Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال سوم، شماره دوم، پاییز و زمستان ۱۴۰۴، صفحات ۱۸۶-۱۷۵

تحلیل و شناسایی ویژگی‌های کلیدی برای طبقه‌بندی عیوب سطحی صفحات فولادی با استفاده از مدل‌های یادگیری ماشین و روش متعادل‌سازی داده‌ها

حبیب خدادادی [۱*]، سید علی مهری [۲] مهسا خدادادی [۳]

[۱] استادیار، گروه مهندسی برق و کامپیوتر، دانشکده فنی و مهندسی، دانشگاه هرمزگان، بندرعباس، ایران

[۲] مهندس کامپیوتر، گروه کامپیوتر، واحد بندرعباس، دانشگاه آزاد اسلامی، بندرعباس، ایران

[۳] مربی، گروه مهندسی مواد و متالورژی، واحد میناب، دانشگاه آزاد اسلامی، میناب، ایران

چکیده

مقاله پژوهشی

در این پژوهش چالش‌های اصلی در تشخیص عیوب سطحی صفحات فولادی ازجمله نامتوازن‌بودن داده‌ها، هم‌پوشانی ویژگی‌ها و دشواری در شناسایی ویژگی‌های مؤثر مورد بررسی قرارگرفته است. هدف اصلی پژوهش، شناسایی مؤثرترین ویژگی‌ها در افزایش دقت مدل‌های طبقه‌بندی عیوب نظیر خراش، لکه و برجستگی است. برای این منظور سه مرحله تحلیل انجام شد: مرحله اول و دوم شامل اجرای الگوریتم‌های جنگل تصادفی و تقویت گرادیان حداکثری بر داده خام و مرحله سوم شامل اعمال همین الگوریتم‌ها بر داده‌های متوازن‌شده با روش SMOTE بود. نتایج نشان داد که استفاده از SMOTE دقت مدل تقویت گرادیان حداکثری را از ۷۹/۱۸٪ به ۹۱/۷٪ و دقت مدل جنگل تصادفی را از ۸۰/۲۱٪ به ۹۱/۰٪ افزایش داده است. نوآوری اصلی این پژوهش در ترکیب روش‌های یادگیری ماشین با متعادل‌سازی داده و تحلیل عددی اهمیت ویژگی‌هاست. همچنین ویژگی‌های مشترکی چون ویژگی‌های مرتبط با ابعاد هندسی، شاخص‌های مکانی و روشنایی به عنوان شاخص‌های پایدار معرفی شدند که می‌توانند مبنای طراحی سامانه‌های هوشمند کنترل کیفیت در صنعت فولاد باشند. تفاوت قابل توجه در فهرست ویژگی‌های مهم بین الگوریتم‌ها و شرایط داده نشان داد که انتخاب ویژگی در این مسئله امری ثابت نبوده و وابسته به روش یادگیری و نحوه آماده‌سازی داده‌ها است. این رویکرد، علاوه بر ارائه بینش علمی عمیق‌تر در حوزه شناسایی عیوب فولاد، می‌تواند به تصمیم‌گیری دقیق‌تر در انتخاب ویژگی‌های کلیدی در کاربردهای صنعتی مشابه کمک کند.

تاریخ دریافت:
۱۴۰۴/۰۵/۲۹
تاریخ پذیرش:
۱۴۰۴/۰۸/۱۴

کلیدواژه‌ها:
شناسایی ویژگی‌های کلیدی، متعادل‌سازی داده‌ها، عیوب سطحی، صنعت فولاد، یادگیری ماشین

