت‌ها پذیرفته شده‌اند. با این و صف، 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 فیاض 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 فیاض 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- توضیحات و نتایج

جریان رو شی که برای این تحقیق پیروی شده ا ست، در شکل ۵ (a) ن شان داده شده ا ست. تق سیم مجموعه داده به دو بخش، ن سبت‌های ا ستفاده شده برای هر دو، ۵۰:۵۰، ۳۴:۶۶ و ۲۵:۷۵ می‌باشد، جایی که توزیع از داده‌های کلاهبرداری به داده‌های غیر-کلاهبرداری است.

تق سیم‌بندی را می‌توان در شکل ۵ (b) دید. علاوه بی شتر، 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.

Table 5
Results obtained after testing from ratio 50:50 [14].

Techniques	Sensitivity	Specificity	Accuracy	Precision	F-measure	AUC
Logistic regression	0.878	0.949	0.912	0.951	0.913	0.914
Naïve Bayes	0.757	0.964	0.854	0.959	0.846	0.860
K-nearest neighbor	0.687	0.669	0.679	0.701	0.694	0.678

Table 6
Results obtained after testing from ratio 34:66 [14].

Techniques	Sensitivity	Specificity	Accuracy	Precision	F-measure	AUC
Logistic regression	0.777	1.0	0.923	1.0	0.875	0.888
Naïve Bayes	0.718	1.0	0.902	1.0	0.836	0.859
K-nearest neighbor	0.477	0.789	0.681	0.544	0.508	0.633

Table 7
Results obtained after testing from ratio 25:75 [14].

Techniques	Sensitivity	Specificity	Accuracy	Precision	F-measure	AUC
Logistic regression	0.839	0.997	0.959	0.991	0.909	0.918
Naïve Bayes	0.664	0.995	0.915	0.979	0.789	0.829
K-nearest neighbor	0.405	0.861	0.751	0.483	0.441	0.633

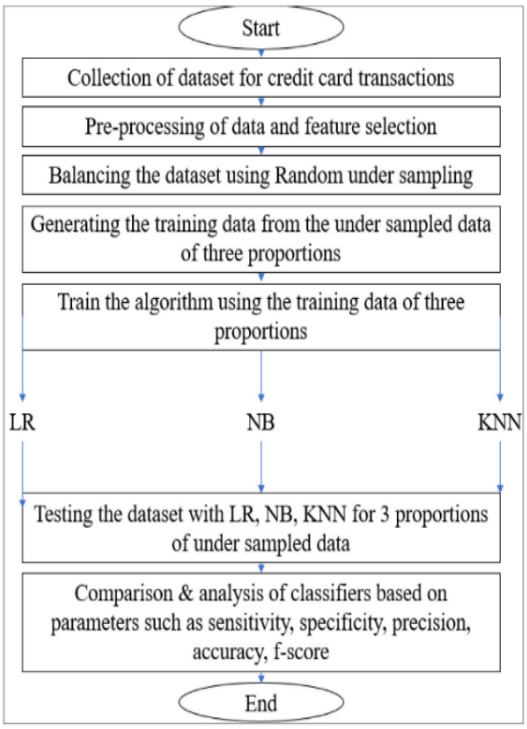


Fig. 5(a). (a) Flow diagram of research work [14].

2-3- مزایای 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.

2-4- 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.

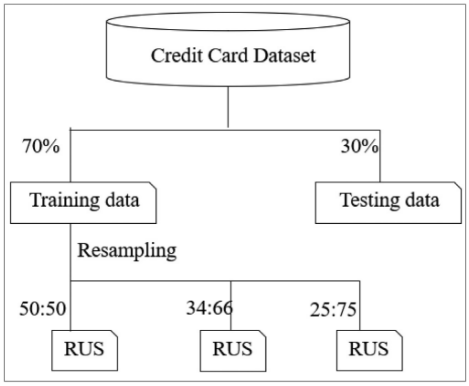


Fig. 5(b). Division of dataset to be used [14].

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

Table 1
Division of dataset by ratio 50:50 [14].

Data division	Training data	Resampling method RUS
Fraud	492	344
Non-fraud	284,315	344
Total	284,807	688

Table 3
Division of dataset by ratio 25:75 [14].

Data division	Training data	Resampling method RUS
Fraud	492	353
Non-fraud	284,315	1024
Total	284,807	1377

Table 2
Division of dataset by ratio 34:66 [14].

Data division	Training data	Resampling method RUS
Fraud	492	341
Non-fraud	284,315	692
Total	284,807	1033

Table 4
Testing dataset preparation [14].

Data proportion	Fraud	Non-Fraud	Total
50:50	35	261	296
34:66	137	306	443
25:75	141	450	591

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.

3-1- فضای جستجو

تمام جمعیت در منطقه خاصی که فضای جستجو نامیده می‌شود محدود شده است. هر موجود حاضر در اینجا دارای یک کلید یا راه حل برای مسئله داده شده است. هر کروموزوم به عنوان یک بردار با طول منظوری رمزگذاری می‌شود. پس از انتخاب و ایجاد نسل اولیه، الگوریتم ژنتیک منجر به تکامل گروه می‌شود، با استفاده از فرایندهای انتخاب، گذردهی و جهش.

3-2- 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.

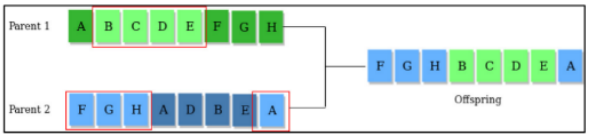
3-3- رمزگذاری

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 ا ست. هر کروموزوم می‌تواند از این فرمت رمزگذاری شود. هر بیت حا ضر در رشته، شامل بخشی از راه‌حل خروجی است.

Chromosome A	**10110010110011100101**
Chromosome B	**11111110000000011111**

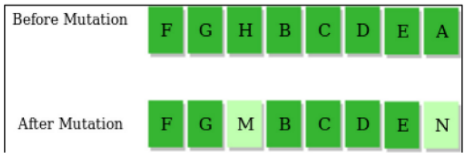
3-4- تلاقی

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.



