

شناسایی عوامل مؤثر بر خطاهای سامانه‌های تشخیص خودکار پلاک خودرو (ANPR) با استفاده از مدل جنگل تصادفی: مطالعه موردی دوربین‌های پلاک‌خوان شهر مشهد

محمدامین ابراهیم زاده^{۱*}، مهدیسه دباغی^۲، نوید نخعی^۲

۱- گروه مهندسی عمران دانشگاه آزاد اسلامی، واحد کرج، ایران.

۲- گروه مهندسی عمران، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

چکیده

دوربین‌های تشخیص خودکار پلاک خودرو (ANPR) به عنوان یکی از اجزای کلیدی سامانه‌های حمل‌ونقل هوشمند، نقش مهمی در پایش جریان ترافیک، ثبت تخلفات و تولید داده‌های مورد نیاز برای مدیریت ترافیک شهری ایفا می‌کنند. با این حال، بروز خطا در ثبت اطلاعات پلاک می‌تواند موجب کاهش کیفیت داده‌های ترافیکی و در نتیجه تصمیم‌گیری‌های نادرست در برنامه‌ریزی و مدیریت حمل‌ونقل شود. هدف این پژوهش، شناسایی و تحلیل عوامل مؤثر بر خطاهای سامانه‌های پلاک‌خوان شهری و ارزیابی توانایی مدل‌های یادگیری ماشین در پیش‌بینی این خطاها است. داده‌های مورد استفاده از دو دوربین پلاک‌خوان منتخب در کلان‌شهر مشهد طی هفته نخست خردادماه ۱۳۹۹ استخراج شد. متغیرهای ورودی شامل شماره دوربین، روز هفته، ساعت ثبت پلاک، سرعت وسیله نقلیه، خط عبوری و حجم ترافیک بود و سه نوع خطا شامل خطای ثبت پلاک، خطای ثبت حرف پلاک و خطای تشخیص زوج یا فرد بودن پلاک به عنوان متغیرهای وابسته مورد بررسی قرار گرفت. به منظور رفع نامتوازنی داده‌ها از الگوریتم SMOTE استفاده شد و عملکرد مدل‌های جنگل تصادفی و شبکه عصبی مصنوعی بر اساس داده‌های آموزشی و آزمون مورد ارزیابی قرار گرفت. نتایج نشان داد که در خطای ثبت پلاک و خطای ثبت حرف پلاک، متغیر ساعت ثبت پلاک بیشترین تأثیر را بر وقوع خطا دارد و مدل جنگل تصادفی به ترتیب با دقت ۸۵ و ۸۶ درصد عملکرد بهتری نسبت به شبکه عصبی مصنوعی ارائه کرده است. همچنین، در خطای تشخیص زوج یا فرد بودن پلاک، متغیر خط عبوری بیشترین اهمیت را داشته و مدل جنگل تصادفی با دقت ۹۶۰۵ درصد بهترین عملکرد را به خود اختصاص داده است. یافته‌های پژوهش نشان می‌دهد که مدل جنگل تصادفی می‌تواند به عنوان ابزاری مؤثر برای شناسایی الگوهای خطا و تحلیل عوامل مؤثر بر عملکرد سامانه‌های ANPR مورد استفاده قرار گیرد و زمینه بهبود کیفیت داده‌های ترافیکی و افزایش قابلیت اطمینان سامانه‌های حمل‌ونقل هوشمند را فراهم سازد. با این حال، به دلیل محدود بودن داده‌ها به دو دوربین و یک بازه زمانی یک‌هفته‌ای، تعمیم نتایج به سایر شرایط مکانی و زمانی مستلزم انجام مطالعات گسترده‌تر است.

کلمات کلیدی: سامانه تشخیص خودکار پلاک خودرو (ANPR)، مدل جنگل تصادفی، سامانه‌های حمل‌ونقل هوشمند،

خطاهای ثبت پلاک، مدیریت ترافیک شهری

* مسئول نویسنده: محمدامین ابراهیم زاده: m.ebrahimzadeh@iau.ir

۱. مقدمه

امروزه با رشد سریع شهرنشینی، افزایش مالکیت خودرو و پیچیده‌تر شدن الگوهای سفر، مدیریت هوشمند ترافیک به یکی از مهم‌ترین چالش‌های کلان‌شهرها تبدیل شده است. در این شرایط، استفاده از سیستم‌های حمل‌ونقل هوشمند به‌عنوان یکی از مهم‌ترین ابزارهای مدیریت شبکه حمل‌ونقل، نقش بسزایی در افزایش ایمنی، کاهش ازدحام، بهبود بهره‌وری شبکه و پشتیبانی از تصمیم‌گیری مدیران شهری ایفا می‌کند. این سامانه‌ها با بهره‌گیری از فناوری‌های نوین نظیر پردازش تصویر، حسگرهای هوشمند، ارتباطات و یادگیری ماشین، امکان جمع‌آوری، پردازش و تحلیل لحظه‌ای داده‌های ترافیکی را فراهم کرده و زمینه توسعه شهرهای هوشمند را مهیا ساخته‌اند.

یکی از مهم‌ترین اجزای سیستم‌های حمل‌ونقل هوشمند، سامانه تشخیص خودکار پلاک خودرو (ANPR) است. این سامانه با ثبت و شناسایی خودکار پلاک وسایل نقلیه، در کاربردهایی نظیر کنترل ترافیک، ثبت تخلفات، مدیریت پارکینگ، اخذ عوارض، شناسایی خودروهای سرقتی و پایش جریان ترافیک مورد استفاده قرار می‌گیرد. با وجود پیشرفت‌های قابل توجه در الگوریتم‌های پردازش تصویر و یادگیری عمیق، عملکرد این سامانه‌ها همچنان تحت تأثیر عواملی نظیر شرایط نوری، کیفیت تصویر، زاویه نصب دوربین، سرعت حرکت وسایل نقلیه، تراکم ترافیک و شرایط محیطی قرار دارد که می‌تواند منجر به بروز خطا در ثبت اطلاعات پلاک و کاهش کیفیت داده‌های ترافیکی شود.

کلان‌شهر مشهد به دلیل جمعیت بالا، حجم قابل توجه سفرهای روزانه و حضور سالانه میلیون‌ها زائر، یکی از پرتراکم‌ترین شهرهای ایران محسوب می‌شود. در چنین شرایطی، صحت عملکرد دوربین‌های پلاک‌خوان نقش مهمی در مدیریت ترافیک، برنامه‌ریزی حمل‌ونقل و تصمیم‌گیری‌های مدیریتی دارد. با این حال، بخش عمده مطالعات پیشین بر توسعه الگوریتم‌های تشخیص پلاک یا بهبود روش‌های پردازش تصویر تمرکز داشته‌اند و بررسی عوامل ترافیکی و عملیاتی مؤثر بر بروز خطاهای سامانه‌های ANPR در شرایط واقعی بهره‌برداری کمتر مورد توجه قرار گرفته است. از سوی دیگر، مطالعات اندکی به‌طور هم‌زمان سه نوع خطای ثبت پلاک، ثبت حرف پلاک و تشخیص زوج یا فرد بودن پلاک را با استفاده از داده‌های واقعی و مدل‌های یادگیری ماشین تحلیل کرده‌اند.

بر این اساس، پژوهش حاضر با بهره‌گیری از داده‌های واقعی ثبت‌شده توسط دوربین‌های پلاک‌خوان شهر مشهد و استفاده از الگوریتم‌های جنگل تصادفی و شبکه عصبی مصنوعی، به شناسایی و رتبه‌بندی عوامل مؤثر بر بروز خطاهای سامانه ANPR می‌پردازد. سؤال اصلی این پژوهش آن است که: «کدام یک از متغیرهای زمانی، ترافیکی و عملیاتی بیشترین تأثیر را بر وقوع خطاهای ثبت پلاک، ثبت حرف پلاک و تشخیص زوج یا فرد بودن پلاک دارند و آیا مدل جنگل تصادفی قادر است این خطاها را با دقت بیشتری نسبت به شبکه عصبی مصنوعی پیش‌بینی کند؟» پاسخ به این سؤال می‌تواند ضمن افزایش دقت سامانه‌های پلاک‌خوان، کیفیت داده‌های مورد استفاده در سامانه‌های حمل‌ونقل هوشمند را بهبود بخشیده و زمینه تصمیم‌گیری دقیق‌تر در مدیریت ترافیک شهری را فراهم سازد.

لازم به ذکر است که هدف این پژوهش توسعه یا بهبود الگوریتم تشخیص پلاک یا پردازش تصویر نیست؛ بلکه تحلیل عوامل عملیاتی و ترافیکی مؤثر بر بروز خطاهای خروجی سامانه ANPR در شرایط واقعی بهره‌برداری است. بنابراین، متغیرهای مورد

استفاده در این پژوهش، متغیرهای مدیریتی و ترافیکی استخراج شده از داده‌های بهره‌برداری سامانه بوده و تحلیل الگوریتم‌های داخلی پردازش تصویر خارج از دامنه این تحقیق است.

۲. مرور ادبیات

مدیریت هوشمند و نظارت بر ترافیک از الزامات اساسی توسعه شهرها محسوب می‌شود. گسترش الگوی شهرنشینی از سال ۱۹۵۰ به بعد و به دنبال آن افزایش شهرنشینی، وضعیت ترافیک در بسیاری از شهرها را چالش برانگیز شده است. براساس گزارش اخیر سازمان ملل متحد، رشد جمعیت و شهرنشینی تا سال ۲۰۵۰ با افزایش ۲٫۹ میلیارد نفری روبه‌رو خواهد بود [۱]. این افزایش جمعیت باعث افزایش تعداد خودروها به ۲٫۵ میلیارد خواهد شد. یکی از کاربردهای فناوری در مدیریت شهری، سیستم‌های هوشمند حمل و نقلی است. سیستم‌های حمل و نقل هوشمند (ITS) با بکارگیری تکنولوژی‌های نوین از قبیل الکترونیک، ارتباطات و سیستم‌های کنترل رایانه‌ای و تلفیق آن‌ها با علوم مهندسی ترافیک و برنامه ریزی، ابزارهای لازم برای کنترل و مدیریت را فراهم ساخته‌اند [۲]. به طور کلی تعریفی از ITS در سال ۱۹۹۸ توسط انجمن حمل و نقل هوشمند آمریکا ارائه شد که به صورت عام مورد قبول قرار گرفت که به بدین شرح است: مردم از تکنولوژی در حمل و نقل برای صرفه جویی در پول و وقت در زندگی روزمره استفاده می‌کنند. تعریف رسمی تری نیز توسط حمل و نقل آمریکا در آوریل سال ۱۹۹۹ نیز منتشر شد که بدین شرح است: سیستم حمل و نقل هوشمند اطلاعات مربوط به جا به جایی مسافر و کالا را جمع‌آوری، نگهداری، پردازش و توزیع می‌نماید. یکی از مهم‌ترین ابزارهای ITS دوربین‌های کنترل ترافیک می‌باشند [۳]. این دوربین‌ها به عنوان یک سیستم نظارتی، کاربردهای متعددی در حمل و نقل دارند که شامل: کنترل ترافیک، ثبت تخلفات، نظارت بر رعایت قوانین راهنمایی و رانندگی از جمله کنترل سرعت، اندازه گیری زمان سفر می‌شود. استفاده از دوربین‌ها سبب افزایش ایمنی در سیستم حمل و نقل درون شهری می‌شوند [۴]. همچنین به کمک این دوربین‌ها می‌توان موضوعاتی از جمله شرایط جوی و گزارش تصادفات را نیز به ثبت رساند. یکی از متداولترین دوربین‌های موجود که در شبکه‌های شهری مورد استفاده قرار می‌گیرد، سامانه خودکار تشخیص پلاک خودرو (ANPR) می‌باشد. این سامانه کاربردهای بسیاری در سامانه‌های نظارت بر ترافیک روزانه و سامانه‌های کنترل عوارض جاده ای دارند. این نوع سیستم‌ها تصاویر خودروها را ضبط کرده و با استفاده از الگوریتم‌های پیشرفته، شماره پلاک خودروها را با دقت بالا ثبت می‌نماید. سپس توسط نرم افزارهای مربوطه، پردازش داده می‌شود. سیستم تشخیص شماره اتوماتیک شماره پلاک ANPR شماره پلاک وسیله نقلیه را مشخص می‌کند. این سیستم در کاربردهای مختلفی از جمله مدیریت ترافیک، مدیریت پارکینگ، جمع‌آوری هزینه پارکینگ، گذرگاه‌های مرزی، یافتن موقعیت سرقت خودروها و غیره کمک می‌کند [۵]. استخراج آمار فراتر از میزان حقیقی آن توسط شمارنده‌های ترافیکی، منجر به تصمیم‌گیری اشتباه برای توسعه راه و تحمیل هزینه‌های زیاد برای گسترش مسیر می‌شود و چنانچه آماری بسیار پایین‌تر از ترافیک واقعی محور ثبت شود، سبب کاهش سطح کیفی ترافیک، کاهش ایمنی، افزایش هزینه‌های تعمیر و نگهداری خواهد شد. بنابر کاربردهای مختلفی که از این دوربین‌ها یاد شده است و نیاز روزافزون به این دوربین‌ها جهت کنترل ترافیک نیاز است که خطاهای این دوربین‌ها شناسایی و سپس میزان هریک برآورد گردد. این امر سبب می‌شود که داده‌های اولیه برای برنامه‌ریزی با دقت بیشتری برداشت گردد تا از خسارات جبران ناپذیر ناشی از برنامه ریزی صحیح جلوگیری شود [۶]. بنابراین پژوهش‌هایی در خصوص بررسی و شناسایی عوامل ایجاد کننده خطا در سامانه خودکار تشخیص پلاک خودرو و نیز ارائه الگوریتم‌ها و روش‌هایی جهت تقلیل خطاهای موجود انجام شده اند. که به طور مثال یکی از دلایل ایجاد خطا مکان نامناسب نصب آن‌ها است. طی پژوهشی یک مدل ریاضی ارائه کرده اند که با در نظر گرفتن امکان ردیابی وسایل نقلیه از طریق تشخیص پلاک و نیز حرکت آنها در طول شبکه راه‌های ایران، بهترین نقاط را برای نصب دوربین‌های پلاک خوان وسایل نقلیه به دست آورد. این مهم از طریق معیارهایی نظیر تعداد دفعات تشخیص در مسیر حرکت نسبت به طول مسیر و همچنین حداکثر تعداد دوربین تشخیص دهنده جسم متحرک در شبکه مورد بررسی قرار گرفته اس [۷]. پژوهش دیگری نیز انجام داده شده است که در آن به ارائه یک مدل بهینه با هدف بیشینه سازی پوشش دوربین‌های کنترل