نویسنده‌مسئول:
khodadadi@hormozgan.ac.ir

: 10.22034/ABMIR.2025.23555.1155

E-ISSN: 2821-2037

۱- مقدمه

عیوب سطحی در صفحات فولادی می‌توانند باعث کاهش کیفیت محصول نهایی و در نتیجه کاهش بهره‌وری در فرآیندهای تولید شوند. تشخیص به‌موقع این عیوب نقش مهمی در ارتقاء کنترل کیفیت و بهینه‌سازی هزینه‌ها دارد. با رشد روزافزون داده‌های صنعتی و پیشرفت الگوریتم‌های یادگیری ماشین[1]، امکان تحلیل دقیق‌تر و خودکار عیوب فراهم شده است. این پژوهش تلاش دارد با تمرکز بر شناسایی ویژگی‌های مؤثر از داده‌های مربوط به صفحات فولادی، به طراحی مدلی هوشمند برای طبقه‌بندی نوع عیوب بپردازد.

در سال‌های اخیر، یادگیری ماشین به عنوان یک ابزار قدرتمند در بهبود کیفیت و افزایش بهره‌وری در صنعت فولاد و به‌ویژه در حوزه تشخیص خودکار عیوب سطحی ورق‌های فولادی مطرح شده است. الگوریتم‌های یادگیری ماشین، به‌طور گسترده‌ای برای طبقه‌بندی و تشخیص عیوبی مانند ترک، حباب، اکسیداسیون و ناهمواری سطحی مورداستفاده قرارگرفته‌اند که دارای دقت تشخیص بالایی بوده و زمان بازرسی را به طور چشمگیری کاهش داده‌اند [۱]. ادغام این فناوری‌ها در سیستم‌های بازرسی مبتنی بر بینایی ماشین، نه تنها کاهش خطاهای انسانی را به همراه داشته، بلکه به بهبود کیفیت نهایی محصول و کاهش ضایعات در خط تولید منجر شده است [۲].

تحقیقات و پژوهش در شناسایی خودکار عیوب ورقه‌های فولادی، به طور پیوسته در جریان بوده و هر روز روش‌های موثرتری ارائه می‌شود. در تحقیقات جدید به طور گسترده از روش‌های یادگیری عمیق جهت شناسایی عیوب، استفاده شده است ازجمله در پژوهش سال ۲۰۱۹ [۳]، نویسندگان ضمن معرفی یک مجموعه داده جدید در این حوزه، مدل یادگیری عمیقی به‌نام Reverse Attention، معرفی‌شده است که باعث بهبود تشخیص نقص‌های کوچک می‌شود. پژوهشی دیگر [٤]، به دنبال تشخیص و طبقه‌بندی خودکار نقص‌های سطحی در فولاد نورد گرم است که در این راستا به جای استفاده از مدل‌های شبکه عصبی عمیق CNN از مدل Vision Transformer بهره می‌گیرد و نشان داده می‌شود که استفاده از مدل‌های مبتنی بر ترانسفورمر می‌تواند عملکرد بهتری نسبت به شبکه‌های کانولوشنی سنتی داشته باشد. در مقاله دیگری [۵] در سال ۲۰۲٤، به تشخیص نقص در فولاد با تاکید بر ویژگی‌های لبه[2] پرداخته شده است چرا که نقص‌های فولادی اغلب دارای لبه‌های ضعیف هستند. در این مقاله از روش‌های سنتی استخراج ویژگی لبه مانند Sobel به همراه شبکه‌های عصبی کوچک و سبک برای طبقه‌بندی استفاده شده است. علاوه بر این، استفاده از شبکه عصبی کانولوشنی بهینه‌شده به همراه پیش‌پردازش‌های پردازش تصویر [٦] و استفاده از مکانیزم توجه[3] در معماری شبکه و بهینه‌سازی فرآیند تشخیص چند مقیاسی [۷] نیز جهت تشخیص عیوب سطحی فولاد استفاده شده است. همچنین در پژوهشی دیگری [۸]، چندین روش تشخیص عیوب سطحی فولاد مقایسه شده است و نویسندگان روش‌های مبتنی بر یادگیری عمیق و مجموعه داده‌های استاندارد و نیز معیارهای ارزیابی را مقایسه و تحلیل می‌کنند که می‌تواند به عنوان یک منبع در حوزه کنترل کیفی صنعت فولاد محسوب شود. در رابطه با انتخاب ویژگی‌های برتر و کلیدی در زمینه شناسایی عیوب سطحی فولاد نیز تاکنون کارهای متعددی انجام شده است. در سال ۲۰۲۲ پژوهشی [۹] به ارائه یک روش مبتنی بر یادگیری ترکیبی[4] برای طبقه‌بندی عیوب صفحات فولادی می‌پردازد و چندین روش یادگیری ماشین را در این زمینه مقایسه کرده و درنهایت از ویژگی‌هایی چون ویژگی‌های بافتی و ویژگی‌های شدت نور به عنوان ویژگی‌های برتر و کلیدی در شناسایی عیوب سطحی فولاد نام برده است. همچنین تحقیق دیگری [۱۰] نیز با کمک چندین روش یادگیری ماشین به طبقه‌بندی عیوب فولاد پرداخته و همچنین پیش‌پردازش‌هایی مانند متعادل‌سازی داده‌های آموزشی نیز در نظر گرفته‌شده است و در اینجا نیز ویژگی‌هایی چون مساحت عیب، شکل عیب و شاخص روشنایی عیب به عنوان ویژگی‌های برتر