3-5- جهش

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



3-6- تشخیص چهره بر اساس بهینه سازی الگوریتم ژنتیک- برنامه کاربردی اخیر

3-6-1 - پس زمینه کار اخیر

"مراد موسی و همکاران (۲۰۱۸) ، بر روی روش تشخیص چهره بر اساس تحلیل مولفه اصلی معروف و روش تبدیل کوزین متمرکز انجام دادند. برای

تو سعه یک عملیات ت شخیص چهره پایدار و قابل اطمینان، لازم ا ست ابتدا به انتخاب ویژگی ها توجه کرد، که مسئول لغو سر و صداهای غیر ضروری، داده های اضـافی و ویژگی های متعدد دیگر نامربوط هسـتند. با این حال، توسـعه الگوریتم ژنتیک، که الگوریتمی نسـبتاً جدیدتر برای انتخاب ویژگی ها اسـت، می تواند برای رفع این مسئله استفاده شود. برای استفاده از الگوریتم ژنتیک به منظور حل یک مسـئله، لازم اسـت راه حل های موثر را در زنجیره های بیت قابل اندازه گیر کد کرد تا شـامل کروموزوم های آمده از نقاط خاص شـوند. هدف نهایی اسـتفاده از اپراتورهای ژنتیک و توسـعه تمییز معقولی میان کروموزوم ها ا ست. موراد مو سی و همکاران (۲۰۱۸) ، یک سیستم ت شخیص چهره را با اسـتفاده از الگوریتم ژنتیک در کنار ترکیبی از تحلیل M تبدیل تبدیل کوزین متمرکز – تحلیل مولفه اصـلی PCA-DCT طراحی کردند که برای کاهش بعد و انتخاب ویژگی بر روی یک مجموعه تصـاویر چهره انسـان به کار برده شـد. نتایج ارائه شـده توسـط موراد موسـی و همکاران (۲۰۱۸) ، کارایی این روش را نسبت به کارهای قبلی نشان می دهد."

3-6-2- شرح و نتایج

موراد موسی و همکاران (۲۰۱۸) ، از سه طرح استاندارد به نام موسسه علم و فناوری دانشـگاه منچسـتر، UMIST آزمایشـگاه تحقیقاتی اولیوتی ORL و ییل که در جدول ۸ زیر ارائه شده است، به همراه چهره‌های نماینده (شکل ۹) که برای ت ست ا ستفاده شدند، ا ستفاده کردند. پایگاه داده‌ها به طور ت صادفی برای آموزش یا برای مجموعه‌های تسـت اسـتفاده شـدند و تمام ترتیبات م صلحی برای پژوهش ا ستفاده شدند. نتایج میانگین و م شاهدات ارائه شده ا ست. برای این پژوهش آزمایشی از نسخه R۲۰۱۵MATLAB a ا ستفاده 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 شـده توسـط موراد موسـی و همکاران (۲۰۱۸) در جدول ۱۰ ارائه شـده اند و رویکرد تعقیب شده توسط آنها کمک کرده است تا این سیستم شناسایی چهره به نرخ ت شخیص ۹۹٪ بر سد. آ شکار ا ست 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 ت شخیص چهره شده و در انتخاب مناسب ضرایب مورد نیاز کمک کرده است.

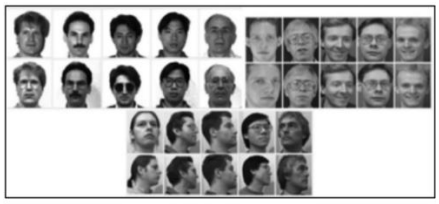


Fig. 9. Sample data [31].

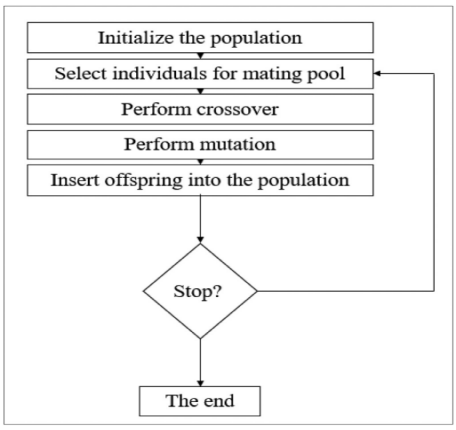


Fig. 10. Entire flow of Genetic Algorithm [30].

3-7- 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.

3-8- معایب الگوریتم ژنتیک

یکی از معایب محتمل الگوریتم ژنتیک این اسـت که اغلب می‌تواند منجر به همگرایی زودرس جمعیت شـود، 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، هایپرمتن و متن توسط الگوریتم ماشین بردار پشـتیبان (SVM) پوشـش داده می‌شـوند. SVM الگوریتمی پیشـرفته اسـت که در تشـخیص متن دسـت‌نوشـته و مرتب‌سـازی پروتئین‌ها در آزمایشگاه‌های زیست شنا سی نیز ا ستفاده می شود. این الگوریتم در حوزه‌های

Table 8
Details of database [31].

Database	Number of classes	Images per class	Size of image
Yale	15	11	243 * 320
ORL	40	10	112 * 92
UMIST	20	24	Variable

Table 9
Parameters for genetic algorithm [31].

Parameters	Values
Population Size	50
Number of generations	100
Crossing probability	0.5
Mutation probability	0.1

Table 10
Results of the research work (Compared with previous results) [31].

Database	Number of classes	Number of train cases	Number of test cases	Recognition rate of previous works	Recognition rate of our works
ORL	40	3	7	92.5%	92.62%
ORL	20	6	4	97.5%	98.45%
UMIST	20	24	Variable	91.66%	99.4%
YALE	15	5	6	93.33%	96.5%
YALE	15	4	7	95.23%	95.5%

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 می شود. 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.

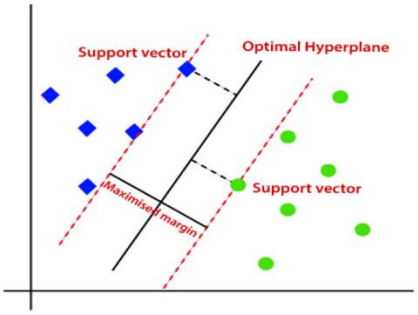


Fig. 12. Implementation of SVM [36].

4-1- انواع بردار پشتیبان

4-1-4- نوع خطی ماشین‌های بردار پشتیبان

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.

4-1-2- نوع غیرخطی ماشین های بردار پشتیبان

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.

4-2- هایپرپلاین و بردار های پشتیبان در الگوریتم SVM

4-2-1- هایپرپلاین

هایپرپلاین در الگوریتم SVM به عنوان بهترین مرز تصمیم تعریف می شود 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.

4-2-2- 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.

4-3- عملکرد الگوریتم SVM

4-3-1- 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 X۱ و X۲ است. یک الگوریتم دسته‌بندی که بتواند جفت مختصات و X۱ و X۲ 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 می شود 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.