سرعت در مسیر اقدام شده است. در این مقاله سعی بر آن شده که در یک مسیر حداقل تعداد دوربین ها برآورد گردد تا در حین به صرفه بودن حداکثر پوشش دهی مسیر را داشته باشیم [۸]. در پژوهشی جهت ساماندهی و مدیریت بهتر پارکینگ ها اقدام به کنترل خودروها به کمک سامانه خودکار تشخیص پلاک خودرو کرده اند. به کمک این سیستم توانستند خودروها را احراز کرده و و فضای خالی داخل پارکینگ را شناسایی نمایند. در نتیجه این تحقیق دسترسی غیرقانونی به مناطق محدود، تهدید به سرقت وسایل نقلیه و تجهیزات داخل آنها با این پیشنهاد به شدت کاهش یافت [۹]. از دیگر خطاها می توان به از دست دادن سوژه تحت ردیابی اشاره نمود، که یک مشکل اساسی و شایع در سیستم ANPR است. در پژوهشی با این پرسش که سوژه از کجا و چه موقع از دست رفته، مورد بررسی قرار گرفته است. در این تحقیق از الگوی تانسور که پسوند ماتریس است، برای مدل سازی داده های ترافیکی استفاده شده است. به دلیل آن که داده ها به طور تصادفی از دست می رود، الگوریتم مبتنی بر تانسور پیشنهاد شده است [۱۱]. به کمک این الگوریتم یک مجموعه داده های سیستم ANPR در داده های واقعی کنترل شده و با سایر روش های مدرن نیز مقایسه شده است. بنابراین با توجه به پژوهش های صورت گرفته در ایران و جهان و مشاهده نتایج حاصله و با توجه به آن که مدت زیادی از نصب و به کارگیری سامانه ANPR در کنترل ترافیک مسیرها نمی گذرد، شناخت کامل و جامعی از نقاط قوت و به خصوص نقاط ضعف آن ها در دسترس نیست و در این زمان محدود فرصت کافی برای تحقیق و شناخت حاصل نشده است [۱۲]. به کارگیری روش های هوش مصنوعی و یادگیری ماشین می تواند فرآیند تحلیل داده های پیچیده را به صورت خودکار و با دقت قابل قبول انجام دهد. این رویکردها با یادگیری از داده های برجسته گذاری شده، ضمن کاهش زمان تحلیل، قابلیت شناسایی الگوها و عوامل مؤثر بر بروز خطا را نیز فراهم می کنند. از این رو، استفاده از این مدل ها در سامانه های حمل و نقل هوشمند و تحلیل عملکرد دوربین های پلاک خوان، رویکردی مؤثر برای افزایش کیفیت داده ها و بهبود تصمیم گیری محسوب می شود [۱۳]. عملکرد سامانه های تشخیص خودکار پلاک علاوه بر نوع الگوریتم، به عواملی نظیر کیفیت تصویر، شرایط نوری، زاویه نصب دوربین، تاری ناشی از حرکت، شرایط محیطی و محدودیت های سخت افزاری وابسته است. همچنین، آن ها تأکید کردند که استفاده از روش های نوین یادگیری ماشین و یادگیری عمیق می تواند دقت و قابلیت اطمینان این سامانه ها را به طور قابل توجهی افزایش دهد [۱۴]. سامانه ای مبتنی بر الگوریتم YOLO و شبکه های عصبی کانولوشنی برای تشخیص خودکار پلاک خودرو ارائه کردند. نتایج پژوهش آن ها نشان داد که استفاده از مدل های یادگیری عمیق و رایانش لبه می تواند دقت و سرعت تشخیص پلاک را در کاربردهای بلادرنگ افزایش داده و کارایی سامانه های امنیتی و حمل و نقل هوشمند را بهبود بخشد [۱۵]. تشخیص و شناسایی پلاک خودرو یکی از اجزای کلیدی سیستم های حمل و نقل هوشمند است، اما عواملی نظیر تغییرات روشنایی، سایه، تنوع قالب پلاک و شرایط محیطی همچنان از مهم ترین چالش های این سامانه ها به شمار می روند. همچنین، آن ها تأکید کردند که توسعه روش های نوین تشخیص پلاک نقش مهمی در افزایش دقت سامانه های ANPR و بهبود مدیریت ترافیک هوشمند دارد [۱۶]. سامانه ای مبتنی بر پردازش تصویر و شبکه های عصبی کانولوشنی (CNN) برای تشخیص و شناسایی خودکار پلاک خودرو ارائه کردند. نتایج پژوهش آن ها نشان داد که ترکیب روش های بهبود کیفیت تصویر و یادگیری عمیق می تواند دقت تشخیص نویسه های پلاک را افزایش داده و عملکرد سامانه های ANPR را در کاربردهای مدیریت ترافیک و کنترل وسایل نقلیه بهبود بخشد [۱۷]. سامانه ای یکپارچه مبتنی بر یادگیری عمیق برای تشخیص هم زمان مدل خودرو و پلاک ارائه کردند که با بهره گیری از الگوریتم های YOLO و OCR به دقت ۹۷٫۵ درصد دست یافت. نتایج پژوهش آن ها نشان داد که استفاده از مدل های یادگیری عمیق علاوه بر افزایش دقت تشخیص، عملکرد سامانه های ANPR را در شرایط مختلف محیطی مانند نور کم، باران و مه بهبود می بخشد [۱۸]. عملکرد دوربین های تشخیص خودکار پلاک خودرو را در شمارش وسایل نقلیه در شبکه شهری مشهد بررسی کرده و خطاهای موجود در داده های ترافیکی را شناسایی می کند. همچنین با استفاده از مدل های شبکه عصبی و جنگل تصادفی، مدلی برای کاهش خطا و عملکرد دوربین های تشخیص خودکار پلاک خودرو را در شمارش وسایل نقلیه در شبکه شهری مشهد بررسی کرده و خطاهای موجود در داده های ترافیکی را شناسایی می کند. همچنین با استفاده از مدل های شبکه عصبی و جنگل تصادفی، مدلی برای کاهش خطا و افزایش دقت شمارش ترافیک ارائه می دهد [۲۰].

با وجود پیشرفت‌های قابل توجه در توسعه سامانه‌های تشخیص خودکار پلاک خودرو و بهبود الگوریتم‌های مبتنی بر یادگیری ماشین و یادگیری عمیق، بخش عمده پژوهش‌های پیشین بر افزایش دقت تشخیص پلاک، بهینه‌سازی الگوریتم‌های پردازش تصویر، بهبود کیفیت شناسایی نویسه‌ها و توسعه سامانه‌های هوشمند تشخیص پلاک متمرکز بوده‌اند. در مقابل، مطالعات محدودی به شناسایی و تحلیل عوامل مؤثر بر وقوع خطاهای ثبت اطلاعات در دوربین‌های پلاک‌خوان شهری پرداخته‌اند و سهم نسبی متغیرهای ترافیکی، زمانی و عملیاتی در بروز این خطاها به‌صورت کمی مورد ارزیابی قرار نگرفته است. ازاین‌رو، پژوهش حاضر با بهره‌گیری از داده‌های واقعی دوربین‌های پلاک‌خوان شهری و استفاده از مدل جنگل تصادفی، به شناسایی، رتبه‌بندی و تحلیل عوامل مؤثر بر سه نوع خطای اصلی شامل خطای ثبت پلاک، خطای ثبت حروف پلاک و خطای تشخیص زوج یا فرد بودن پلاک می‌پردازد. نتایج این پژوهش می‌تواند ضمن افزایش قابلیت اطمینان داده‌های حاصل از سامانه‌های ANPR، مبنایی برای بهبود بهره‌برداری، نگهداری و تصمیم‌گیری در سامانه‌های حمل‌ونقل هوشمند فراهم سازد.

۳. روش شناسی

در این پژوهش، به‌منظور شناسایی عوامل مؤثر بر خطاهای سامانه‌های تشخیص خودکار پلاک خودرو، از رویکرد داده‌کاوی و مدل‌سازی طبقه‌بندی استفاده شد. هدف اصلی مدل‌سازی، پیش‌بینی وقوع یا عدم وقوع خطا در داده‌های ثبت‌شده توسط دوربین‌های پلاک‌خوان و تعیین اهمیت نسبی متغیرهای مؤثر بر هر نوع خطا بود. متغیرهای وابسته پژوهش شامل سه نوع خطا، یعنی خطای ثبت پلاک، خطای ثبت حرف پلاک و خطای تشخیص زوج یا فرد بودن پلاک است. متغیرهای مستقل نیز شامل شماره دوربین، روز هفته، ساعت ثبت پلاک، سرعت وسیله نقلیه، خط عبوری و حجم ترافیک در بازه زمانی مورد مطالعه است. فرآیند اجرای پژوهش شامل آماده‌سازی داده‌ها، تعریف متغیرهای خطا، متوازن‌سازی داده‌ها، تقسیم داده‌ها به مجموعه آموزشی و آزمون، آموزش مدل‌ها، ارزیابی عملکرد و تحلیل اهمیت متغیرها بود. در مرحله نخست، داده‌های خام ثبت‌شده توسط دوربین‌های پلاک‌خوان بررسی و متغیرهای مورد نیاز استخراج شد. سپس برای هر یک از سه نوع خطا، متغیر پاسخ به‌صورت دودویی تعریف شد؛ به‌گونه‌ای که مقدار صفر نشان‌دهنده عدم وقوع خطا و مقدار یک نشان‌دهنده وقوع خطا بود. با توجه به نامتوازن بودن توزیع کلاس‌های خطا و عدم خطا، از الگوریتم SMOTE برای متوازن‌سازی داده‌ها استفاده شد. این روش با تولید نمونه‌های مصنوعی برای کلاس اقلیت بر اساس نزدیک‌ترین همسایه‌ها، توزیع داده‌ها را متعادل می‌کند و از سوگیری مدل به سمت کلاس اکثریت جلوگیری می‌نماید. در این پژوهش، تعداد نزدیک‌ترین همسایه‌ها برابر با $k = 5$ و نسبت نمونه‌برداری برابر با ۱۰٪ در نظر گرفته شد؛ به‌گونه‌ای که تعداد نمونه‌های کلاس اقلیت و اکثریت پس از متوازن‌سازی تقریباً برابر شود. پس از آماده‌سازی و متوازن‌سازی داده‌ها، مجموعه داده به دو بخش آموزشی و آزمون تقسیم شد؛ به‌طوری‌که ۷۰ درصد داده‌ها برای آموزش مدل و ۳۰ درصد برای ارزیابی عملکرد آن مورد استفاده قرار گرفت. انتخاب این نسبت به دلیل استفاده گسترده آن در مطالعات طبقه‌بندی و یادگیری ماشین و ایجاد تعادل مناسب میان حجم داده‌های آموزشی و آزمون انجام شد. افزایش حجم داده‌های آموزشی موجب یادگیری بهتر الگوها و روابط میان متغیرها می‌شود، در حالی که اختصاص بخشی مستقل از داده‌ها به آزمون، امکان ارزیابی واقع‌بینانه عملکرد مدل را بر روی داده‌های دیده‌نشده فراهم می‌کند. همچنین، مجموعه داده آزمون در فرآیند آموزش مدل مورد استفاده قرار نگرفت تا از بروز سوگیری در ارزیابی جلوگیری شود. با توجه به حجم نسبتاً زیاد داده‌های این پژوهش و هدف مقایسه عملکرد مدل‌ها بر روی یک مجموعه آزمون مستقل، از روش تقسیم آموزش-آزمون استفاده شد [۲۱]. در این پژوهش، مدل جنگل تصادفی به‌عنوان مدل اصلی طبقه‌بندی مورد استفاده قرار گرفت. جنگل تصادفی یکی از الگوریتم‌های یادگیری ماشین مبتنی بر یادگیری گروهی است که با ایجاد مجموعه‌ای از درختان تصمیم و ترکیب نتایج آن‌ها، توانایی مناسبی در مدل‌سازی روابط غیرخطی، کاهش بیش‌برازش و تحلیل اهمیت متغیرها دارد. در این الگوریتم، هر درخت با استفاده از نمونه‌گیری تصادفی از داده‌های آموزشی ساخته می‌شود و در هر گره، زیرمجموعه‌ای تصادفی از متغیرهای ورودی برای انتخاب بهترین تقسیم مورد بررسی قرار می‌گیرد. در نهایت، طبقه نهایی هر مشاهده بر اساس رأی اکثریت درختان تعیین می‌شود [۲۲].

برای پیاده‌سازی مدل جنگل تصادفی، تعداد درخت‌ها برابر با ۵۰۰ در نظر گرفته شد. معیار تقسیم گره‌ها بر اساس شاخص جینی انتخاب گردید. تعداد متغیرهای بررسی‌شده در هر گره برابر با جذر تعداد متغیرهای ورودی تعیین شد. از آنجا که عدم توازن کلاس‌ها پیش از آموزش مدل با استفاده از SMOTE برطرف شد، وزن‌دهی مجدد به کلاس‌ها اعمال نگردید. همچنین، محدودیتی برای حداکثر عمق درخت‌ها در نظر گرفته نشد و رشد درخت‌ها تا رسیدن به معیار توقف پیش‌فرض الگوریتم ادامه یافت. مقدار Random Seed نیز برای اجرای مدل برابر با ۴۲ تنظیم شد تا نتایج قابلیت بازتولید داشته باشند.

برای مقایسه عملکرد، مدل شبکه عصبی مصنوعی نیز اجرا شد. هدف از به‌کارگیری این مدل، ارزیابی عملکرد مدل جنگل تصادفی در مقایسه با یک روش یادگیری ماشین دیگر بود. عملکرد مدل‌ها بر اساس ماتریس خطا، دقت طبقه‌بندی و منحنی ROC ارزیابی شد. همچنین، در مدل جنگل تصادفی از شاخص کاهش جینی برای تعیین اهمیت نسبی متغیرها استفاده شد. این شاخص نشان می‌دهد حذف یا کاهش نقش هر متغیر تا چه اندازه می‌تواند دقت مدل را تحت تأثیر قرار دهد؛ بنابراین، از آن برای شناسایی مهم‌ترین عوامل مؤثر بر وقوع خطا در سامانه‌های پلاک‌خوان استفاده شد. به‌منظور مقایسه عملکرد مدل جنگل تصادفی، از یک شبکه عصبی مصنوعی پیش‌خور نیز استفاده شد. ساختار شبکه شامل یک لایه ورودی متناسب با تعداد متغیرهای مستقل، یک لایه پنهان و یک لایه خروجی برای طبقه‌بندی دودویی بود. در لایه پنهان از تابع فعال‌سازی سیگموئید و در لایه خروجی نیز از تابع سیگموئید برای پیش‌بینی احتمال وقوع خطا استفاده شد. آموزش شبکه با استفاده از الگوریتم پس‌انتشار خطا انجام گرفت و وزن‌های شبکه تا رسیدن به معیار توقف همگرا به‌روزرسانی شدند. به‌منظور جلوگیری از بیش‌برازش، عملکرد مدل با استفاده از مجموعه داده آزمون ارزیابی و با نتایج مدل جنگل تصادفی مقایسه شد.