[1] Machine Learning
[2] Edge Features
[3] Attention Mechanism
[4] Ensemble Learning

شناسایی‌شده است. هر دو پژوهش از داده رایگان Steel Plates Faults[1] [۱۱] استفاده کرده‌اند. دوربنه و همکاران [۱۲]، پوشش جامعی از الگوریتم‌های کلاسیک و دقت آن‌ها را در شناسایی نقص‌های سطحی فولاد ارائه داده‌اند و چندین الگوریتم مورد مقایسه صورت گرفته است؛ اما عدم تحلیل اهمیت ویژگی‌ها و عدم مرتفع سازی عدم تعادل داده‌ها از نقاط ضعف این مقاله است. از دیگر پژوهش‌های انجام شده، مقاله‌ای در سال ۲۰۲۲ توسط فرناندز و همکاران [۱۳] ارائه‌شده است که یک مرور وسیع از کاربرد یادگیری ماشین در تشخیص نقص‌های صنعتی است و در اینجا نشان داده می‌شود که در مسائل عملی نیازمند بررسی‌هایی مثل عدم تعادل داده و مانند آن هستیم.

علی‌رغم پیشرفت‌های قابل توجه در استفاده از روش‌های یادگیری ماشین برای شناسایی عیوب فولاد، چند چالش مهم همچنان باقی است. نخست، توزیع نامتوازن داده‌ها میان کلاس‌های مختلف عیب موجب می‌شود که مدل‌ها به سمت کلاس‌های پرتکرار گرایش پیدا کنند و دقت شناسایی عیوب نادر کاهش یابد. دوم، بسیاری از ویژگی‌های موجود در داده دارای هم‌پوشانی و هم‌بستگی بالا هستند و این امر انتخاب ویژگی‌های مؤثر را دشوار می‌سازد. سوم، مرور پژوهش‌های پیشین نشان می‌دهد که اجماع روشنی در مورد اینکه کدام ویژگی‌ها بیشترین تأثیر را بر دقت مدل دارند وجود ندارد. چهارم، در اغلب مطالعات، تأثیر متوازن‌سازی داده‌ها و انتخاب ویژگی‌ها به صورت عددی و تطبیقی بررسی نشده است. در این پژوهش تلاش شده است تا با ترکیب الگوریتم‌های جنگل تصادفی[2] و تقویت گرادیان حداکثری[3] (XGBoost) و روش متوازن‌سازی داده‌ها، این چالش‌ها به‌صورت نظام‌مند تحلیل و تا حد زیادی برطرف شود.