4-3-2 SVM غیرخطی

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 می شود. 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.

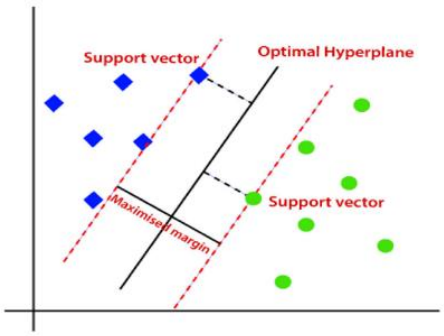


Fig. 12. Implementation of SVM [36].

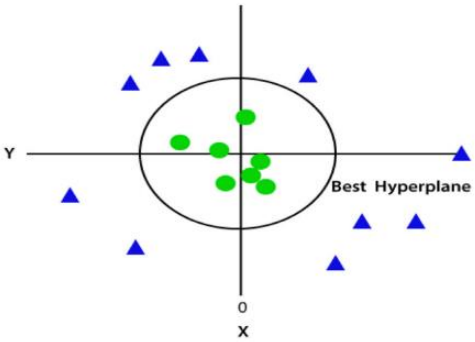


Fig. 13. 3-D Arrangement [36].

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

4-4-1- پیش‌زمینه کاراخیر

جینی ای. ام. ســـایدی-گیبونز و همکاران (2019) تحقیقات دقیقی در مورد الگوریتم‌های خاص یادگیری ما شین که می‌توانند برای پیش‌بینی سرطان، به ویژه سرطان پستان، استفاده شوند، انجام دادند. الگوریتم‌های یادگیری ماشین بـ سیار کارآمد ه ستند و می‌توانند در علوم پز شکی برای ت شخیص زودرس یا پیش‌داوری بسیاری از بیماری‌های فاتال مفید باشند.

در تحقیقات آزمایشـــی آن‌ها، طرح‌های پیش‌بینی مبتنی بر الگوریتم‌های مختلف برای ت شخیص سرطان بر ا ساس موادی که از جرم پستان ا ستخراج شـده، انجام شـد. الگوریتم‌های اسـتفاده شـده در کار تحقیقاتی آن‌ها شـامل شبکه‌های عصبی مصنوعی تک لایه، الگوریتم ماشین بردار پشتیبان با هسته تابع پا یه گرد، و مدل تخطی عمومی (GLM) بود. تقریباً 456 نمو نه از جرم‌های پستان برای ارزیابی و 227 نمونه برای اعتبار سنجی ا ستفاده شدند. قبل از آزمایش الگوریتم‌ها و مدل‌ها در مجموعه اعتبار سنجی برای ت شخیص بیماری، آن‌ها با استفاده از نمونه‌های ارزیابی آموزش دیده شدند.

به منظور مقایســـه عملکردهای مدل‌های الگوریتمی موردنظر، معیارهای ارزیابی که تو سط جینی ای. ام. سایدی-گیبونز و همکاران (2019) ا ستفاده شد، حسا سیت، خصو صیت و دقت بودند. پس از تحقیقات انجام شده تو سط آن‌ها، مشخص شد که الگوریتم SVM بیشینه مساحت زیر منحنی و دقت را نسبت به دو الگوریتم دیگر فراهم کرد.

4-4-2- توضیحات و نتایج

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

ویژگی های جرم های پســـتان که در این تحقیق به آن ها تمرکز داده شده‌است، هسته‌های سلولی هستند که تحت انجام سوزن-نمونه‌برداری دقیق (FNA)، یک روش ت شخیـ صی پز شکی ا ستاندارد در زمینه انکولوژی تجربی شدند. یک شکل نمونه از جرم پستان در شکل 14 ن شان داده شده ا ست. تعداد نمونه‌های ا ستفاده شده برای این تحقیق در واقع تعداد نمونه‌ها ست که به یک شنا سه یکتا (ID) اخت صاص یافته‌اند و به جز آن، ویژگی‌های مختلف دیگر متصل شده‌اند. ستون کلاس نمایش داده شده در شکل حاصل، تشخیص ا ست که بیماری سرطانی بدخیم یا مهلک ا ست که بـ ستگی 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 ن شان داده شده‌است، که هر ویژگی بر اساس یک مقیاس از 1 تا 10 ارزیابی شده است. هرچه مقدار به 10 نزدیک‌تر باشـد، ویژگی طبیعتاً بدخیم اسـت و هرچه مقدار به 1 نزدیک‌تر باشد، ویژگی طبیعتاً مهلک است.

با وجود اینکه تمام الگوریتم‌ها یک ســبک کاری بســیار متنوع دارند، آن‌ها سطحی منا سب از دقت، حـ سا سیت و خ صو صیت را در عملکرد خود ن شان داده‌ا ند و الگوریتم SVM بهترین عملکرد را با دقت 96.0 در جدول 12 نشان داده شده‌است.

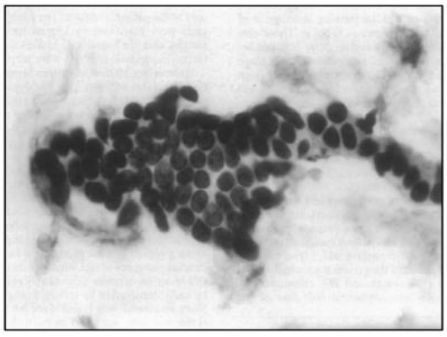


Fig. 14. Sample image of breast mass used for extracting other features [37].

Table 11
Attributes of the data set that was used for the experiment [37].

Instance No.	Sample ID	Thickness	Cell shape	Cell size	Adhesion	Epithelial size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	2
2	1002945	5	4	4	5	7	10	3	2	1	2
3	1015425	3	1	1	1	2	2	3	1	1	2
.											
.											
699	897471	4	8	8	5	4	5	10	4	1	4

Table 12
Evaluated performance metrics of the algorithms on cancer dataset [37].

			Actual outcomes		Sensitivity	Specificity	Accuracy
			Benign (0)	Malignant (1)			
Predicted outcomes	GLM	Benign (0)	148	10	0.99	0.87	0.95
		Malignant (1)	42	67			
	SVM	Benign (0)	1146	5	0.97	0.94	0.96
		Malignant (1)	4	72			
	ANN	Benign (0)	148	11	0.99	0.86	0.94
		Malignant (1)	2	66			

4-5- مزایای الگوریتم SVM

الگوریتم SVM بیشـــترین تطابق را در مواردی ارائه می‌دهد که تقســـیم وا ضحی بین کلاس‌ها وجود دارد. SVM عملکرد بهتری در فضاهای با ابعاد بالاتر نشـان می‌دهد و در مواقعی که تعداد فضـاهای بعدی از مقدار نمونه‌های موجود بی شتر با شد، به طور مؤثر عمل می‌کند. از نظر حافظه، SVM نسبت به سایر ویژگی‌ها کاربردی واقعی است.