۳.۱. تعریف برچسب‌ها و تشکیل مجموعه داده‌ها

به‌منظور آموزش مدل‌های طبقه‌بندی، ابتدا داده‌های خام استخراج‌شده از سامانه تشخیص خودکار پلاک مورد بررسی و پاک‌سازی قرار گرفتند. سپس، برای هر یک از سه متغیر وابسته شامل خطای ثبت پلاک، خطای ثبت حرف پلاک و خطای تشخیص زوج یا فرد بودن پلاک، برچسب‌های طبقه‌بندی به‌صورت دودویی تعریف شد. در این پژوهش، مقدار صفر نشان‌دهنده عدم وقوع خطا و مقدار یک نشان‌دهنده وقوع خطا در نظر گرفته شد. تعیین این برچسب‌ها بر اساس مقایسه اطلاعات ثبت‌شده توسط سامانه ANPR با اطلاعات مرجع و صحت ثبت هر رکورد انجام شد. اطلاعات مرجع مورد استفاده برای تعیین برچسب‌ها، داده‌های تأییدشده سامانه کنترل ترافیک و نتایج بازبینی ثبت هر رکورد توسط سامانه بهره‌بردار بوده است.

پس از تعریف برچسب‌ها، متغیرهای مستقل شامل شماره دوربین، روز هفته، ساعت ثبت، سرعت وسیله نقلیه، حجم ترافیک و خط عبوری استخراج و به همراه متغیر وابسته در قالب یک مجموعه داده یکپارچه سازمان‌دهی شدند. در ادامه، داده‌ها از نظر مقادیر گمشده، داده‌های تکراری و مقادیر غیرمنطقی بررسی و اصلاح شدند تا کیفیت داده‌های ورودی برای آموزش مدل افزایش یابد.

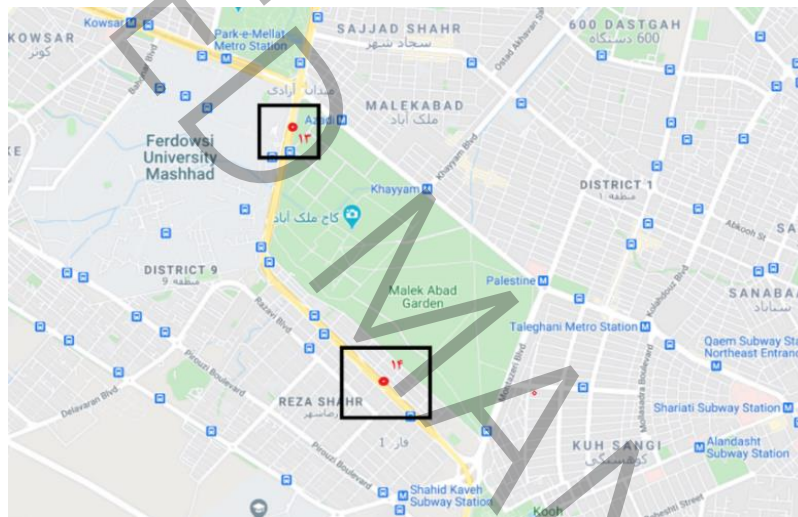
با توجه به نامتوازن بودن توزیع کلاس‌های «خطا» و «عدم خطا»، ابتدا تنها مجموعه داده آموزشی با استفاده از الگوریتم SMOTE متوازن شد تا از ایجاد سوگیری در فرآیند یادگیری جلوگیری شود. سپس، مجموعه داده به روش Hold-out Validation به دو بخش شامل ۷۰ درصد داده‌های آموزشی و ۳۰ درصد داده‌های آزمون تقسیم شد. مجموعه آموزشی برای یادگیری مدل‌ها و مجموعه آزمون که در فرآیند آموزش استفاده نشد، صرفاً برای ارزیابی عملکرد و بررسی قابلیت تعمیم مدل بر روی داده‌های دیده‌نشده مورد استفاده قرار گرفت.

در نهایت، مدل‌های جنگل تصادفی و شبکه عصبی مصنوعی با استفاده از داده‌های آموزشی توسعه یافتند و عملکرد آن‌ها بر اساس داده‌های آزمون با استفاده از شاخص‌هایی نظیر ماتریس درهم‌ریختگی، دقت و سطح زیر منحنی ROC (AUC) مورد ارزیابی قرار گرفت.

۲.۳. معرفی شهر مورد مطالعه

مشهد، مرکز استان خراسان رضوی، با مساحت حدود ۳۵۱ کیلومتر مربع و جمعیتی بالغ بر ۳ میلیون نفر (بر اساس سرشماری سال ۱۳۹۵)، دومین کلان‌شهر پرجمعیت ایران پس از تهران محسوب می‌شود. این شهر به دلیل وجود بارگاه مطهر حضرت امام رضا (ع)، مهم‌ترین قطب زیارتی کشور بوده و سالانه پذیرای میلیون‌ها زائر داخلی و خارجی است. حجم بالای سفرهای درون‌شهری و برون‌شهری، اهمیت مدیریت هوشمند ترافیک و استفاده از سامانه‌های حمل‌ونقل هوشمند را در این کلان‌شهر دوچندان کرده است.

در این پژوهش، به‌منظور بررسی عوامل مؤثر بر خطاهای ثبت اطلاعات در سامانه‌های تشخیص خودکار پلاک، از داده‌های واقعی ثبت‌شده توسط دو دوربین پلاک‌خوان شهری استفاده شد. داده‌های مورد بررسی مربوط به اطلاعات ترافیکی ثبت‌شده توسط دوربین شماره ۱۳ واقع در بزرگراه شهید کلانتری، مقابل درب شرقی دانشگاه فردوسی (جهت جنوب‌شرقی) و دوربین شماره ۱۴ واقع در بزرگراه شهید کلانتری، تقاطع بلوار صدا و سیما (جهت شمال‌غربی) است. اطلاعات این دو دوربین طی هفته نخست خردادماه سال ۱۳۹۹ استخراج و برای تحلیل مورد استفاده قرار گرفت. موقعیت مکانی دوربین‌های مورد مطالعه در شکل (۱) نمایش داده شده است.



شکل ۱: جانیابی دوربین‌های مورد استفاده در این تحقیق خوانش سامانه خودکار تشخیص پلاک خودرو

Figure 1: Placement of cameras used in this study for the automatic license plate recognition system.

۴. مدل‌سازی

پژوهش حاضر که با استفاده از متدولوژی CRISP-DM^۶ انجام گرفته است. با توجه به اینکه هدف اصلی تحقیق حاضر، شناخت خطاهای موجود در عملکرد سامانه ANPR و بررسی تأثیر متغیرهای مستقل نظیر: اطلاعات خودرو، زمان، سرعت خودرو و خط عبوری بر میزان خطای دوربین‌های ANPR می‌باشد، هر موضوعی مرتبط با دوربین‌های ترافیکی ANPR بهره‌مند از نتایج این تحقیق بوده و در دامنه اهداف تجاری-عملیاتی این طرح داده‌کاوی قرار می‌گیرد. شناسایی خطاهای موجود، عوامل مؤثر بر این خطاها، کشف ارتباط بین متغیرهای ممکن و بررسی الگوی رفتاری داده‌های ثبتی دوربین‌های مذکور از طریق انجام فرآیند داده‌کاوی نقشی اساسی در مدیریت این دوربین‌ها و افزایش کارایی آنها خواهد داشت. روزانه حجم عظیمی از اطلاعات و داده‌ها توسط دوربین‌های ترافیکی ثبت می‌شود که این اطلاعات نقشی حیاتی در تصمیم‌گیری‌ها، برنامه‌ریزی‌ها و شناسایی چالش‌های موجود شهری دارد. عدم استفاده از این اطلاعات، کارایی و نقش سازمان بزرگی چون سازمان ترافیک را در برنامه‌ریزی‌های شهری

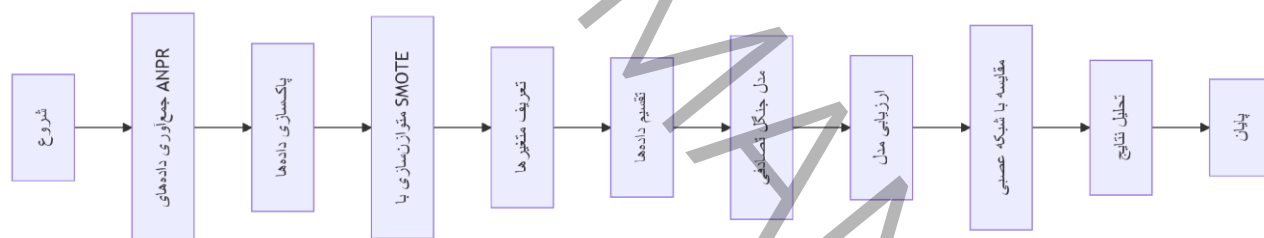
کاهش خواهد داد و سازمان دچار مشکل و بعضاً بحران خواهد شد. این مراحل به ترتیب برای خطای موجود در دوربین ها مورد بررسی قرار گرفته است. که شامل:

- Y₁: خطای ثبت پلاک
- Y₂: خطای ثبت حرف پلاک
- Y₅: زوج یا فرد بودن پلاک

در ادامه نتایج الگوریتم کلاس بندی جنگل تصادفی به تفصیل گزارش گردیده و بر اساس معیارهای ارزیابی بهترین مدل انتخاب می شود. متغیرهای مستقل مورد مطالعه در این مدل ها که انتظار می رود بر خطای دوربین های ترافیکی، به عنوان تابعی از این متغیرها، تأثیرگذار باشند، عبارتند از:

- X₁: شماره دوربین
- X₂: روز هفته
- X₃: ساعت حرکت خودرو
- X₄: سرعت حرکت خودرو،
- X₅: خط عبوری،
- X₆: تراکم یا حجم خودرو،

در تمامی مدل ها نسبت داده های آموزشی و آزمایشی به ترتیب ۰/۷۰ و ۰/۳۰ است. همچنین در انتهای روش تحقیق نمودار روند نما به شرح زیر ارائه شده است.



شکل ۲: نمودار روند نما مراحل اجرای پژوهش

Figure 2: Flowchart of the research implementation stages.

۵. نتایج مدل سازی

۱.۵. بررسی میزان تأثیر زمان و سایر متغیرهای ترافیکی بر خطای ثبت پلاک

مرحله اول) متوازن کردن داده: متغیرهای زمانی شامل روزهای هفته و ساعات مختلف شبانه روز است. در این بخش، تأثیر متغیرهای زمانی و سایر متغیرهای مستقل شامل سرعت حرکت خودرو، حجم ترافیک، شماره دوربین و خط عبوری بر وقوع خطاهای دوربین های پلاک خوان مورد بررسی قرار گرفت. از آنجا که توزیع کلاس های خطا و عدم خطا نامتوازن بود، برای رفع این مشکل از الگوریتم (SMOTE) استفاده شد. این الگوریتم با تولید نمونه های مصنوعی برای کلاس اقلیت بر اساس نزدیک ترین همسایه ها، تعادل مناسبی بین کلاس ها ایجاد می کند. در این پژوهش، تعداد نزدیک ترین همسایه ها برابر با $k=5$ در نظر گرفته شد و نسبت نمونه برداری (Sampling Strategy = 1) به گونه ای تنظیم گردید که تعداد نمونه های کلاس اقلیت و اکثریت پس از متوازن سازی تقریباً برابر شود. همچنین، فرآیند متوازن سازی تنها بر روی داده های آموزشی اعمال شد تا از نشت اطلاعات در مرحله ارزیابی مدل

جلوگیری شود. در نهایت، داده‌های متوازن شده بر مبنای متغیر خطای ثبت پلاک شامل ۲۴۲ رکورد بود که برای آموزش و ارزیابی مدل مورد استفاده قرار گرفت.

۲.۵. بررسی میزان تأثیر زمان و سایر متغیرهای ترافیکی بر خطای ثبت پلاک

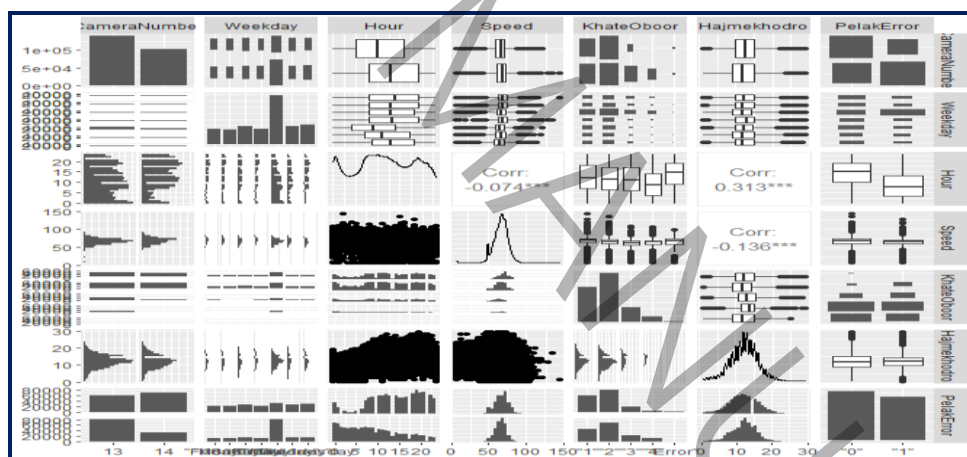
مرحله اول) متوازن کردن داده‌ها: داده‌های متوازنی که بر مبنای متغیر خطای ثبت پلاک و با استفاده از الگوریتم SMOTE ساخته و تولید گردید، شامل ۲۴۲۹۳۸ رکورد می‌باشد.

جدول ۱: توزیع فراوانی خطای ثبت پلاک خودرو

Table 1: Frequency distribution of vehicle license plate recording errors

خطا	کد	تعداد	درصد
ثبت پلاک خودرو	عدم خطا=۰	۱۲۸۶۱۴	۵۳
عدم ثبت پلاک خودرو	خطا=۱	۱۱۴۳۲۴	۴۷
جمع		۲۴۲۹۳۸	۱۰۰

مرحله دوم) تعریف متغیرها: ابتدا در ماتریس نمودار پراکنش^۲ که در شکل ذیل ارائه گردیده است، رابطه بین هر دو متغیر مذکور مورد مطالعه قرار می‌گیرد.