داده‌های مورداستفاده در این پژوهش از نوع عددی هستند و مقادیر هر ویژگی حاصل پردازش‌های هندسی و نوری از تصاویر خطوط تولید فولاد است. بنابراین هدف پژوهش نه استخراج ویژگی‌های جدید بلکه شناسایی ویژگی‌های کلیدی موجود در داده‌ها برای افزایش دقت مدل است. با توجه به اهمیت موضوع شناسایی ویژگی‌های کلیدی در فرآیند تشخیص عیوب سطحی فولاد و کاستی‌های تحقیقات قبلی، این پژوهش به این موضوع می‌پردازد.

به طور خلاصه نوآوری‌های این پژوهش شامل موارد زیر است:

- شناسایی تفاوت لیست ویژگی‌های مهم در شرایط مختلف: در این پژوهش نشان داده می‌شود که لیست ویژگی‌های مهم وابسته به الگوریتم و روش متعادل‌سازی داده است و ثابت نیست. این موضوع برای حوزه فولاد یک پیام مهم دارد که انتخاب ویژگی‌ها باید متناسب با شرایط داده و مدل انجام شود.
- مقایسه سیستماتیک چند رویکرد انتخاب ویژگی در سه مرحله متفاوت که امکان مقایسه تأثیر متعادل‌سازی داده و الگوریتم انتخاب ویژگی بر نتیجه نهایی را فراهم می‌کند و معرفی مجموعه‌ای از ویژگی‌های پایدار و مشترک میان چند روش مختلف.
- ارائه مقادیر عددی دقیق اهمیت ویژگی‌ها در هر مرحله: در اینجا برعکس بیشتر مقالات که تنها به ارائه نمودار اهمیت ویژگی‌ها بسنده کرده‌اند، برای هر مرحله، جدول عددی اهمیت ویژگی‌ها استخراج‌شده است و مقایسه کمی بین مراحل را ممکن و علمی‌تر می‌کند.
- مقایسه عملکرد مدل‌ها با تمام ویژگی‌ها و با ۱۰ ویژگی برتر و بررسی ترکیب متعادل‌سازی داده‌ها با مدل‌های یادگیری ماشین بر روی داده‌های فولادی.

ادامه مقاله به این صورت سازمان‌دهی شده است: در بخش دوم، الگوریتم‌ها و روش‌های استفاده شده در این مقاله شرح داده می‌شوند. بخش سوم آزمایش‌ها و یافته‌ها مورد بررسی قرار می‌گیرد و مجموعه داده موردنظر جهت آزمایش‌ها معرفی می‌گردد. درنهایت، بخش چهارم به نتیجه‌گیری می‌پردازد.

۲- مواد و روش‌ها

در این مقاله از الگوریتم‌های جنگل تصادفی و تقویت گرادیان حداکثری (XGBoost) جهت طبقه‌بندی عیوب سطحی فولاد

[1] https://archive.ics.uci.edu/dataset/198/steel+plates+faults
[2] Random Forest
[3] Extreme Gradient Boosting (XGBoost)

استفاده شده است و جهت متعادل‌سازی داده‌ها نیز از روش افزایش نمونه‌های مصنوعی اقلیت[1] (SMOTE) استفاده شده است. در ادامه جزئیات این روش‌ها شرح داده‌شده است.

۲-۱ الگوریتم جنگل تصادفی

الگوریتم‌های جنگل تصادفی از قدرتمندترین الگوریتم‌های یادگیری ماشین در حوزه طبقه‌بندی و رگرسیون هستند که از ترکیب تعدادی درخت تصمیم استفاده کرده و ایده «یادگیری تجمیعی» را برای افزایش دقت پیش‌بینی و کاهش بیش‌برازش به کار می‌گیرد.