4-6- معایب الگوریتم SVM

الگوریتم SVM مناسـب برای مجموعه داده‌های بزرگ نیسـت. در شـرایطی که مجموعه داده‌ها دارای حجم زیادی از سر و صدا هستند، عملکرد منا سبی ندارد که این مســئله خیلی اغلب در عمل رخ می‌دهد. SVM در مواردی که مقدار عددی ویژگی‌های هر نقطه داده بیشتر از نمونه‌های داده آموز شی با شد، عملکرد ضعیفی دارد. الگوریتم SVM برای طبقه‌بندی صورت گرفته تو سط آن توجیه احتمالی نمی‌دهد.

5- درخت تصمیم

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

دو نوع گره وجود دارد: اول، گره تصـمیم که برای تصـمیم‌گیری اسـتفاده می‌شـود و شـاخه‌های 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 ا ست که به شاخه‌های مختلفی تو سعه 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.

5-1- طبقه‌بندی درخت‌های تصمیم

5-1-1- درخت تصمیم دارای متغیر خوشه‌ای

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

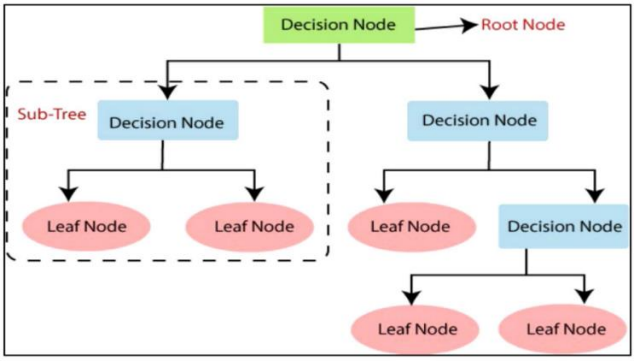


Fig. 15. Schematic diagram of a decision tree [40].

5-1-2- درخت تصمیم دارای متغیر ثابت

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

5-2- اصطلاحات درخت تصمیم

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 تقســیم درختی امکان پذیر نیســت.

تقسیم: شامل فرآیند تقسیم گره ا صلی به زیرگره‌ها بر ا ساس محدودیت‌های ارائه شده می‌شود.

زیردرخت: تقسیم یک سلسله‌مراتب به یک زیردرخت یا شاخه منجر می‌شود.

قالب‌بندی: شـامل حذف شـاخه‌های بیش از حد از درخت تصـمیم به منظور به‌دسـت‌آوردن نتایج بهینه اسـت. در واقع، اندازه درخت را بدون تأثیر بر دقت کاهش می‌دهد. این از دو نوع، قابلیت هزینه و قبول خطا تقسـیم‌بندی اسـت.

گره والدین و کودک: این گره پایه نامیده می‌شـود که همچنین گره والدین نیز نامیده می‌شـود، درحالی که گره‌های باقی‌مانده به سـادگی گره‌های کودک نامیده می‌شوند.

5-3- اندازه گیری انتخاب ویژگی ها

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

5-3-1- بهره اطلاعات

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

5-3-2- شاخص Gini

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

5-4- 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- پیش بینی بیماری کبد با استفاده از تکنیک های مختلف درخت تصمیم-برنامه نویسی اخیر

5-5-1- 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. الگوریتم های 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.

5-5-2- توضیحات و نتایج

هدف اصلی این تحقیق کشف این است که آیا یک بیمار تحت تأثیر بیماری های 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.

5-6- مزایای الگوریتم درخت تصمیم

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.

5-7- 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- الگوریتم 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.

6-1- ساختار LSTM

ساختار LSTM که شامل چهار شبکه عصبی و بلوک های حافظه مختلفی به نام سلول ها است، در زیر تصویب شده است. دروازه ها تغییرات حافظه را بر داده های ذخیره شده در سلول ها انجام می دهند. دروازه ها سه نوع هستند.

6-1-1- 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 بایاس، به بیرون می‌روند. برای دریافت یک خروجی دودویی، نتیجه از آنالیز یک عملکرد فعال‌ســـازی عبور می‌کند. اگر خروجی '۱' با شد، اطلاعات در حالت سلول حفظ می شود و اگر خروجی '۰' باشد، پاک می‌شود.

6-1-2- 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.

6-1-3- 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.

6-2- عملکرد LSTM

مرحله اول نیاز به تصمیم‌گیری در مورد حذف اطلاعات غیرضروری از حالت سلولی دارد. این تصمیم‌گیری‌ها توسط "لایه 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.

6-3- مدل RNN-LSTM برای پیش بینی نیاز به بار برق - کاربرد اخیر

6-3-1- زمینه کار اخیر

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.

6-3-2- شرح نتایج

Bouktif Salah و همکاران (۲۰۱۸)، یک مدل توسعه دادند که از سیستم پوشش دهنده و متنوع، تاخیر زمانی منطقی، و لایه های برای مدل LSTM استفاده می کند، و در نهایت، الگوریتم ژنتیک آن ها را قادر به کنترل بیش اندازه گیری می کند و به دست آوردن پیش بینی دقیقتر و قابل اعتماد تر. آن ها مجموعه داده های بزرگی را برای یک فضای شهری که یک بازه زمانی حدود ۹ سال در تعریف ۳۰ دقیقه داشت، جمع آوری کردند، استفاده کردند که با استفاده از آن، یک سیستم شامل RNN-LSTM را آموزش دادند تا میزان متوسط نیاز به بار برق الکتریکی را پیش بینی کنند.

معیارهای ارزیابی که برای تحلیل استفاده کردند شامل ضریب واریانس، خطای مربع میانگین ریشه، و خطای مطلق میانگین بودند. به عنوان بخشی از تحقیقات تجربی خود، سیستم طراحی شده RNN-LSTM با استاندارد یادگیری ماشین مقایسه شد، و مدل طراحی شده نتایج بهتری را در میان مدل های غیرخطی و خطی ارائه داد.