شکل ۳: ماتریس نمودار پراکنش متغیرهای حاضر در مدل‌های کلاس‌بندی مورد استفاده برای متغیر ثبت پلاک

Figure 3: Scatter plot matrix of variables included in the classification models used for the license plate registration variable.

همان‌گونه که در شکل (۳) مشاهده می‌شود، ماتریس نمودار پراکنش نمایی اولیه از نحوه توزیع داده‌ها و روابط میان متغیرهای مستقل مورد استفاده در مدل را ارائه می‌دهد. بررسی این نمودار نشان می‌دهد که الگوی توزیع متغیرها از روابط خطی ساده تبعیت نمی‌کند و در برخی متغیرها پراکندگی قابل توجهی مشاهده می‌شود. همچنین، همپوشانی کلاس‌های خطا و عدم خطا در برخی متغیرها بیانگر آن است که تشخیص الگوهای مؤثر بر وقوع خطا نیازمند استفاده از مدل‌های یادگیری ماشین با توانایی مدل‌سازی روابط غیرخطی،

^۲ . Scatterplot Matrix

نظیر جنگل تصادفی، است. از این رو، در ادامه عملکرد مدل‌های طبقه‌بندی برای شناسایی عوامل مؤثر بر خطای ثبت پلاک مورد ارزیابی قرار گرفته است.

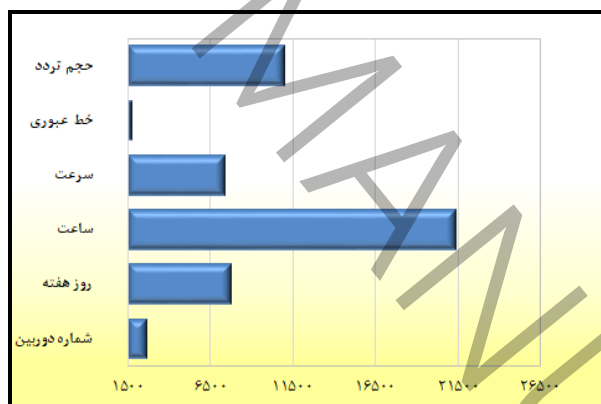
مرحله سوم) ورود اطلاعات به مدل‌ها و تحلیل نتایج: نتایج دو الگوریتم کلاس‌بندی (جنگل تصادفی، شبکه عصبی) مورد بررسی قرار گرفته است.

جنگل تصادفی: جنگل تصادفی با هدف رتبه‌بندی ویژگی‌های مورد بررسی در مدل کلاس‌بندی برای متغیر خطای ثبت پلاک استفاده شده است، که ماتریس انطباقی آن به همراه نمودار ضرایب کاهش جینی در ادامه گزارش می‌گردد.

جدول ۲: ماتریس انطباقی داده‌های آزمایشی جنگل تصادفی برای متغیر خطای ثبت پلاک

Table 2: Random Forest experimental data matching matrix for the license plate registration error variable.

واقعی		ماتریس انطباقی داده‌های آزمایشی	
خطا	عدم خطا		
۲۱۵۲۴	۸۶۷۵۴	خطا	پیش‌بینی شده
۵۸۵۰۳	۳۲۷۶	عدم خطا	
دقت = ۰/۸۵۴ و MSE=۰/۱۵۱			

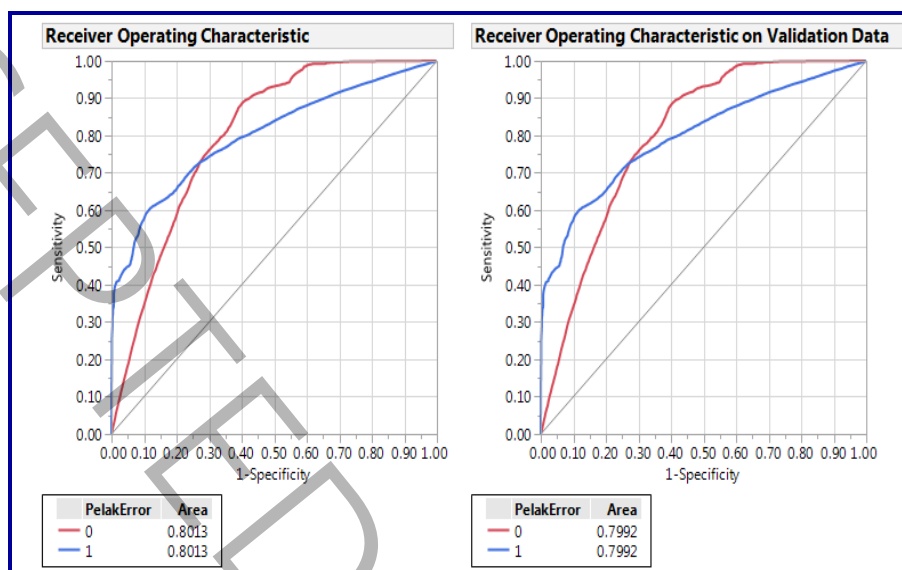


شکل ۴: نمودار میله‌ای میزان اهمیت متغیرهای مدل کلاس‌بندی جنگل تصادفی برای متغیر خطای ثبت پلاک

Figure 4: Bar chart showing the importance of variables in the Random Forest classification model for the license plate registration error variable.

همان‌گونه که در شکل (۴) مشاهده می‌شود، متغیر ساعت بیشترین مقدار شاخص کاهش جینی را به خود اختصاص داده است؛ بنابراین، این متغیر مهم‌ترین عامل مؤثر بر وقوع خطای ثبت پلاک محسوب می‌شود. پس از آن، متغیرهای حجم ترافیک، سرعت وسیله نقلیه و شماره دوربین در رتبه‌های بعدی قرار دارند که نشان می‌دهد شرایط زمانی و ویژگی‌های ترافیکی نقش قابل توجهی در افزایش احتمال وقوع خطاهای ثبت پلاک دارند.

مرحله چهارم) ارزیابی و انتخاب مدل مناسب: با توجه به میزان دقت و خطای مدل‌های مورد بررسی، جنگل تصادفی با دقت ۸۵ درصد بهترین کلاس‌بند برای خطای ثبت پلاک را ارائه می‌دهد.



شکل ۵: منحنی‌های ROC برای داده‌های آموزشی و آزمایشی در مدل جنگل تصادفی برای متغیر خطای ثبت پلاک

Figure 5: ROC curves for training and testing data in the Random Forest model for the license plate recognition error variable.

همان‌گونه که در شکل (۵) مشاهده می‌شود، منحنی‌های ROC برای داده‌های آموزش و اعتبارسنجی عملکرد قابل‌قبولی را در تفکیک کلاس‌های خطا و عدم خطا نشان می‌دهند. مقدار سطح زیر منحنی (AUC) برای داده‌های آموزشی برابر با ۰/۸۰۱۳ و برای داده‌های اعتبارسنجی برابر با ۰/۷۹۹۲ به دست آمده است. نزدیک بودن مقادیر AUC در دو مجموعه آموزشی و اعتبارسنجی بیانگر پایداری عملکرد مدل و تعمیم‌پذیری مناسب آن بر روی داده‌های دیده‌نشده است. همچنین، قرار گرفتن منحنی‌ها در بالای خط مرجع (قطر نمودار) نشان می‌دهد که مدل جنگل تصادفی توانایی مناسبی در تفکیک نمونه‌های دارای خطا و بدون خطا داشته و از قدرت پیش‌بینی قابل‌قبولی برخوردار است.

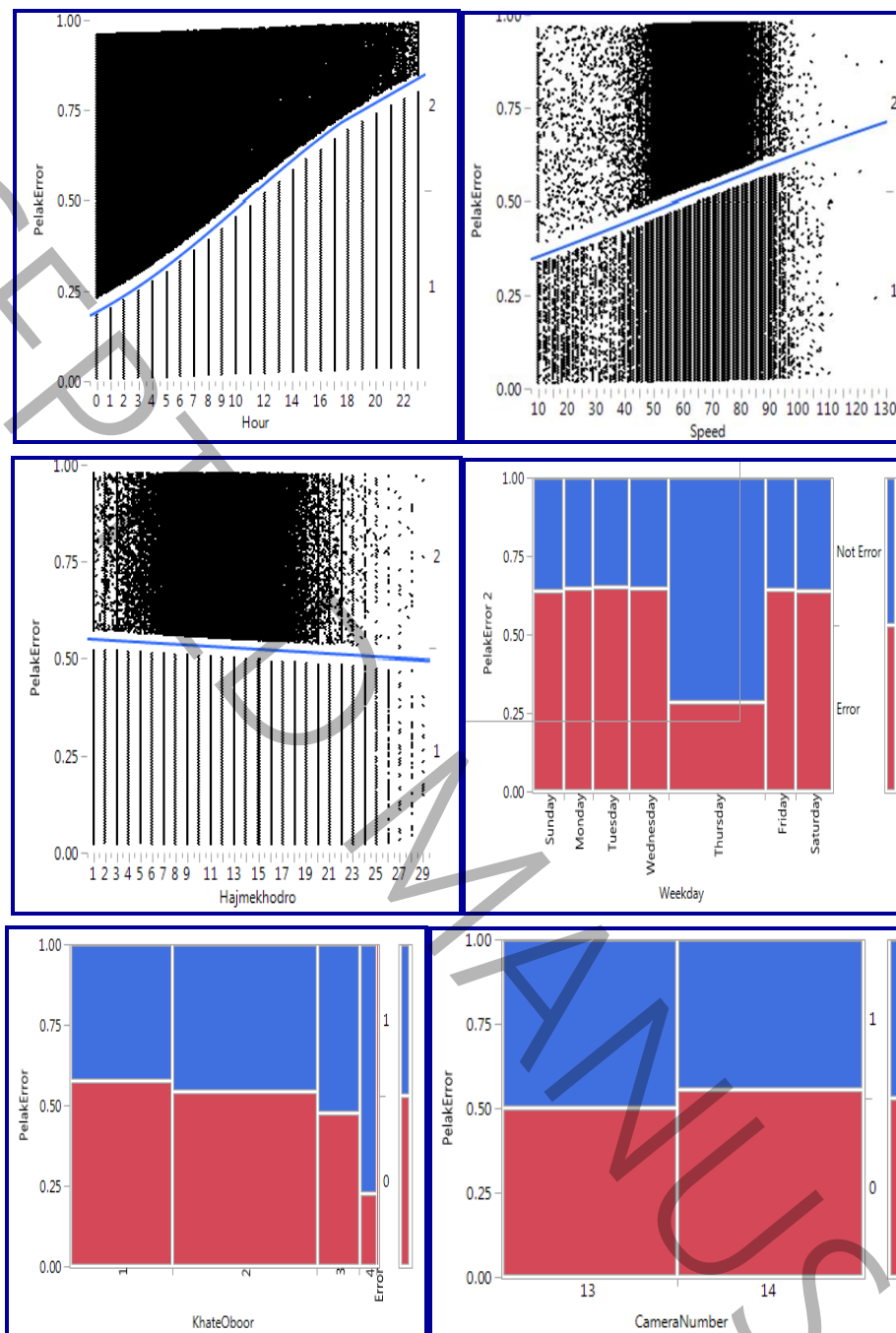
۳.۵. بررسی میزان تأثیر زمان و سایر متغیرهای ترافیکی بر خطای ثبت حرف پلاک

مرحله اول) متوازن کردن متغیرها: داده‌های متوازی که بر مبنای متغیر خطای ثبت حرف پلاک و با استفاده از الگوریتم SMOTE ساخته و تولید گردید، شامل ۵۱۰۱۹۵ رکورد می‌باشد.

جدول ۳: توزیع فراوانی خطای ثبت حرف پلاک خودرو

Table 3: Frequency distribution of errors in recording vehicle license plate characters

خطا	کد	تعداد	درصد
ثبت حرف پلاک خودرو	عدم خطا=۰	۲۷۰۱۰۳	۵۳
عدم ثبت حرف پلاک خودرو	خطا=۱	۲۴۰۰۹۲	۴۷
جمع		۵۱۰۱۹۵	۱۰۰



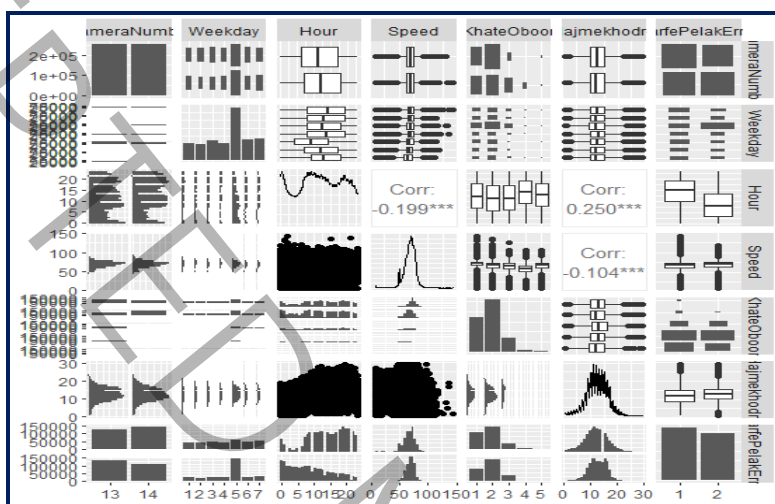
شکل ۶: نمودار پراکنش رابطه بین خطا در ثبت پلاک و متغیرهای تحقیق

Figure 6: Scatter plot of the relationship between license plate recognition error and research variables.

همان گونه که در شکل (۶) مشاهده می شود، نمودارهای پراکنش نحوه توزیع مشاهدات و ارتباط اولیه میان متغیرهای مستقل با متغیر خطای ثبت پلاک را نشان می دهند. بررسی این نمودارها بیانگر آن است که الگوی وقوع خطا از یک رابطه خطی ساده تبعیت نمی کند و میزان پراکندگی داده ها در متغیرهای مختلف متفاوت است. همچنین، تمرکز و پراکندگی نقاط در برخی بازه های زمانی و ترافیکی نشان می دهد که متغیرهایی نظیر ساعت ثبت، حجم ترافیک، سرعت وسیله نقلیه و شماره دوربین می توانند در بروز خطای ثبت پلاک

نقش مؤثری داشته باشند. این موضوع ضرورت استفاده از مدل‌های یادگیری ماشین با قابلیت مدل‌سازی روابط غیرخطی، از جمله جنگل تصادفی، را برای شناسایی الگوهای مؤثر بر وقوع خطا تأیید می‌کند.