در این روش، دو مفهوم «نمونه‌برداری بوت استرپ» و انتخاب ویژگی در هر تقسیم نود، مجموعه‌ای از درخت‌های تصمیم متنوع و مستقل تولید می‌شود و نتایج آن‌ها برای تشکیل پیش‌بینی نهایی ادغام می‌شود. نحوه عملکرد این الگوریتم به شرح زیر است [۱۴]:

- ابتدا از مجموعه داده اصلی، چند زیرمجموعه نمونه‌برداری شده با جایگزینی (بوت‌استرپ) ساخته می‌شود.
- هر درخت تصمیم به صورت جداگانه روی یکی از این زیرمجموعه‌ها آموزش داده می‌شود.
- در هر گره تقسیم، تنها بخشی تصادفی از ویژگی‌ها برای پیدا کردن بهترین نقطه تقسیم بررسی می‌شوند، نه همه ویژگی‌ها.
- در انتها، برای مسائل طبقه‌بندی، پیش‌بینی نهایی بر اساس رأی اکثریت درخت‌ها تعیین می‌شود و برای مسائل رگرسیون، میانگین خروجی درخت‌ها گرفته می‌شود.

ازجمله مزایای این الگوریتم، مقاومت بالا در برابر بیش‌برازش[2] نسبت به درخت تصمیم تنها، کارایی مناسب در مسائل با داده‌ها و ویژگی‌های زیاد، قابلیت محاسبه اهمیت هر ویژگی برای تبیین مدل و مناسب برای داده‌های با نویز و داده‌های نامتوازن را می‌توان برشمرد [۱۵].

۲-۲ الگوریتم تقویت گرادیان حداکثری

الگوریتم تقویت گرادیان حداکثری (XGBoost) یکی از روش‌های یادگیری ماشین مدرن است که بر اساس الگوریتم تقویت گرادیان[3] توسعه‌یافته و استفاده بهتری از آن می‌کند و نوعی پیاده‌سازی از درخت تصمیم‌گیری با تقویت شیب است. در اینجا درخت‌های تصمیم به صورت مرحله‌به‌مرحله با هم ترکیب می‌شوند تا یک مدل قوی‌تر ساخته‌شده و در هر مرحله یک مدل جدید با هدف کاهش خطاهای مدل‌های قبلی آموزش داده می‌شود [۱۶].

این روش تأثیر زیادی در بهبود عملکرد مدل‌های یادگیری ماشین داشته و به صورت استاندارد در بسیاری از کاربردها به کار می‌رود. بخشی از کار XGBoost برپایه روش تقویت گرادیان است که ساختار هر مدل جدید بر روی خطاهای مدل‌های قبلی تمرکز می‌کند. این الگوریتم با استفاده از توابع هدف و امکانات بهینه‌سازی کلاس‌های مختلف، مدیریت فرایند آموزش به صورت موازی، مدیریت داده‌های ناقص و استفاده از روش‌های جلوگیری از بیش برازش، قادر به کار با داده‌های کم و کاهش پیچیدگی مدل است. از ویژگی‌های مهم XGBoost می‌توان به موارد زیر اشاره کرد [۱۷]:

- دقت بالا و قابلیت تعمیم‌پذیری خوب به‌واسطه یادگیری مرحله‌به‌مرحله و بهینه سازی دقیق.
- کارایی محاسباتی بالا با استفاده از موازی‌سازی و بهینه‌سازی‌های داخلی.
- مقابله با داده‌های نامتوازن از طریق تنظیم پارامترها.
- امکان ذخیره و بارگذاری سریع مدل‌ها برای کاربردهای عملی.

۲-۳ روش SMOTE

یکی از روش‌های پیشرفته برای مدیریت مشکل عدم تعادل داده‌ها در یادگیری ماشین، به‌ویژه در مسائل طبقه‌بندی، افزایش نمونه‌های مصنوعی اقلیت (SMOTE) است. در بسیاری از مسائل، تعداد نمونه‌ها در کلاس اقلیت بسیار کمتر از کلاس اکثریت است. این عدم تعادل در داده‌ها دقت مدل‌ها را کاهش می‌دهد و باعث تمایل به سمت کلاس اکثریت می‌شود. که این مشکل را با ایجاد نمونه‌های مصنوعی جدید برای کلاس اقلیت کاهش می‌یابد.