نمودارها و نتایج مربوط به الگوریتم LSTM بدین صورت است: نمودار نشان دهنده مصرف برق به مگاوات از ژانویه تا فوریه سال ۲۰۱۱ است. جداول ۱۵، ۱۶ عملکرد مدل های یادگیری ماشین از جمله مدل LSTM در مجموعه آزمایشی را نشان می دهد، که با دیدن می توان فهمید که مدل LSTM در این مورد نسبت به دیگر مدل ها برتر است، خطای مربع میانگین ریشه RMSE و خطای مطلق میانگین به طور مقایسه ای با مدل LSTM بسیار کوچک تر است. علاوه بر این، نمودار نشان داده شده در شکل ۲۰ تفاوت بین بار واقعی و پیش بینی شده توسط مدل LSTM را نشان می دهد. تحقیق انجام شده توسط Salah Bouktif و همکاران (۲۰۱۸) نشان می دهد که مدل LSTM نتایج بسیار دقیقی برای پیش بینی بار برق فراهم کرده است.

Table 15
Performance metrics of other machine learning models on test set [51].

Model	RMSE	CV (RMSE)	MAE
Linear Regression	847.62	1.55	630.76
Ridge	877.35	1.60	655.70
K-Nearest Neighbor	1655.70	3.02	1239.35
Random Forest	539.08	0.98	370.09
Gradient Boosting	1021.55	1.86	746.24
Neural Network	2741.91	5.01	2180.89
Extra Trees	466.88	0.85	322.04

Table 16
Performance metrics of LSTM model on test set [51].

Metrics	LSTM metrics 30 lags	LSTM metrics optimal time lags	Extra tree model metrics	Error reduction (%)
RMSE	353.38	341.40	428.01	20.3
CV (RMSE)	0.643	0.622	0.78	20.3
MAE	263.14	249.53	292.49	14.9

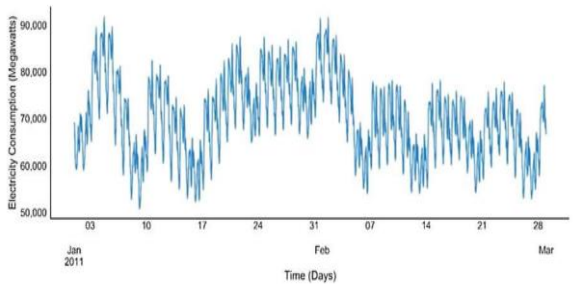


Fig. 20. Electricity load (France Metropolitan) vs. time (January-February 2011) [51].

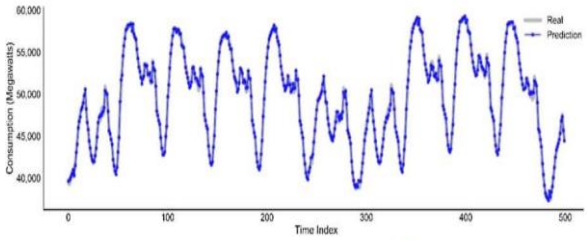


Fig. 21. Actual vs. predicted forecast by the LSTM model [51].

6-4- مزایای الگوریتم 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.

6-5- معایب الگوریتم LSTM

LSTM ناتوان در حل مشکلات ناپدیدشدن گرادیان به طور کامل هست زیرا سلول پیچیده تر شده است. LSTM ها زمان و منابع به مقدار زیادی نیاز دارند برای آموزش، به عبارت دیگر، نیاز به پهنای باند حافظه بسیار بالا دارند. بنابراین، در مورد سخت افزار ناکارامد می باشند. با افزایش تقاضا برای استخراج داده، جستجویی برای مدل های دارای زمان ذخیره سازی بلندتر وجود دارد. شروع وزن بر LSTM ها تصادفی تاثیر می گذارد و باعث می شود آن ها شبیه یک شبکه عصبی خوراک به نظر برسند. مشکلات نصب که حتی با الگوریتم قطع شبکه برطرف نمی شوند.

7- مقایسه کمی الگوریتم ها

7-1- مطالعه موردی برای مقایسه الگوریتم های K-NN، SVM و درخت تصمیم

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.

Table 17
Details of the dataset used [52].

No.	Feature title	Variable data type	Feature categorization
1	GP	Continuous	0-4
2	GPA	Continuous	0-4
3	Hometown	Categorical	1: City close from campus 0: City far from campus
4	Type of school	Categorical	1: Public School 0: Private School
5	Major	Categorical	1: Computer/Informatics 2: Science Major 3: Others
6	Parents' Job	Categorical	1: Civil Servant 2: Employee 3: Entrepreneur 4: Farmer/Fisherman 5: Others
7	Active	Categorical	1: Active 0: Others

Table 18
Accuracy before testing [52].

Algorithm	Result	Accuracy
KNN	K = 3	94.50%
SVM	Value C = 1	95.09%
Decision Tree	Cp = 0.6689113	95.65%

Table 19
Confusion matrix for KNN [52].

Prediction	Reference	
	Active	Non-active
Active	309	14
Non-Active	7	52

Table 20
Confusion matrix for SVM [52].

Prediction	Reference	
	Active	Non-active
Active	311	13
Non-Active	5	53

Table 21
Confusion matrix for Decision Tree [52].

Prediction	Reference	
	Active	Non-active
Active	308	18
Non-Active	4	48

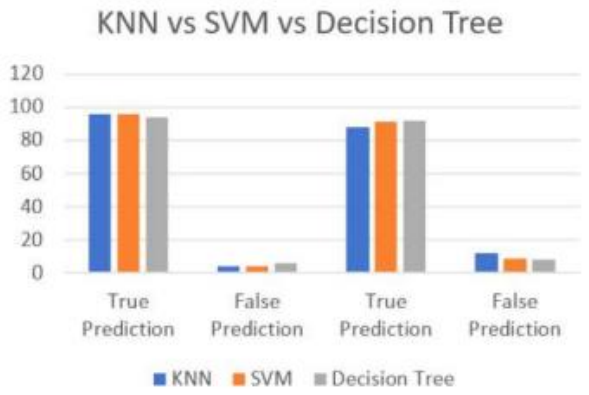


Fig. 22. Comparison of testing accuracy [52].

Table 22
Comparison of confusion matrices [52].

Prediction		KNN	SVM	Decision Tree
Active	True	96%	96%	94%
	False	4%	4%	6%
Non-Active	True	88%	91%	92%
	False	12%	9%	8%

Table 23
Quantitative comparison of performance of various ML algorithms.

Parameter	KNN	SVM	Decision Tree
Accuracy	94.5%	95%	93%
Sensitivity	95%	96.34%	94.75%
Specificity	98.32%	97.8%	97.3%
Precision	97.09%	98.5%	97.76%

Table 24
Characteristics of the dataset that was considered for research.