مرحله دوم) تعریف متغیرها: در این بخش ابتدا در ماتریس نمودار پراکنش^۳ که در شکل ذیل ارائه گردیده است، رابطه بین هر دو متغیر مذکور مورد مطالعه قرار می‌گیرد. شایان ذکر است که در تمامی مدل‌های ارائه شده در ادامه میزان دقت و خطا در داده‌های آموزشی و داده‌های آزمایشی نزدیک هم هستند، لذا در ادامه تنها موارد مربوط به داده‌های آزمایشی را گزارش می‌کنیم.



شکل ۷: ماتریس نمودار پراکنش متغیرهای حاضر در مدل‌های کلاس‌بندی مورد استفاده برای متغیر خطای ثبت حرف پلاک

Figure 7: Scatter plot matrix of variables included in the classification models used for the license plate character recognition error variable.

همان‌گونه که در شکل (۷) مشاهده می‌شود، ماتریس نمودارهای پراکنش، توزیع هر یک از متغیرها و روابط دوبه‌دو میان متغیرهای مستقل و متغیر وابسته را نمایش می‌دهد. نمودارهای قطر اصلی، نحوه توزیع فراوانی متغیرها را نشان می‌دهند، در حالی که نمودارهای خارج از قطر بیانگر الگوی پراکندگی و میزان وابستگی بین متغیرها هستند. همچنین، ضرایب همبستگی محاسبه‌شده نشان می‌دهد که بین برخی متغیرها همبستگی خطی ضعیفی وجود دارد؛ به‌عنوان نمونه، بین متغیر سرعت و ساعت همبستگی منفی ضعیفی حدود ۰/۲۰ و بین حجم ترافیک و ساعت همبستگی مثبت ضعیفی حدود ۰/۲۵ مشاهده می‌شود. این نتایج بیانگر آن است که روابط میان متغیرها عمدتاً غیرخطی بوده و هیچ‌گونه هم‌خطی شدید بین متغیرهای ورودی وجود ندارد. از این‌رو، استفاده از مدل‌های یادگیری ماشین مبتنی بر روابط غیرخطی، مانند جنگل تصادفی، برای تحلیل عوامل مؤثر بر خطاهای ثبت پلاک، انتخابی مناسب و قابل اتکا محسوب می‌شود.

مرحله سوم) ورود اطلاعات به مدل‌ها و تحلیل نتایج: نتایج چهار الگوریتم کلاس‌بندی جنگل تصادفی مورد بررسی قرار گرفته است. برای انجام جنگل تصادفی با هدف رتبه‌بندی ویژگی‌های تأثیرگذار بر متغیر خطای ثبت حرف پلاک، به دلیل ضعف سخت‌افزاری مجبور به کاهش حجم داده‌ها شدیم. از داده‌ها ابتدا نمونه‌ای ۷۰ درصدی گرفتیم. که ماتریس انطباقی آن به‌همراه نمودار ضرایب

^۳. Scatterplot Matrix

اهمیت کاهش جینی^۴ در ادامه گزارش می‌گردد. ضریب کاهش جینی یک متغیر میزان اهمیت آن متغیر را در مدل نشان می‌دهد، به‌طوری که بیان می‌دارد با حذف آن متغیر چه میزان از دقت مدل کم می‌شود.

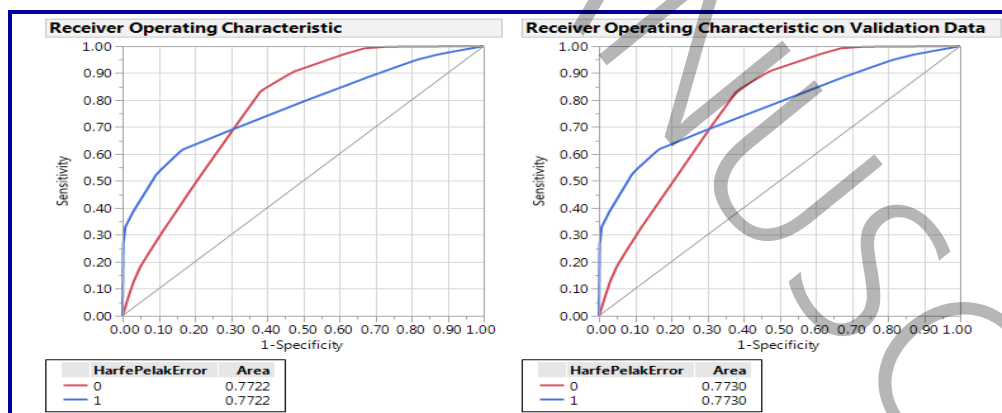
جدول ۴- ماتریس انطباقی داده‌های آزمایشی جنگل تصادفی برای خطای ثبت حرف پلاک

Table 4 – Random Forest experimental data matching matrix for license plate character recognition error.

واقعی		ماتریس انطباقی داده‌های آزمایشی	
خطا	عدم خطا	خطا	پیش‌بینی شده
۸۷۵۷۳	۴۰۳۰		
۳۰۰۷۲	۱۲۸۳۲۰	عدم خطا	
دقت = ۰/۸۶۳ و MSE = ۰/۱۴۰			

با توجه به جدول فوق، میزان دقت و خطای کلاس‌بند طراحی شده توسط الگوریتم جنگل تصادفی برای متغیر خطای ثبت حرف پلاک به ترتیب ۸۶ درصد و ۱۴ درصد می‌باشد، که مقادیر معقولی هستند و نشان از مناسبت مدل دارند.

مرحله چهارم) ارزیابی و انتخاب مدل مناسب: یکی از معیارهای ارزیابی مناسبت مدل برابری خطا و دقت در داده‌های آموزشی و آزمایشی می‌باشد، که این برابری برای تمامی مدل‌های کلاس‌بندی اجرا شده برای خطای ثبت حرف پلاک برقرار بوده است. همچنین، با توجه به میزان دقت و خطای مدل‌های مورد بررسی، جنگل تصادفی با دقت ۸۶ درصد بهترین کلاس‌بند برای خطای ثبت حرف پلاک را ارائه می‌دهد. همان‌طور که شکل (۹) نشان می‌دهد، مساحت زیر منحنی‌ها بیش از ۷۰ درصد است که نشان می‌دهد مدل جنگل تصادفی ساخته شده بسیار مناسب و دقیق است.

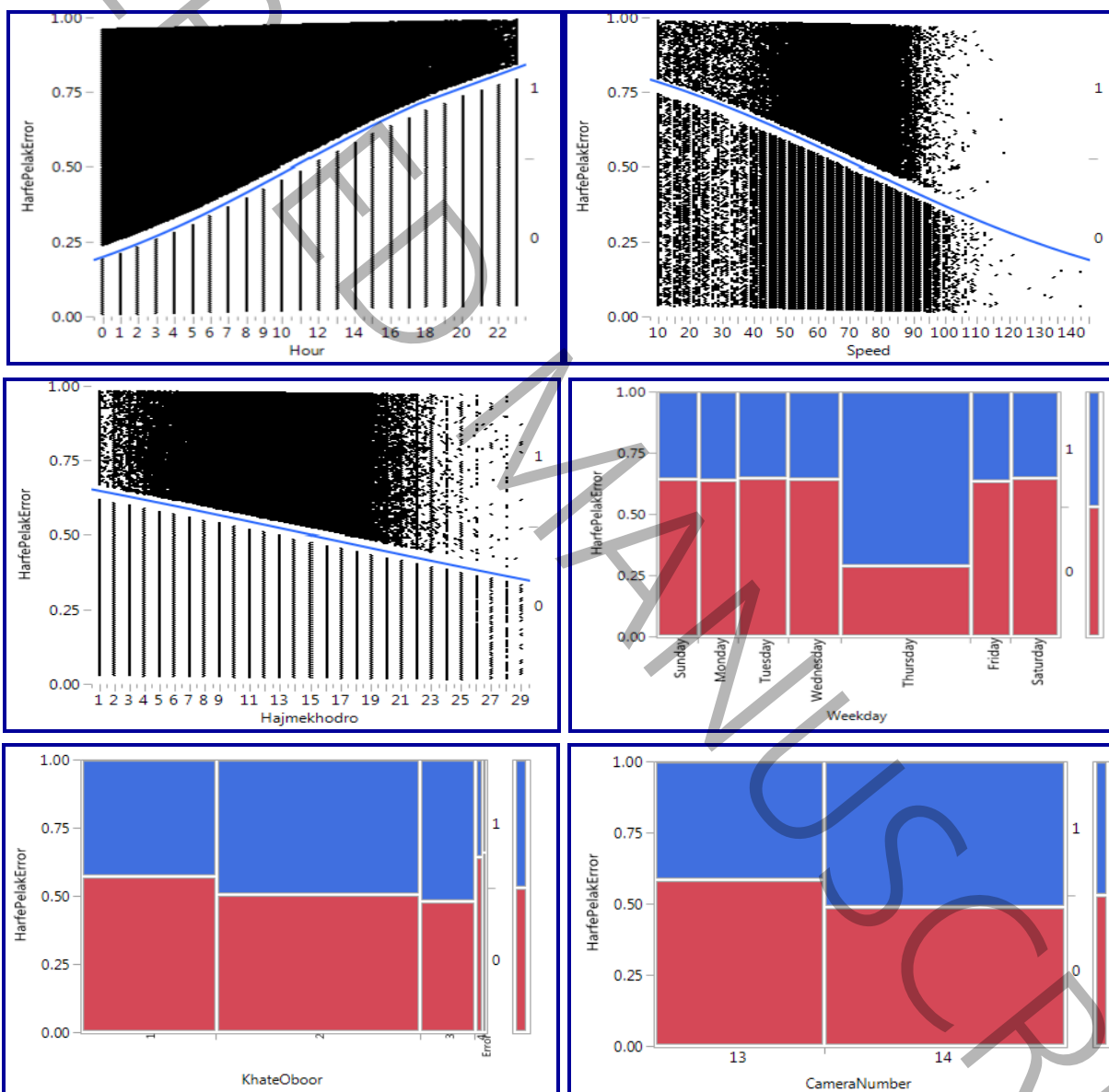


شکل ۹: منحنی‌های ROC برای داده‌های آموزشی و آزمایشی در مدل جنگل تصادفی برای خطای ثبت حرف پلاک

Figure 9: ROC curves for training and testing data in the Random Forest model for license plate character recognition error.

⁴ . MeanDecreaseGini

همان گونه که در شکل (۹) مشاهده می‌شود، منحنی‌های ROC مربوط به داده‌های آموزش و اعتبارسنجی، عملکرد مدل جنگل تصادفی را در تشخیص خطای ثبت حروف پلاک نشان می‌دهند. مقدار سطح زیر منحنی (AUC) برای داده‌های آموزشی برابر با ۰/۷۷۲ و برای داده‌های اعتبارسنجی برابر با ۰/۷۷۳ به دست آمده است. نزدیکی این دو مقدار بیانگر پایداری عملکرد مدل و توانایی مناسب آن در تعمیم به داده‌های جدید است. همچنین، قرار گرفتن منحنی‌های ROC بالاتر از خط مرجع نشان می‌دهد که مدل از قدرت قابل قبولی در تفکیک نمونه‌های دارای خطا و بدون خطا برخوردار است. اگرچه مقدار AUC در مقایسه با مدل مربوط به خطای ثبت پلاک اندکی کمتر است، اما همچنان عملکرد مدل برای پیش‌بینی خطای ثبت حروف پلاک در محدوده قابل قبول قرار دارد و می‌تواند به عنوان ابزاری مناسب برای تحلیل این نوع خطا مورد استفاده قرار گیرد.



شکل ۱۰: نمودار پراکنش رابطه بین خطا در ثبت حروف پلاک و متغیرهای تحقیق

Figure 10: Scatter plot of the relationship between the error in recording the license plate letter and the research variables.

همان گونه که در شکل (۱۰) مشاهده می شود، نمودارهای پراکنش، نحوه توزیع داده ها و ارتباط اولیه بین متغیرهای تحقیق و متغیر خطای ثبت حرف پلاک را نمایش می دهند. بررسی این نمودارها نشان می دهد که الگوی وقوع خطا در ثبت حروف پلاک از یک رابطه خطی مشخص پیروی نمی کند و میزان پراکندگی داده ها در متغیرهای مختلف متفاوت است. همچنین، تمرکز مشاهدات در برخی بازه های زمانی و تغییرات حجم ترافیک، سرعت وسیله نقلیه، شماره دوربین و خط عبوری بیانگر آن است که این متغیرها می توانند بر احتمال وقوع خطای ثبت حرف پلاک تأثیرگذار باشند. علاوه بر این، همپوشانی نسبی مشاهدات مربوط به کلاس های «خطا» و «عدم خطا» نشان می دهد که تفکیک این دو کلاس با روش های آماری خطی به سادگی امکان پذیر نیست و استفاده از مدل های یادگیری ماشین با قابلیت مدل سازی روابط غیرخطی، مانند جنگل تصادفی، برای شناسایی الگوهای مؤثر بر این نوع خطا ضروری است.

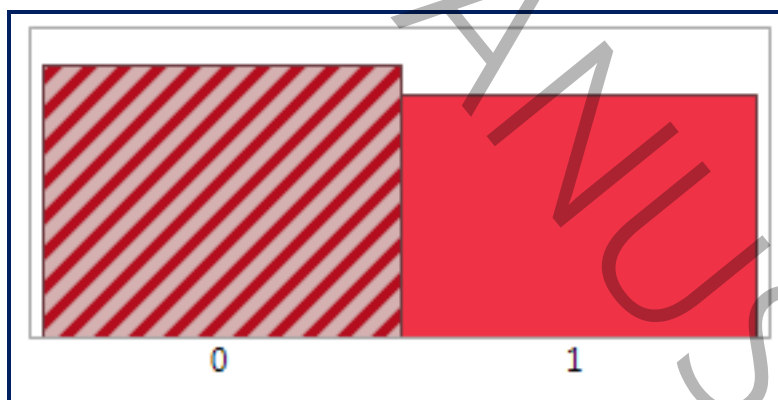
۴.۵. بررسی میزان تأثیر زمان و سایر متغیرهای ترافیکی بر خطای ثبت زوج یا فرد بودن پلاک

مرحله اول) متوازن کردن داده ها: داده های متوازنی که بر مبنای متغیر خطای ثبت زوج یا فرد بودن پلاک و با استفاده از الگوریتم SMOTE ساخته و تولید گردید، شامل ۴۳۲۵۶ رکورد می باشد.