[1] Synthetic Minority Oversampling Technique (SMOTE)

[2] Overfitting

[3] Gradient Boosting

اولین معرفی این روش به نوا داتار چلا و همکارانش در سال ۲۰۰۲ نسبت داده می‌شود [۱۸]. این روش عمدتاً به دنبال حل مشکل بیش برازش ناشی از تکرار تصادفی نمونه‌های کلاس اقلیت بود. شیوه کار SMOTE به این صورت است که نمونه‌های جدیدی با استفاده از درون‌یابی خطی بین نمونه‌های کلاس اقلیت و نزدیک‌ترین همسایه‌های آن‌ها در فضای ویژگی ایجاد می‌کند. این امر به افزایش تنوع نمونه‌های کلاس اقلیت کمک کرده و به مدل‌های یادگیری ماشین امکان می‌دهد تا مرزهای دقیق‌تری برای طبقه‌بندی ایجاد کنند. در این روش سه مرحله زیر انجام می‌شود:

۱- برای هر نمونه از کلاس اقلیت، kنزدیک‌ترین همسایه آن در فضای ویژگی‌ها شناسایی می‌شود

۲- نمونه‌های مصنوعی جدید با ترکیب ویژگی‌های نمونه اصلی و یکی از همسایه‌های نزدیک آن به صورت تصادفی بر روی خط بین این دو نمونه ایجاد می‌شوند.

۳- این فرآیند تا جایی ادامه می‌یابد که تعداد نمونه‌های کلاس اقلیت به میزان دلخواه یا متناسب با کلاس اکثریت برسد.

این رویکرد برخلاف روش‌های ساده نمونه‌برداری، باعث افزایش تنوع نمونه‌های اقلیت بدون تکرار مستقیم می‌شود و به بهبود عملکرد مدل‌ها به‌خصوص در داده‌های نامتوازن کمک می‌کند [۱۸].

۳- آزمایش‌ها و یافته‌ها

۳-۱ مجموعه‌داده

در این پژوهش از مجموعه داده Steel Plates Faults [۱۱] که یک داده معتبر در حوزه تشخیص عیوب سطحی فولاد است، بهره گرفته‌شده است. در جدول ۱ نام ویژگی‌ها و معادل فارسی آن را مشاهده می‌کنید. این مجموعه شامل ۱۹۴۱ نمونه از صفحات فولادی است که برای هر نمونه ۲۷ ویژگی عددی و کیفی اندازه‌گیری شده و هدف آن تشخیص نوع عیب سطحی در صفحات فولاد است. داده‌ها در هفت کلاس مختلف برچسب‌گذاری شده‌اند که بیانگر انواع عیوب رایج در فرآیند تولید فولاد هستند.

در این مجموعه داده علاوه بر ویژگی‌های هندسی و نوری پایه نظیر مساحت، محیط و شاخص‌های مکانی، تعدادی ویژگی مشتق‌شده نیز وجود دارد که توسط تهیه‌کنندگان داده طراحی شده‌اند. مانند SigmoidOfAreas و SigmoidOfLuminosity که حاصل اعمال تابع سیگموید بر مقادیر خام هستند تا مقیاس داده‌ها به بازه بین ۰ و ۱ محدود شده و اثر مقادیر پرت کاهش یابد. همچنین، برخی ویژگی‌ها مانند LogOfAreas و Log_X_Index حاصل تبدیل لگاریتمی هستند که با هدف کاهش چولگی توزیع داده و افزایش تفکیک‌پذیری بین نمونه‌ها ایجاد شده‌اند. علاوه بر این، ویژگی‌هایی نظیر Square_Index یا Orientation_Index بیانگر ترکیب‌های هندسی خاصی از متغیرهای اولیه‌اند که اطلاعات ساختاری بیشتری درباره عیوب سطحی فولاد ارائه می‌کنند.