Impact factors	Characteristics
Tourist flow related historical data	The number of tourists yesterday The number of tourists the day before yesterday The number of tourists 365 days ago The number of tourists same day last week The number of tickets
Time factors	Monday to Sunday Holiday or working day
Meteorological factors	Weather Wind Speed Average Temperature Average Humidity
Baidu search index	Huangshan Scenic Spot Huangshan Travel Guide Huangshan Hong Village Huangshan Travel Map Huangshan Day Tour

7-2- مطالعه موردی برای مقایسه الگوریتم ژنتیک و الگوریتم 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.

Table 25
Performance comparison based on the parameter MAPE.

Test	GA-LSTM-CNN	LSTM-CNN	LSTM	GA
1	20.73	22.90	24.92	29.81
2	20.50	22.29	23.96	29.81
3	20.86	22.56	26.64	29.80
4	20.79	22.64	24.54	29.81
5	20.96	22.74	24.56	29.81
Average	20.77	22.63	24.92	29.81

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

این مقاله یک مطالعه مقایسه‌ای از الگوریتم‌های یادگیری ماشین، KNN ژنتیک SVM درخت تصمیم و LSTM به همراه برخی از کاربردهای نوآورانه اخیرشان را ارائه می‌دهد که در زمینه تحقیقات آینده دارای دامنه ی عظیمی می باشد. الگوریتم‌ها و مفاهیم مربوطه به جزئیات خیلی زیادی توضیح داده شده اند، از آغاز تا مهمترین کاربردهای جدیدشان. این مقاله نوری بر بسیاری از جنبه‌های حیاتی افکار پرتاب می‌کند، مانند زمانی و در چه شرایطی الگوریتم‌ها بر هایجسته نمودند و چگونه در سناریوی امروزی مفید بودند برای کاربردهای پیش بینی واقعی زمان و کاربردهای دیگر. بینش‌های متعلق به روش‌های پیاده سازی این الگوریتم‌ها نیز به طور جزئی مورد بحث قرار گرفته است و نتایج و عملکرد آن ها در کارهای تحقیقی اصیل و نو آور بحث شده است.

مقایسه‌ی دقیقی از انواع الگوریتم‌های یادگیری ماشین براساس مبنای کیفی و کمی صورت گرفته است و همچنین به صورت جدولی خلاصه شده است. پس از انجام یک بررسی جامع و تحقیق در این حوزه، ما قادر بودیم به نتایج مهمی برسیم که شبکه LSTM و الگوریتم SVM از بهترین نتایج ارائه داده اند زمانی که به پیش‌بینی تحلیلی در برنامه‌های واقعی زمانی مرتبط با حوزه‌های

چندرشته‌ی مانند پزشکی، تقلب‌های بانکی، شناسایی چهره، پیش‌بینی عملکرد دانش‌آموز، پیش‌بینی مصرف برق و غیره می‌پردازیم.

شبکه LSTM شبکه یادگیری عمیق با بازخورد است و دارای این مزیت است که اطلاعات مورد نیاز را حفظ می‌کند، که این امر به آن امکان می‌دهد تا نتایج بسیار دقیقی را ارائه بدهد در زمینه پیش‌بینی. در نهایت، دامنه ی آینده تاکید می‌کند که تقاضای مورد انتظار و محبوبیت یادگیری ماشین و هوش مصنوعی در آینده، که انتظار می‌رود یا انسان‌ها را در زمینه‌های مختلف پشتیبانی نماید یا کاملاً آن‌ها را جایگزین کند و تمام روند اتوماتیک کردن را در مقیاس و سرعت بزرگتر با کمک تحقیقات پیشرفته و دقیقتری فراهم آورد.

Table 28
Qualitative comparison of various ML algorithms.

Algorithm → Parameter ↓	KNN	GA	SVM	DT	LSTM
Type	Supervised Classification algorithm	Supervised algorithm	Supervised, Classification algorithm	Supervised algorithm	Unsupervised algorithm
Advantages	Easy to apply. Tolerant and resistant to the noise. Fast and easy to interpret. Effective for large datasets.	Highly robust against local maxima or minima. Can improvise enormous-sized space state. Resistant to varying inputs and noise. Does not require imitative data. Elaborative and ideal in nature.	Efficient when there is a clear margin of separation between classes. Good for high-dimensional spaces. Useful when no. of dimensions exceed the no. of samples Efficient in terms of memory.	Simple and easy to comprehend. Beneficial during exploration of data. Less cleaning and segregation of data is required It is flexible in terms of data type Uses Non-Parametric Method and does not assume in spatial distribution.	Time lags are bridged by constant error backpropagation within the memory cell itself Robust for vanishing gradients and do not require parameter fine-tuning Can handle long-term sequential dependencies and save memory for longer durations Accurate prediction
Disadvantages	Deciding a suitable value for K is a challenge. Calculating the Euclidean distance between all the points leads to high cost of computation.	Can lead to untimely convergence of the population Tough to design objective function and achieve the operations. GA is time consuming to apply.	Inefficient for huge data set. Does not work with noisy data set. Unsuitable when no. of features exceed no. of training data samples.	Overfitting is a major drawback in the decision tree model, can be solved by pruning. Unsuitable for continuous variables.	Due to high complexity of the cell, the vanishing gradient problems are unsolved. Resource and time-intensive during training. Requires high memory R.W, hence inefficient in terms of hardware.

Table 29
Novel applications of various ML algorithms.

Algorithm → Parameter ↓	KNN	GA	SVM	DT	LSTM
Novel Application	Comparison of LR, NB & KNN ML algorithms for credit card fraud detection.	Face Recognition Based on Genetic Algorithm Optimization	Use of General Linear Model, Artificial Neural Networks & Support Vector Machine algorithm for breast cancer detection	Liver disease prediction using different decision tree techniques	LSTM-RNN model for prediction of electric load requirement
Year of publishing	2020	2018	2019	2018	2018
Author (s)	Fayaz Itoo, Meenakshi, Satwinder Singh	Mourad Moussa, Maha Hamila, Ali Douik	Sidey-Gibbons, J. Sidey-Gibbons, C.	Nazmun Nahar, Ferdous Ara	Salah Bouktif, Ali Fiaz, Ali Ouni, Mohamed Adel Serhani
Major Findings	K-NN showed the least accuracy, due to small training set.	This novel approach increased the accuracy by 8% than previous works related to this field.	Out of the three algorithms that were compared, SVM represented the most accurate results.	Decision stump outperformed the rest of the techniques by gaining an accuracy of 70.67%.	LSTM model is superior to other models in this case, as the RMSE and the MAE are comparatively quite small in case of LSTM.