جدول ۵: توزیع فراوانی خطای ثبت زوج یا فرد بودن پلاک

Table 5: Frequency distribution of errors in recording license plate parity (odd/even)

خطا	کد	تعداد	درصد
ثبت زوج یا فرد بودن پلاک	عدم خطا=۰	۲۲۹۰۰	۵۳
عدم ثبت زوج یا فرد بودن پلاک	خطا=۱	۲۰۳۵۶	۴۷
جمع		۴۳۲۵۶	۱۰۰



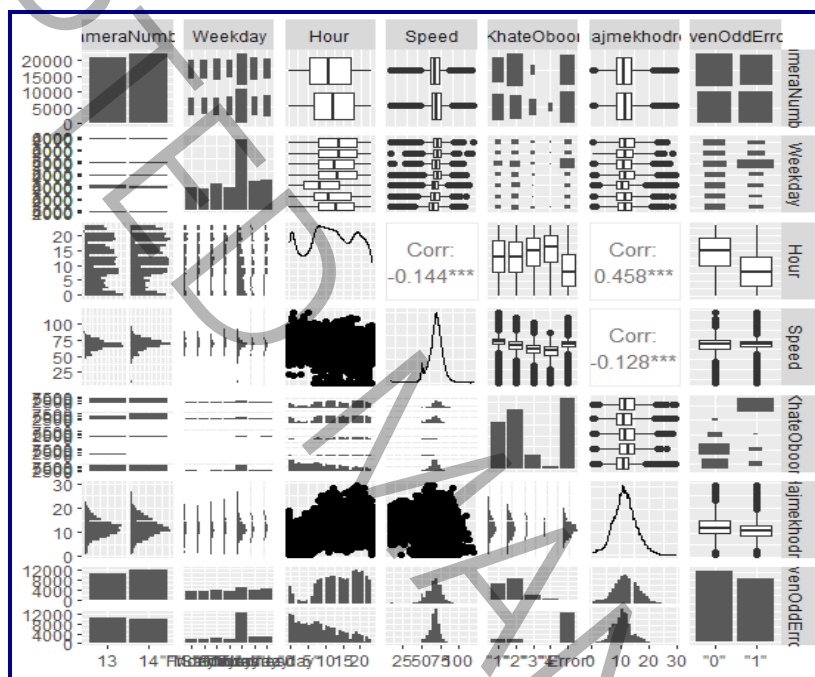
شکل ۱۱: نمودار ستونی توزیع فراوانی خطای ثبت زوج یا فرد بودن پلاک در داده های متوازن

Figure 11: Bar chart showing the frequency distribution of errors in recording license plate parity (odd/even) in the balanced dataset.

همان گونه که در شکل (۱۱) مشاهده می شود، پس از اعمال الگوریتم SMOTE، توزیع فراوانی کلاس های «خطا» و «عدم خطا» در متغیر زوج یا فرد بودن پلاک تقریباً متعادل شده است. متوازن سازی داده ها موجب کاهش سوگیری مدل نسبت به کلاس اکثریت شده و شرایط مناسبی را برای آموزش الگوریتم های یادگیری ماشین فراهم می کند. در نتیجه، مدل قادر خواهد بود الگوهای مربوط به

هر دو کلاس را با دقت بیشتری یاد گرفته و عملکرد قابل اعتمادتری در پیش‌بینی خطای تشخیص زوج یا فرد بودن پلاک بر روی داده‌های جدید ارائه دهد.

مرحله دوم) تعریف متغیرها: برای داده‌های متوازن مربوط به خطای ثبت زوج یا فرد بودن پلاک ۳۰۲۷۹ رکورد برای داده‌های آموزشی^۵ و ۱۲۹۷۷ رکورد برای داده‌های آزمایشی^۶ در نظر گرفته شده است. در این بخش ابتدا در ماتریس نمودار پراکنش^۷ که در شکل (۱۲) ارائه گردیده است، رابطه بین هر دو متغیر مذکور مورد مطالعه قرار می‌گیرد. شایان ذکر است که در تمامی کلاس‌بندی‌های ارائه شده در ادامه برای متغیر خطای ثبت زوج یا فرد بودن پلاک، میزان دقت و خطا در داده‌های آموزشی و داده‌های آزمایشی نزدیک هم هستند، لذا در ادامه تنها موارد مربوط به داده‌های آزمایشی را گزارش می‌کنیم.



شکل ۱۲: ماتریس نمودار پراکنش متغیرهای حاضر در مدل‌های کلاس‌بندی برای متغیر خطای ثبت زوج یا فرد بودن پلاک

Figure 12: Scatter plot matrix of variables included in the classification models for the license plate parity recording error variable.

همان‌گونه که در شکل (۱۲) مشاهده می‌شود، ماتریس نمودارهای پراکنش، نحوه توزیع متغیرهای پژوهش و روابط دوطرفه میان آن‌ها را در ارتباط با متغیر خطای تشخیص زوج یا فرد بودن پلاک نمایش می‌دهد. نمودارهای روی قطر اصلی، توزیع فراوانی هر متغیر را نشان می‌دهند، در حالی که نمودارهای خارج از قطر بیانگر الگوی پراکندگی و میزان وابستگی بین متغیرها هستند. نتایج بیانگر آن است که بین برخی متغیرها همبستگی خطی ضعیفی وجود دارد؛ به‌عنوان نمونه، بین ساعت ثبت پلاک و سرعت وسیله نقلیه همبستگی مثبت ضعیفی در حدود ۰,۱۴۴ و بین ساعت ثبت پلاک و خطای تشخیص زوج یا فرد بودن پلاک همبستگی

^۵ . Training data

^۶ . Test data

^۷ . Scatterplot Matrix

مثبت در حدود ۰,۴۵۸ مشاهده می‌شود، در حالی که بین سرعت وسیله نقلیه و خطای تشخیص زوج یا فرد بودن پلاک همبستگی منفی ضعیفی (حدود -۰,۱۲۸) وجود دارد. به‌طور کلی، مقادیر پایین ضرایب همبستگی نشان می‌دهد که روابط میان متغیرها عمدتاً غیرخطی بوده و هم‌خطی شدید بین متغیرهای ورودی مشاهده نمی‌شود؛ بنابراین، استفاده از الگوریتم‌های یادگیری ماشین مبتنی بر روابط غیرخطی، از جمله جنگل تصادفی، برای تحلیل عوامل مؤثر بر خطای تشخیص زوج یا فرد بودن پلاک مناسب و قابل توجیه است.

مرحله سوم) ورود اطلاعات به مدل ها و تحلیل نتایج: نتایج الگوریتم کلاس‌بندی جنگل تصادفی به شرح زیر مورد بررسی قرار گرفته است.

برای رتبه‌بندی ویژگی‌های تأثیرگذار بر متغیر خطای ثبت زوج و فرد بودن پلاک از جنگل تصادفی استفاده شده است که ماتریس انطباقی آن به‌همراه نمودار ضرایب کاهش جینی^۸ در ادامه گزارش می‌گردد.

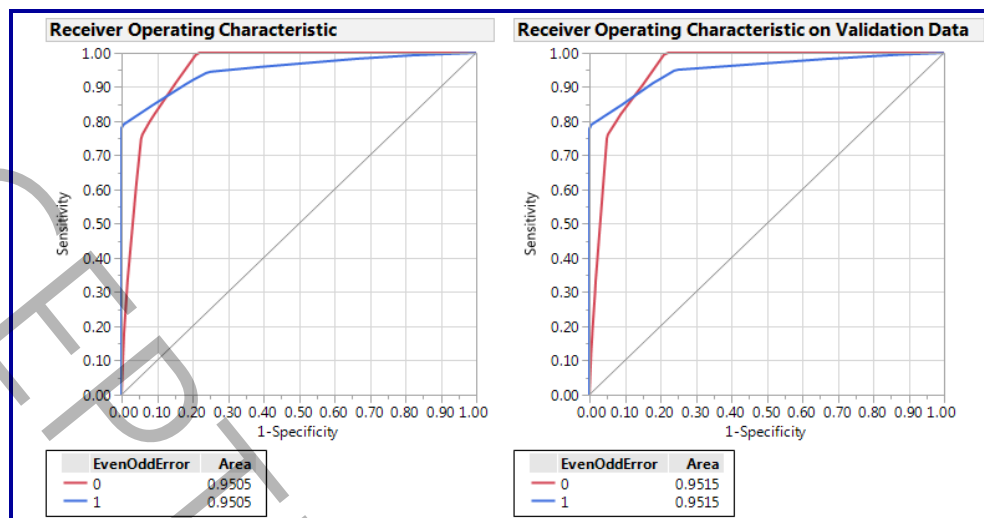
جدول ۶: ماتریس انطباقی داده‌های آزمایشی جنگل تصادفی برای خطای ثبت زوج و فرد بودن پلاک

Table 6: Random Forest experimental data matching matrix for license plate odd/even registration errors.

واقعی		ماتریس انطباقی داده‌های آزمایشی	
خطا	عدم خطا		
۳۳	۱۳۲۲۲	خطا	پیش‌بینی شده
۱۵۹۹۷	۱۰۲۷	عدم خطا	
دقت = ۰/۹۶۵ و MSE=۰/۰۴۳			

مرحله چهارم) ارزیابی و انتخاب مدل مناسب: یکی از معیارهای ارزیابی مناسب مد برابری خطا و دقت در داده‌های آموزشی و آزمایشی می‌باشد، که این برابری برای تمامی مدل‌های کلاس‌بندی اجرا شده برای خطای ثبت زوج و فرد بودن پلاک برقرار بوده است. همچنین، با توجه به میزان دقت و خطای مدل‌های مورد بررسی، تمامی مدل‌ها دقت بالایی داشته‌اند. دقت مدل جنگل تصادفی نسبت به سایر مدل‌ها با ۹۶/۵ درصد دارای دقت بالاتری است.

^۸ . MeanDecreaseGini

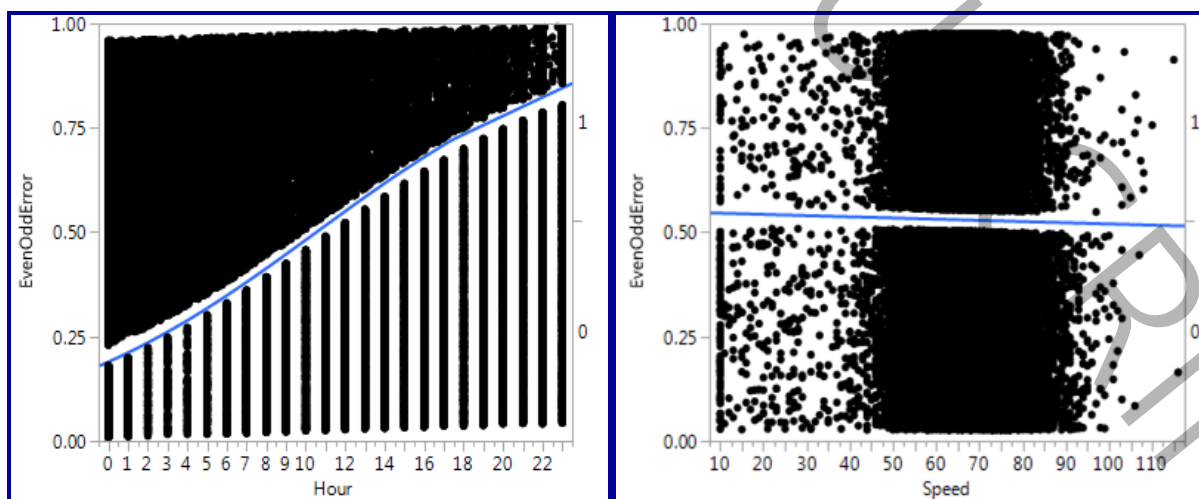


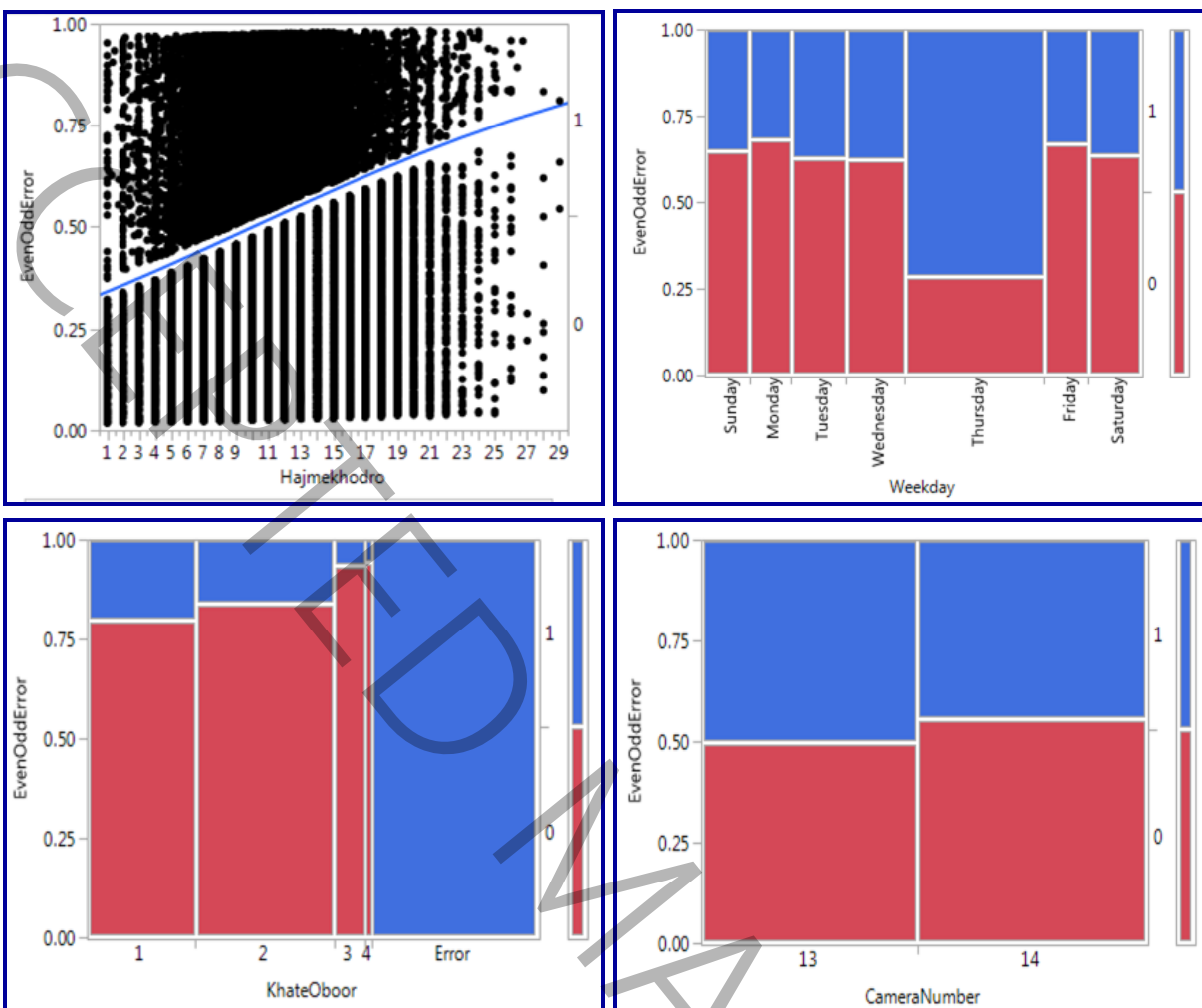
شکل ۱۳: منحنی‌های ROC برای داده‌های آموزشی و آزمایشی در مدل جنگل تصادفی برای خطای ثبت زوج و فرد بودن پلاک

Figure 13: ROC curves for training and testing data in the Random Forest model for license plate odd/even registration errors.