جدول (۱): ویژگی‌های مجموعه‌داده Steel Plates Faults

ردیف	نام ویژگی	معادل فارسی
۱	X_Minimum	حداقل مختصات X
۲	X_Maximum	حداکثر مختصات X
۳	Y_Minimum	حداقل مختصات Y
٤	Y_Maximum	حداکثر مختصات Y
٥	Pixels_Areas	تعداد پیکسل‌های ناحیه
٦	X_Perimeter	محیط در راستای X
۷	Y_Perimeter	محیط در راستای Y
۸	Sum_of_Luminosity	مجموع روشنایی
۹	Minimum_of_Luminosity	حداقل روشنایی
۱۰	Maximum_of_Luminosity	حداکثر روشنایی
۱۱	Length_of_Conveyer	طول نوار نقاله
۱۲	TypeOfSteel_A300	نوع فولاد A300
۱۳	TypeOfSteel_A400	نوع فولاد A400
۱٤	Steel_Plate_Thickness	ضخامت صفحه فولادی
۱٥	Edges_Index	شاخص لبه‌ها
۱٦	Empty_Index	شاخص فضای خالی
۱۷	Square_Index	شاخص مربعی
۱۸	Outside_X_Index	شاخص بیرونی X
۱۹	Edges_X_Index	شاخص لبه در راستای X
۲۰	Edges_Y_Index	شاخص لبه در راستای Y
۲۱	Outside_Global_Index	شاخص کلی بیرونی
۲۲	LogOfAreas	لگاریتم مساحت ناحیه
۲۳	Log_X_Index	لگاریتم شاخص X
۲٤	Orientation_Index	شاخص جهت‌گیری
۲٥	Luminosity_Index	شاخص روشنایی
۲٦	SigmoidOfAreas	سیگموید مساحت‌ها
۲۷	SigmoidOfLuminosity	سیگموید روشنایی

استفاده از این نوع ویژگی‌های مهندسی‌شده در کنار متغیرهای پایه، امکان تحلیل دقیق‌تر و مدل‌سازی مؤثرتر عیوب را فراهم می‌سازد. همان‌طور که بیان شد، در این مجموعه داده، داده‌ها عددی و شامل ویژگی‌های هندسی و نوری استخراج‌شده از تصاویر هستند که در مجموعه داده موردنظر وجود دارد؛ ازاین‌رو نیازی به استخراج ویژگی جدید نبوده و تمرکز مقاله بر انتخاب ویژگی‌های کلیدی بوده است. به همین دلیل، در این پژوهش از الگوریتم‌های یادگیری ماشین کلاسیک استفاده شده است، چرا که این نوع داده‌ها برای مدل‌های شبکه عصبی عمیق مناسب نیستند و یادگیری عمیق عموماً برای داده‌های تصویری خام کاربرد دارد.

در جدول ۲ اسامی نقص‌ها و تعداد رکورد مربوط به هر نقص نشان داده‌شده است. همان‌گونه که از اطلاعات جدول ۲ مشخص است، این مجموعه داده یک مجموعه داده نامتوازن است که ممکن است موجب سوگیری مدل به سمت یادگیری کلاس‌های غالب شود.

جدول (۲): کلاس‌های خروجی (انواع عیوب سطح فولاد)

ردیف	نام عیب	معادل فارسی	تعداد رکورد
۱	Pastry	خمیر مانند (پوسته‌ای)	۱۵۸
۲	Z_Scratch	خط و خش نوع Z	۱۹۰
۳	K_Scatch	خط و خش نوع K	۳۹۱
٤	Stains	لکه‌ها	۷۲
٥	Dirtiness	کثیفی‌ها	٥٥
٦	Bumps	برجستگی‌ها	٤۰۲
۷	Other_Faults	سایر عیوب	٦۷۳