منابع

[1] https://www.javatpoint.com/machine-learning.
[2] https://images.app.goo.gl/eLBR6gBjRGnSyJ7S9.
[3] E. Fix, J.L. Hodges, Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties, Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
[4] M. Bansal, H. Singh, The genre of applications requiring the use of IoT in Day-to-Day Life, Int. J. Innov. Adv. Comput. Sci. (IJIACS) 6 (11) (2017) 147–152.
[5] M.E. Hellman, The nearest neighbor classification rule with a reject option, IEEE Trans. Syst. Man Cybern. 3 (1970) 179–185.
[6] K. Fukunaga, L. Hostetler, K-nearest-neighbor bayes-risk estimation, IEEE Trans. Inform. Theory 21 (3) (1975) 285–293.
[7] S.A. Dudani, The distance-weighted k-nearest-neighbor rule, IEEE Trans. Syst. Man Cybern. SMC-6 (1976) 325–327.
[8] M. Bansal, M. Nanda, M.N. Husain, Security and privacy aspects for internet of things (IoT), in: 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 199–204, http://dx.doi.org/10.1109/ICICT50816.2021.9358665.
[9] M. Bansal, V. Sirpal, Fog computing-based internet of things and its applications in healthcare, J. Phys.: Conf. Ser 916 (2021) 012041, pp. 1-9.
[10] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.
[11] https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-ofknn-algorithm/.
[12] https://images.app.goo.gl/Lx61DdipcQyXZ2287.
[13] https://images.app.goo.gl/WQbK8Ak4KaFzQs6r9.
[14] F. Itoo, S. Meenakshi Singh, Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection, Int. J. Inf. Technol. 13 (2021) 1503–1511, http://dx.doi.org/10.1007/s41870-020-00430.
[15] K. Taunk, S. De, S. Verma, A. Swetapadma, A brief review of nearest neighbor algorithm for learning and classification, in: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255–1260, http://dx.doi.org/10.1109/ICCS45141.2019.9065747.
[16] M. Turing, I.—Computing machinery and intelligence, Mind LIX (236) (1950) 433–460, http://dx.doi.org/10.1093/mind/LIX.236.433.
[17] N.A. Barricelli, Numerical testing of evolution theories, Acta Biotheor. 16 (1962) 69–98, http://dx.doi.org/10.1007/BF01556771.
[18] M. Bansal, S. Gupta, S. Mathur, Comparison of ECC and RSA algorithm with DNA encoding for IoT security, in: 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1340–1343, http://dx.doi.org/10.1109/ICICT50816.2021.9358591.
[19] M. Bansal, S. Garg, Internet of things (IoT) based assistive devices, in: 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1006–1009, http://dx.doi.org/10.1109/ICICT50816.2021.9358662.
[20] A. Bernstein, H. Rubin, Artificial evolution of problem-solvers, Amer. Behav. Sci. 8 (9) (1965) 19–23, http://dx.doi.org/10.1177/000276426500800907.
[21] AS. Fraser, Simulation of genetic systems by automatic digital computers I. Introduction, Aust. J. Biol. Sci. 10 (1957) 484–491, http://dx.doi.org/10.1071/BI9570484.
[22] Alex Fraser, Donald Burnell, Computer Models in Genetics [by] Alex Fraser [and] Donald Burnell, McGraw-Hill, New York, 1970.
[23] Jack L. Crosby, Computer Simulation in Genetics, John Wiley & Sons, London, ISBN: 978-0-471-18880-3, 1973.
[24] David B. Fogel (Ed.), Evolutionary Computation: The Fossil Record, IEEE Press, New York, ISBN: 978-0-7803-3481-6, 1998.

[53] M. Bansal, T. Chopra, S. Biswas, Organ simulation and healthcare services: An application of IoT, in: 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 205–208, http://dx.doi.org/10.1109/ICICT50816.2021.9358677.

[54] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[55] https://www.analyticsvidhya.com/blog/2021/02/machine-learning-101-decision-tree-algorithm-for-classification/.

[56] https://images.app.goo.gl/Ugwp564wYFUqjPsd8.

[57] Nazmun Nahar, Ferdous Ara, Liver disease prediction by using different decision tree techniques, Int. J. Data Min. Knowl. Manage. Process 8 (2018) 01–09, http://dx.doi.org/10.5121/ijdkp.2018.8201.

[58] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[59] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2017) 2222–2232, http://dx.doi.org/10.1109/TNNLS.2016.2582924.

[60] M. Bansal, . Priya, Application layer protocols for internet of healthcare things (IoHT), in: 2020 Fourth International Conference on Inventive Systems and Control (ICISC), 2020, pp. 369–376, http://dx.doi.org/10.1109/ICISC47916.2020.9171092.

[61] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with LSTM, in: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), Vol. 2, 1999, pp. 850–855, http://dx.doi.org/10.1049/cp:19991218.

[62] https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.

[63] B. Lantz, Machine Learning with R, Packt Publishing Ltd, Birmingham, 2013.

[64] M. Ciolacu, A. Tehrani, R. Beer, H. Popp, Education 4.0fostering student's performance with machine learning methods, in: 2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME), 2017, pp. 438–443.

[65] A. Khasanah, A. Harwati, A comparative study to predict students performance using educational data mining techniques,, IOP Conf. Ser.: Mater. Sci. Eng. 215 (2) (2017) 1–7.

[66] A. Vihavainen, Predicting students' performance in an introductory programming course using data from students' own programming process, in: 2013 IEEE 13th International Conference on Advanced Learning Technologies, 2013, pp. 498–499.

[67] M. Quadri, N. Kalyankar, Drop out feature of student data for academic performance using decision tree techniques,, Glob. J. Comput. Sci. Technol. (2010).

[68] T. Devasia, T. Vinushree, V. Hegde, Prediction of students performance using educational data mining, in: 2016 International Conference on Data Mining and Advanced Computing (Sapience), 2016, pp. 91–95.

[69] M. Bansal, N. Adarsh, N. Kumar, M. Meena, 24×7 Smart IoT based integrated home security system, in: 2020 Fourth International Conference on Inventive Systems and Control (ICISC), 2020, pp. 477–481, http://dx.doi.org/10.1109/ICISC47916.2020.9171051.