همان‌گونه که در شکل (۱۳) مشاهده می‌شود، منحنی‌های ROC مربوط به داده‌های آموزشی و آزمایشی، عملکرد مدل جنگل تصادفی را در پیش‌بینی خطای ثبت زوج و فرد بودن پلاک نشان می‌دهند. مقدار سطح زیر منحنی (AUC) برای داده‌های آموزشی برابر با ۰/۹۵۰۵ و برای داده‌های آزمایشی برابر با ۰/۹۵۱۵ به دست آمده است. این مقادیر بیانگر توانایی بسیار بالای مدل در تفکیک صحیح نمونه‌های دارای خطا و بدون خطا است. همچنین، اختلاف ناچیز بین مقادیر AUC در دو مجموعه آموزشی و آزمایشی نشان می‌دهد که مدل از پایداری و قابلیت تعمیم مناسبی برخوردار بوده و دچار بیش‌برازش نشده است. قرار گرفتن منحنی‌های ROC در فاصله قابل توجهی بالاتر از خط مرجع نیز مؤید قدرت پیش‌بینی مطلوب مدل جنگل تصادفی در شناسایی خطای ثبت زوج و فرد بودن پلاک است.

همان‌طور که شکل (۱۵) نشان می‌دهد، مساحت زیر منحنی‌های ROC برای این مدل بیش از ۹۵ درصد است که نشان می‌دهد مدل جنگل تصادفی ساخته شده برای متغیر خطای ثبت زوج و فرد بودن پلاک بسیار مناسب و دقیق است.





شکل ۱۴: نمودار پراکنش رابطه بین خطای ثبت زوج و فرد بودن پلاک و متغیرهای تحقیق

Figure 14: Scatter plot of the relationship between the error in recording license plate parity (odd/even) and the research variables.

همان‌گونه که در شکل (۱۴) مشاهده می‌شود، نمودارهای پراکنش، نحوه توزیع داده‌ها و ارتباط اولیه میان متغیرهای مستقل با متغیر خطای ثبت زوج و فرد بودن پلاک را نشان می‌دهند. بررسی این نمودارها بیانگر آن است که الگوی وقوع خطا از یک رابطه خطی مشخص تبعیت نمی‌کند و میزان پراکندگی داده‌ها در متغیرهای مختلف متفاوت است. همچنین، تمرکز داده‌ها در برخی بازه‌های زمانی، تغییرات حجم ترافیک، سرعت حرکت وسایل نقلیه، شماره دوربین و خط عبوری نشان می‌دهد که این متغیرها می‌توانند در بروز خطای تشخیص زوج و فرد بودن پلاک نقش مؤثری داشته باشند. علاوه بر این، همپوشانی نسبی نمونه‌های مربوط به کلاس‌های «خطا» و «عدم خطا» بیانگر آن است که تفکیک این دو کلاس با استفاده از روش‌های آماری خطی به‌سادگی امکان‌پذیر نیست. از این‌رو، استفاده از مدل‌های یادگیری ماشین مبتنی بر روابط غیرخطی، از جمله الگوریتم جنگل تصادفی، می‌تواند در شناسایی الگوهای مؤثر بر این نوع خطا عملکرد مناسب‌تری داشته باشد.

نتایج این پژوهش نشان داد که الگوریتم جنگل تصادفی در هر سه مسئله طبقه‌بندی شامل خطای ثبت پلاک، خطای ثبت حرف پلاک و خطای تشخیص زوج یا فرد بودن پلاک عملکرد بهتری نسبت به شبکه عصبی مصنوعی ارائه کرده است. این برتری را می‌توان به ساختار یادگیری گروهی در جنگل تصادفی نسبت داد؛ زیرا این الگوریتم با ترکیب نتایج تعداد زیادی درخت تصمیم، قادر است روابط پیچیده و غیرخطی میان متغیرهای ترافیکی را بهتر مدل‌سازی کرده و حساسیت کمتری نسبت به نویز داده‌ها و بیش‌برازش داشته باشد. در مقابل، عملکرد شبکه عصبی مصنوعی به انتخاب ساختار شبکه، تنظیم پارامترهای آموزشی و حجم داده وابسته است و در مجموعه داده‌های محدود یا دارای ناهمگنی، ممکن است به اندازه الگوریتم‌های مبتنی بر درخت تصمیم کارایی نداشته باشد. از این‌رو، نتایج این پژوهش بیانگر آن است که در تحلیل داده‌های حاصل از سامانه‌های پلاک‌خوان شهری، استفاده از روش‌های مبتنی بر جنگل تصادفی می‌تواند گزینه مناسب‌تری برای پیش‌بینی و تحلیل خطاها باشد.

بررسی اهمیت متغیرها نیز نشان داد که **ساعت ثبت پلاک** بیشترین تأثیر را بر وقوع خطای ثبت پلاک و خطای ثبت حرف پلاک دارد، در حالی که **خط عبوری** مهم‌ترین عامل در بروز خطای تشخیص زوج یا فرد بودن پلاک بوده است. این نتایج را می‌توان با شرایط واقعی بهره‌برداری از سامانه‌های تشخیص خودکار پلاک تبیین کرد. تغییرات شدت نور در ساعات مختلف شبانه‌روز، تابش مستقیم خورشید، بازتاب نور از بدنه خودروها، کاهش روشنایی در ساعات شب و تغییر کیفیت تصاویر ثبت‌شده می‌تواند موجب کاهش دقت الگوریتم‌های تشخیص نویسه شود. همچنین، افزایش سرعت وسایل نقلیه و حجم ترافیک در ساعات اوج، احتمال ایجاد **تاری حرکتی**، هم‌پوشانی خودروها و کاهش وضوح تصویر را افزایش می‌دهد. از سوی دیگر، تأثیر بالای متغیر خط عبوری نشان می‌دهد که زاویه دید دوربین، فاصله دوربین تا مسیر حرکت خودرو، نحوه استقرار وسایل نقلیه در خطوط مختلف و حتی کیفیت تصویربرداری در هر خط عبوری می‌تواند بر عملکرد سامانه ANPR تأثیرگذار باشد. بنابراین، بخشی از خطاهای مشاهده‌شده علاوه بر ویژگی‌های داده، ناشی از شرایط واقعی نصب و بهره‌برداری دوربین‌های پلاک‌خوان است.

یافته‌های این پژوهش با نتایج مطالعات پیشین همخوانی قابل توجهی دارد. پژوهش‌های اخیر نشان داده‌اند که الگوریتم‌های یادگیری گروهی و مدل‌های مبتنی بر درخت تصمیم در بسیاری از مسائل مرتبط با سامانه‌های حمل‌ونقل هوشمند، از جمله تشخیص پلاک خودرو، دقت و پایداری بیشتری نسبت به مدل‌های منفرد ارائه می‌کنند. همچنین، مطالعات مختلف گزارش کرده‌اند که عواملی مانند شرایط نوری، کیفیت تصویر، زاویه نصب دوربین، سرعت وسایل نقلیه، شرایط جوی و تراکم ترافیک از مهم‌ترین عوامل مؤثر بر عملکرد سامانه‌های ANPR هستند. نتایج این پژوهش نیز این یافته‌ها را تأیید کرده و نشان می‌دهد که متغیرهای زمانی و ترافیکی می‌توانند به‌عنوان شاخص‌هایی غیرمستقیم برای ارزیابی کیفیت عملکرد دوربین‌های پلاک‌خوان مورد استفاده قرار گیرند. از دیدگاه کاربردی، نتایج این تحقیق بیانگر آن است که کاهش خطاهای سامانه‌های ANPR تنها از طریق توسعه الگوریتم‌های هوش مصنوعی امکان‌پذیر نیست، بلکه بهینه‌سازی محل نصب دوربین، تنظیم زاویه دید، بهبود کیفیت تجهیزات تصویربرداری، کنترل شرایط نورپردازی و نگهداری دوره‌ای تجهیزات نیز نقش مهمی در افزایش دقت این سامانه‌ها ایفا می‌کند.

با وجود نتایج مطلوب، این پژوهش دارای محدودیت‌هایی نیز است. داده‌های مورد استفاده تنها مربوط به دو دوربین پلاک‌خوان و یک دوره زمانی یک‌هفته‌ای بوده و از این‌رو تعمیم نتایج به سایر معابر شهری یا شرایط محیطی متفاوت باید با احتیاط انجام شود. علاوه بر این، در این مطالعه تمرکز بر تحلیل داده‌های خروجی سامانه بوده و پارامترهای داخلی الگوریتم‌های پردازش تصویر،

کیفیت تصاویر خام، شرایط آبوهوایی و تنظیمات نرم‌افزار تشخیص پلاک به صورت مستقیم بررسی نشده است. بنابراین، پیشنهاد می‌شود در پژوهش‌های آینده با استفاده از داده‌های بلندمدت، تعداد بیشتری از دوربین‌ها و ترکیب متغیرهای ترافیکی با شاخص‌های محیطی و پارامترهای پردازش تصویر، مدل‌های جامع‌تری برای پیش‌بینی و کاهش خطاهای سامانه‌های تشخیص خودکار پلاک توسعه یابد.

به منظور تفسیر بهتر نتایج و ارزیابی جایگاه یافته‌های پژوهش حاضر، نتایج حاصل با مطالعات پیشین در حوزه سامانه‌های تشخیص خودکار پلاک و کاربرد روش‌های یادگیری ماشین در سیستم‌های حمل‌ونقل هوشمند مقایسه شد. این مقایسه، علاوه بر بررسی میزان همخوانی نتایج، تفاوت‌ها و نوآوری‌های پژوهش حاضر را نسبت به مطالعات پیشین آشکار می‌سازد.

جدول ۷. مقایسه نتایج پژوهش حاضر با مطالعات پیشین

Table 7. Comparison of the present study's results with previous studies

مقایسه با پژوهش حاضر	مهم‌ترین یافته	مدل مورد استفاده	هدف پژوهش	مرجع
نتایج پژوهش حاضر نیز نشان داد که کاهش خطاهای ثبت پلاک می‌تواند قابلیت اعتماد داده‌های ترافیکی را افزایش دهد.	کیفیت داده‌های ANPR بر دقت تحلیل‌های ترافیکی تأثیر مستقیم دارد.	تحلیل آماری	اعتبارسنجی داده‌های ANPR در مدل‌های ترافیکی	Robinson and Venter (2019)
پژوهش حاضر نیز برتری مدل جنگل تصادفی را نسبت به شبکه عصبی در داده‌های ANPR نشان داد.	روش‌های مبتنی بر یادگیری ماشین عملکرد مناسبی در داده‌های پیچیده دارند.	الگوریتم‌های هوشمند	کاربرد روش‌های هوشمند در ITS	Gaber et al. (2018)
یافته‌های این پژوهش نیز اهمیت کنترل کیفیت داده‌های ثبت‌شده توسط دوربین‌های پلاک‌خوان را تأیید می‌کند.	کیفیت داده‌های اولیه نقش تعیین‌کننده‌ای در دقت تحلیل‌ها دارد.	مدل‌های مبتنی بر Tensor	تکمیل داده‌های ناقص ترافیکی	Tan et al. (2013)
علاوه بر مقایسه مدل‌ها، عوامل مؤثر بر وقوع هر نوع خطا نیز رتبه‌بندی و تحلیل شدند که یکی از نوآوری‌های اصلی این پژوهش است.	دقت مدل جنگل تصادفی برای خطاهای ثبت پلاک، ثبت حرف پلاک و زوج/فرد به ترتیب ۸۶، ۸۵ و ۹۶٫۵ درصد به دست آمد.	Random Forest و ANN	تحلیل خطاهای دوربین‌های پلاک‌خوان شهری	پژوهش حاضر

همان گونه که جدول ۷ نشان می دهد، نتایج پژوهش حاضر با مطالعات پیشین از نظر اهمیت کیفیت داده های سامانه های ANPR و کارایی روش های یادگیری ماشین همخوانی دارد. با این حال، برخلاف اغلب مطالعات که تمرکز آن ها بر بهبود الگوریتم تشخیص پلاک یا ارزیابی کلی عملکرد سامانه بوده است، پژوهش حاضر به تحلیل عوامل مؤثر بر انواع مختلف خطاهای ثبت اطلاعات پرداخته و اهمیت نسبی متغیرهای ترافیکی و زمانی را با استفاده از مدل جنگل تصادفی استخراج کرده است. این ویژگی، پژوهش حاضر را از مطالعات پیشین متمایز ساخته و می تواند در مدیریت بهره برداری، کالیبراسیون و پایش عملکرد دوربین های پلاک خوان شهری مورد استفاده قرار گیرد.

یکی از محدودیت های اصلی این پژوهش، استفاده از داده های مربوط به دو دوربین پلاک خوان و یک بازه زمانی یک هفته ای است. این محدودیت ناشی از دسترسی به داده های معتبر و کنترل شده سامانه ANPR بوده است. هدف اصلی این پژوهش، ارائه یک مطالعه اکتشافی برای شناسایی عوامل مؤثر بر خطاهای سامانه و مقایسه عملکرد مدل های یادگیری ماشین بوده است، نه ارائه مدلی با قابلیت تعمیم به تمامی دوربین های شهری. از این رو، تعمیم نتایج به سایر معابر، شهرها و دوره های زمانی نیازمند استفاده از داده های گسترده تر در مطالعات آینده است.