[70] B. Albreiki, N. Zaki, H. Alashwal, A systematic literature review of student' performance prediction using machine learning techniques, Educ. Sci. 11 (9) (2021) 552, http://dx.doi.org/10.3390/educsci11090552.

[71] W. Lu, H. Rui, C. Liang, L. Jiang, S. Zhao, K. Li, A method based on GA-CNN-LSTM for daily tourist flow prediction at scenic spots, Entropy 22 (2020) 261, http://dx.doi.org/10.3390/e22030261.

[72] Abhinav Jain, et al., Overview and importance of data quality for machine learning tasks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020.

[73] Archana Singh, Rakesh Kumar, Heart disease prediction using machine learning algorithms, in: 2020 International Conference on Electrical and Electronics Engineering (ICE3), IEEE, 2020.

[74] https://intellipaat.com/blog/future-scope-of-machine-learning.

[75] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[25] M. Bansal, P. Priya, Performance comparison of MQTT and CoAP protocols in different simulation environments, in: G. Ranganathan, J. Chen, Á. Rocha (Eds.), Inventive Communication and Computational Technologies, in: Lecture Notes in Networks and Systems, vol. 145, Springer, Singapore, 2021, pp. 549–560, http://dx.doi.org/10.1007/978-981-15-7345-3_47.

[26] M. Bansal, P. Priya, Machine learning perspective in VLSI computer aided design at different abstraction levels, in: S. Shakya, et al. (Eds.), Mobile Computing and Sustainable Informatics, in: Lecture Notes on Data Engineering and Communications Technologies, vol. 68, Springer, Singapore, 2021, pp. 95–112, http://dx.doi.org/10.1007/978-981-16-1866-6_6.

[27] N.A. Barricelli, Numerical testing of evolution theories, Acta Biotheor. 16 (1963) 99–126, http://dx.doi.org/10.1007/BF01556602.

[28] Ingo Rechenberg, Evolutionsstrategie, Holzmann-Froboog, Stuttgart, ISBN: 978-3-7728-0373-4, 1973.

[29] Hans-Paul Schwefel, Numerische Optimierung Von Computer-Modellen (Ph.D. thesis), 1974.

[30] https://www.geeksforgeeks.org/encoding-methods-in-genetic-algorithm/.

[31] Mourad Moussa, Maha Hamila, Ali Douik, A novel face recognition approach based on genetic algorithm optimization, Stud. Inform. Control (ISSN: 1220-1766) 27 (1) (2018) 127–134, http://dx.doi.org/10.24846/v27i1y201813.

[32] Hans-Paul Schwefel, Numerische Optimierung Von Computor-Modellen Mittels Der Evolutionsstrategie : Mit Einer Vergleichenden Einführung in Die Hill-Climbing- Und Zufallsstrategie, Birkhäuser, Basel; Stuttgart, ISBN: 978-3-7643-0876-6, 1977.

[33] Hans-Paul Schwefel, Numerical optimization of computer models, in: (Translation of 1977 Numerische Optimierung Von Computor-Modellen Mittels Der Evolutionsstrategie, Wiley, Chichester; New York, ISBN: 978-0-471-09988-8, 1981.

[34] https://www.geeksforgeeks.org/simple-genetic-algorithm-sga/.

[35] . Lambora, K. Gupta, K. Chopra, Genetic algorithm- A literature review, in: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 380–384.

[36] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[37] J. Sidey-Gibbons, C. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Methodol. 19 (2019) 64, http://dx.doi.org/10.1186/s12874-019-0681-4.

[38] M. Bansal, S. Malik, M. Kumar, N. Meena, Arduino based smart walking cane for visually impaired people, in: 2020 Fourth International Conference on Inventive Systems and Control (ICISC), 2020, pp. 462–465, http://dx.doi.org/10.1109/ICISC47916.2020.9171209.

[39] M. Bansal, . Prince, R. Yadav, P.K. Ujjwal, Palmistry using machine learning and opencv, in: 2020 Fourth International Conference on Inventive Systems and Control (ICISC), 2020, pp. 536–539, http://dx.doi.org/10.1109/ICISC47916.2020.9171158.

[40] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[41] https://www.analyticsvidhya.com/blog/2021/02/machine-learning-101-decision-tree-algorithm-for-classification/.

[42] https://images.app.goo.gl/Ugwp564wYFUqjPsd8.

[43] Nazmun Nahar, Ferdous Ara, Liver disease prediction by using different decision tree techniques, Int. J. Data Min. Knowl. Manage. Process 8 (2018) 01–09, http://dx.doi.org/10.5121/ijdkp.2018.8201.

[44] M. Bansal, A. Goyal, A. Choudhary, Industrial internet of things (iIoT): A vivid perspective, in: V. Suma, J.IZ. Chen, Z. Baig, H. Wang (Eds.), Inventive Systems and Control, in: Lecture Notes in Networks and Systems, vol. 204, Springer, Singapore, 2021, pp. 939–949, http://dx.doi.org/10.1007/978-981-16-1395-1_68.

[45] M. Bansal, N. Oberoi, M. Sameer, IoT In online banking, J. Ubiquit. Comput. Commun. Technol. (UCCT) 2 (4) (2020) 219–222.

[46] Sasan Karamizadeh, Shahidan Abdullah, Mehran Asl, Jafar Shayan, Mohammad Rajabi, Advantage and drawback of support vector machine functionality, in: I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings, 2014, http://dx.doi.org/10.1109/I4CT.2014.6914146.

[47] Ms M.P. Phalak, D.R. Bhandari, . Sharma, Analysis of deci- sion tree-A survey, Int. J. Eng. Res. Technol. (IJERT) 03 (03) (2014).

[48] M. Bansal, V. Sirpal, M.K. Choudhary, Advancing e- government using internet of things, in: S. Shakya, et al. (Eds.), Mobile Computing and Sustainable Informatics, in: Lecture Notes on Data Engineering and Communications Technologies, vol. 68, Springer, Singapore, 2021, pp. 123–137, http://dx.doi.org/10.1007/978-981-16-1866-6_8.

[49] https://images.app.goo.gl/QiZ2cYj6MfrZqCbGA.

[50] http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[51] S. Bouktif, A. Fiaz, A. Ouni, M. Serhani, Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches †, Energies 11 (1636) (2018).

[52] Slamet Wiyono, Taufiq Abidin, Comparative study of machine learning knn, svm, and decision tree algorithm to predict student's performance, Int. J. Res. - Granthaalayah 7 (2019) 190–196, http://dx.doi.org/10.29121/granthaalayah.v7.i1.2019.1048.