۷. نتیجه گیری

سامانه های خودکار تشخیص پلاک خودرو به عنوان یکی از اجزای اصلی سیستم های حمل و نقل هوشمند، نقش مهمی در مدیریت ترافیک، پایش تخلفات و تولید داده های مورد استفاده در برنامه ریزی های حمل و نقل ایفا می کنند. با این وجود، وجود خطا در فرآیند ثبت اطلاعات این سامانه ها می تواند بر کیفیت داده های ترافیکی و در نتیجه بر تصمیم گیری مدیران شهری تأثیر منفی بگذارد. از این رو، شناسایی عوامل مؤثر بر وقوع این خطاها و تحلیل الگوهای آن ها از اهمیت بالایی برخوردار است.

در این پژوهش، عملکرد دو دوربین پلاک خوان دارای بیشترین نرخ خطا در شهر مشهد مورد بررسی قرار گرفت و سه نوع خطای اصلی شامل خطای ثبت پلاک، خطای ثبت حرف پلاک و خطای تشخیص زوج یا فرد بودن پلاک تحلیل شد. برای این منظور، از متغیرهای شماره دوربین، روز هفته، ساعت ثبت، سرعت وسیله نقلیه، خط عبوری و حجم ترافیک به عنوان متغیرهای مستقل استفاده شد. همچنین به منظور رفع مشکل نامتوازن بودن داده ها، از روش SMOTE بهره گرفته شد و عملکرد مدل جنگل تصادفی با مدل شبکه عصبی مصنوعی مقایسه گردید.

نتایج نشان داد که مدل جنگل تصادفی در هر سه مسئله طبقه بندی، عملکرد بهتری نسبت به مدل شبکه عصبی مصنوعی داشته است. دقت این مدل برای پیش بینی خطای ثبت پلاک، خطای ثبت حرف پلاک و خطای تشخیص زوج یا فرد بودن پلاک به ترتیب برابر با ۸۵، ۸۶ و ۵/۹۶ درصد به دست آمد. بررسی اهمیت متغیرها نیز نشان داد که در خطاهای ثبت پلاک و ثبت حرف پلاک، متغیر ساعت ثبت بیشترین نقش را در وقوع خطا دارد، در حالی که در خطای تشخیص زوج یا فرد بودن پلاک، متغیر خط عبوری مهم ترین عامل تأثیرگذار شناخته شد. این نتایج بیانگر آن است که علاوه بر ویژگی های فنی سامانه، شرایط بهره برداری و ویژگی های ترافیکی نیز می توانند بر میزان خطای سامانه های ANPR اثرگذار باشند.

از دیدگاه کاربردی، یافته های این پژوهش می تواند در اولویت بندی اقدامات نگهداری و کالیبراسیون دوربین های پلاک خوان، پایش کیفیت داده های ترافیکی و بهبود قابلیت اعتماد سامانه های حمل و نقل هوشمند مورد استفاده سازمان های مدیریت ترافیک قرار گیرد. همچنین شناسایی متغیرهای اثرگذار بر وقوع خطا می تواند به مدیران شهری در تدوین راهکارهای کاهش خطا و افزایش دقت برداشت اطلاعات کمک کند.

با وجود نتایج قابل قبول، این پژوهش دارای محدودیت‌هایی نیز است. داده‌های مورد استفاده تنها مربوط به دو دوربین پلاک‌خوان و یک بازه زمانی یک‌هفته‌ای بوده است؛ بنابراین تعمیم نتایج به سایر دوربین‌ها، شرایط محیطی و بازه‌های زمانی مختلف باید با احتیاط انجام شود. همچنین در این پژوهش، تمرکز بر تحلیل داده‌های خروجی سامانه بوده و پارامترهای داخلی الگوریتم پردازش تصویر و فرآیند تشخیص پلاک مورد بررسی قرار نگرفته است.

پیشنهاد می‌شود در پژوهش‌های آینده، داده‌های مربوط به تعداد بیشتری از دوربین‌ها، دوره‌های زمانی طولانی‌تر و شرایط محیطی مختلف نظیر تغییرات نور، شرایط جوی، زاویه نصب دوربین، کیفیت تصاویر و تراکم ترافیک مورد بررسی قرار گیرد. همچنین استفاده از مدل‌های یادگیری عمیق و روش‌های ترکیبی می‌تواند زمینه مناسبی برای افزایش دقت تشخیص خطا و توسعه سامانه‌های هوشمند مدیریت ترافیک فراهم سازد.

۸. محدودیت تحقیق

پژوهش حاضر نیز همانند سایر مطالعات دارای محدودیت‌هایی است که باید در تفسیر نتایج مورد توجه قرار گیرد. نخست، داده‌های مورد استفاده تنها مربوط به دو دوربین پلاک‌خوان شهری در کلان‌شهر مشهد و بازه زمانی یک هفته بوده است؛ از این رو، تعمیم نتایج به سایر معابر، دوربین‌ها و دوره‌های زمانی مختلف باید با احتیاط انجام شود.

دوم، در این پژوهش تحلیل‌ها بر اساس داده‌های خروجی سامانه تشخیص خودکار پلاک (ANPR) انجام شد و پارامترهای داخلی الگوریتم پردازش تصویر، تنظیمات نرم‌افزار، کیفیت تصاویر خام و مشخصات سخت‌افزاری دوربین‌ها در دسترس نبود؛ بنابراین، امکان بررسی مستقیم تأثیر این عوامل بر بروز خطا فراهم نشد.

سوم، برخی عوامل محیطی نظیر شدت نور، شرایط جوی، بارندگی، مه، آلودگی لنز دوربین، زاویه نصب دوربین و کیفیت تصویربرداری به صورت مستقیم اندازه‌گیری نشده و تنها اثر غیرمستقیم آن‌ها از طریق متغیرهای زمانی و ترافیکی مورد بررسی قرار گرفت.

چهارم، در این مطالعه تنها عملکرد دو الگوریتم جنگل تصادفی و شبکه عصبی مصنوعی مقایسه شد؛ بنابراین، امکان مقایسه نتایج با سایر روش‌های پیشرفته یادگیری ماشین و یادگیری عمیق مانند XGBoost، LightGBM، CatBoost یا شبکه‌های عصبی کانولوشنی (CNN) وجود نداشت.

با وجود این محدودیت‌ها، استفاده از داده‌های واقعی سامانه‌های پلاک‌خوان شهری و تحلیل هم‌زمان سه نوع خطای ثبت پلاک، ثبت حرف پلاک و تشخیص زوج یا فرد بودن پلاک موجب شده است نتایج این پژوهش بتواند در شناسایی عوامل مؤثر بر عملکرد سامانه‌های ANPR و بهبود کیفیت داده‌های مورد استفاده در سامانه‌های حمل‌ونقل هوشمند، کاربرد عملی مناسبی داشته باشد.

۹. پیشنهادهای پژوهش

با توجه به نتایج این پژوهش، پیشنهادهای زیر برای تحقیقات آینده ارائه می‌شود:

۱. گسترش دامنه داده‌های مورد مطالعه: پیشنهاد می‌شود در پژوهش‌های آینده از داده‌های مربوط به تعداد بیشتری از دوربین‌های پلاک‌خوان، معابر مختلف شهری و بازه‌های زمانی طولانی‌تر استفاده شود تا امکان تعمیم نتایج و بررسی پایداری مدل‌ها در شرایط گوناگون فراهم گردد.

۲. بررسی عوامل محیطی و فنی: توصیه می‌شود تأثیر متغیرهایی نظیر شدت نور، شرایط جوی، زاویه نصب دوربین، فاصله دوربین از مسیر حرکت، کیفیت تصاویر، تاری ناشی از حرکت (Motion Blur) و آلودگی لنز به صورت مستقیم در مدل‌های پیش‌بینی وارد شود تا نقش این عوامل در بروز خطاهای سامانه‌های ANPR با دقت بیشتری ارزیابی گردد.

۳. مقایسه با سایر الگوریتم‌های یادگیری ماشین: در پژوهش‌های آتی، عملکرد الگوریتم جنگل تصادفی با سایر روش‌های پیشرفته مانند XGBoost، LightGBM، CatBoost، Support Vector Machine (SVM) و مدل‌های یادگیری

عمیق نظیر شبکه‌های عصبی کانولوشنی (CNN) مقایسه شود تا مناسب‌ترین مدل برای پیش‌بینی خطاهای سامانه‌های پلاک‌خوان تعیین گردد .

۴. تحلیل تصاویر خام دوربین‌ها :پیشنهاد می‌شود علاوه بر تحلیل داده‌های خروجی سامانه، تصاویر خام ثبت‌شده توسط دوربین‌ها نیز مورد بررسی قرار گیرد تا ارتباط بین ویژگی‌های تصویر و میزان خطای تشخیص پلاک به‌طور مستقیم تحلیل شود .

۵. ارائه مدل‌های پیش‌بینی بلادرنگ :توسعه مدل‌هایی که بتوانند احتمال وقوع خطا را به‌صورت برخط (Real-Time) پیش‌بینی کرده و پیش از ثبت اطلاعات، اپراتور یا سامانه را از احتمال بروز خطا آگاه سازند، می‌تواند گام مؤثری در افزایش دقت سامانه‌های تشخیص خودکار پلاک باشد .

۶. کاربرد نتایج در مدیریت سامانه‌های حمل‌ونقل هوشمند :پیشنهاد می‌شود از نتایج این پژوهش برای بهینه‌سازی محل نصب دوربین‌ها، تنظیم پارامترهای بهره‌برداری، زمان‌بندی عملیات نگهداری و کالیبراسیون تجهیزات و همچنین ارتقای کیفیت داده‌های مورد استفاده در سامانه‌های حمل‌ونقل هوشمند استفاده شود .

۷. توسعه مدل‌های ترکیبی :استفاده از مدل‌های هیبریدی که قابلیت ترکیب الگوریتم‌های یادگیری ماشین با روش‌های پردازش تصویر را دارند، می‌تواند ضمن افزایش دقت تشخیص، پایداری سامانه‌های ANPR را در شرایط محیطی مختلف بهبود بخشد.

فهرست علائم

- 1- ANPR: Automatic Number Plate Recognition
- 2- ITS: Intelligent Transportation Systems
- 3- ANN: Artificial Neural Network
- 4- RF: Random Forest
- 5- SMOTE: Synthetic Minority Over-sampling Technique
- 6- ROC: Receiver Operating Characteristic
- 7- AUC: Area Under the Curve
- 8- CRISP-DM: Cross-Industry Standard Process for Data Mining

1. M. T. KhQadri, M. Asif, Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition, Proceedings of the International Conference on Education Technology and Computer, pp. 335–338, (2009).
2. A. Roy, D. P. Ghoshal, Number Plate Recognition for Use in Different Countries Using an Improved Segmentation, Proceedings of the National Conference on Emerging Trends and Applications in Computer Science, pp. 1–5, (2011).
3. H. Tan, G. Feng, J. Feng, W. Wang, Y. J. Zhang, F. Li, A Tensor-Based Method for Missing Traffic Data Completion, Transportation Research Part C: Emerging Technologies, 28, pp. 15–27, (2013).
4. T. Gaber, S. Abdelwahab, M. Elhoseny, A. E. Hassanien, Trust-Based Secure Clustering in WSN-Based Intelligent Transportation Systems, Computer Networks, 146, pp. 151–168, (2018).
5. A. Robinson, C. Venter, Validating Traffic Models Using Large-Scale Automatic Number Plate Recognition (ANPR) Data, Journal of the South African Institution of Civil Engineering, 61(3), pp. 45–57, (2019).
6. B. Gunay, M. Saltan, A. Ilgaz, Y. Albayrak, Area-Wide ANPR Coverage with a Small Number of Instrumented Vehicles, Proceedings of the International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 379–386, (2015).
7. A. Homayounfar, A. T. S. Ho, N. Zhu, G. Head, P. Palmer, Multi-Vehicle Convoy Analysis Based on ANPR Data, Proceedings of the International Conference on Imaging for Crime Detection and Prevention, (2011).
8. S. S. Anand, A. G. Büchner, Decision Support Using Data Mining, Financial Times Management, London, UK, (1998).
9. D. Soole, J. Fleiter, B. Watson, Point-to-Point Speed Enforcement, Austroads Project No. SS1649, Austroads Ltd., Sydney, Australia, (2012).
10. C. Patel, D. Shah, A. Patel, Automatic Number Plate Recognition System (ANPR): A Survey, International Journal of Computer Applications, 69(9), (2013).
11. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, 39(11), pp. 27–34, (1996).
12. S. J. Rigatti, Random Forest, Journal of Insurance Medicine, 47(1), pp. 31–39, (2017).
13. C. Morais, K. L. Yung, K. Johnson, R. Moura, M. Beer, E. Patelli, Identification of Human Errors and Influencing Factors: A Machine Learning Approach, Safety Science, 146, 105528, (2022).
14. Lubna, N. Mufti, S. A. A. Shah, Automatic Number Plate Recognition: A Detailed Survey of Relevant Algorithms, Sensors, 21(9), 3028, (2021).
15. M. A. K. Khani, M. Usama, A. S. Shah, A. Shah, S. H. Abbas, A. Maqsood, A. A. Laghari, Intelligent Vehicle Number Plate Recognition System Using YOLO for Enhanced Security in Smart Buildings, Journal of ICT, Design, Engineering and Technological Science, 8(2), pp. 11–17, (2024).
16. S. Parvin, L. J. Rozario, M. E. Islam, Vehicle Number Plate Detection and Recognition Techniques: A Review, Advances in Science, Technology and Engineering Systems Journal, 6(2), pp. 423–438, (2021).
17. G. Pandey, N. Lamichhane, U. Subedi, CNN-Based System for Automatic Number Plate Recognition, Journal of Soft Computing Paradigm, 5(4), pp. 347–364, (2023).
18. T. Mustafa, M. Karabatak, Real-Time Car Model and Plate Detection System by Using Deep Learning Architectures, IEEE Access, 12, pp. 107616–107630, (2024).
19. M. A. Ebrahimzadeh, N. Nakhaie, M. Dabbaqi, A. Abdi Kordani, Analysis and Evaluation of Vehicle Traffic Counting in Inner-City Automatic Number Plate Recognition by Neural Network and Random Forest Models (Case Study of Mashhad Metropolis), Road, Articles in Press, (2024).
20. M. A. Ebrahimzadeh, M. Dabbaqi, N. Nakhaie, Investigating the Reduction of the Error Caused by the Counting of Vehicle Traffic in City License Plate Reader Cameras by Mathematical Models (Case Study: Mashhad City), (2025).
21. T. Hastie, G. James, D. Witten, R. Tibshirani, An Introduction to Statistical Learning, Springer, New York, (2013).
22. L. Breiman, Random Forests, Machine Learning, 45(1), pp. 5–32, (2001).