۳-۲ ارزیابی عملکرد مدل‌ها با داده خام

ازآنجاکه بسیاری از مدل‌های یادگیری ماشین از داده‌هایی با برچسب کلاس عددی استفاده می‌کنند، بنابراین قبل از استفاده از این مجموعه داده، ابتدا برچسب‌های رشته‌ای کلاس‌ها را به صورت اعداد ۰ تا ٦ کدگذاری می‌کنیم؛ یعنی نام کلاس هدف مثل Stains, Bumps و غیره به صورت اعداد ۰ تا ٦ تنظیم می‌شود. همچنین در بعضی از نسخه‌های این داده، برچسب کلاس‌ها به صورت ۷ ستون جدا تعبیه شده است به این معنی که در هر ردیف پس از مقادیر ویژگی‌ها، ۷ ستون کلاس اضافه‌شده است که مقادیر همه کلاس‌ها ۰ و فقط کلاس موردنظر ۱ است؛ بنابراین این مورد هم در مجموعه داده، اصلاح‌شده و ۷ ستون به یک ستون با مقادیر ۰ تا ٦ که نماینده یکی از کلاس‌های ذکرشده است، تبدیل می‌شود.

در گام نخست، کلیه تحلیل‌ها و مدل‌سازی‌ها بر روی داده‌های خام انجام شد، بدون آن‌که هیچ‌گونه پیش‌پردازش یا روش‌های افزایش داده به کار گرفته شود. هدف از این مرحله، بررسی توانایی مدل‌های یادگیری ماشین در شناسایی الگوها و عیوب سطحی فولاد بر اساس تمامی ویژگی‌های اولیه بود.

در مرحله ابتدایی پژوهش، به منظور مقایسه عملکرد مدل‌های مختلف یادگیری ماشین، چهار الگوریتم پایه شامل جنگل تصادفی، XGBoost، ماشین بردار پشتیبان[1] و K- نزدیک‌ترین همسایه[2] مورد آزمون بر روی داده خام قرار گرفتند. نتایج اولیه نشان داد که مدل‌های جنگل تصادفی و XGBoost با دقت‌های حدود ۸۰٪، عملکرد بهتری نسبت به مدل‌های SVM و KNN (با دقت کمتر از ٤۰٪) دارند. بر همین اساس، در مراحل بعدی تمرکز بر دو مدل برتر قرار گرفت تا تحلیل دقیق‌تری از اهمیت ویژگی‌ها و نقش متوازن‌سازی داده‌ها ارائه شود.

برای استخراج مهم‌ترین ویژگی‌ها، از الگوریتم‌های جنگل تصادفی و همچنین XGBoost بهره گرفته شد. ازآنجاکه هر کدام از این دو الگوریتم نماینده دو گروه از الگوریتم‌های ترکیبی هستند، از آن‌ها استفاده شد؛ جنگل تصادفی نماینده روش‌های نمونه‌برداری تجمعی[3] و XGBoost نماینده روش‌های تقویت‌کننده[4] هستند. به‌دلیل توانایی هر دو روش در مدل‌سازی داده‌ها و درعین‌حال امکان استخراج اهمیت ویژگی‌ها مورداستفاده قرار گرفتند. هر یک از این الگوریتم‌ها، پس از آموزش بر روی داده‌ها، وزن یا اهمیتی به هر ویژگی اختصاص می‌دهند که بیانگر سهم آن ویژگی در فرآیند تفکیک کلاس‌ها است. به عنوان مثال، درXGBoost این اهمیت‌ها بر اساس کاهش خطای طبقه‌بندی ناشی از استفاده هر ویژگی محاسبه می‌گردد و در جنگل تصادفی این مقادیر با معیارهایی نظیر کاهش ناخالصی جینی[5] یا کاهش خطا در تقسیم‌بندی داده‌ها تعیین می‌شوند. سپس ویژگی‌ها بر اساس مقادیر

[1] Support Vector Machine (SVM)
[2] K-Nearest Neighbors (KNN)
[3] Bagging
[4] Boosting
[5] Gini Impurity Decrease

اهمیت مرتب شدند و ۱۰ ویژگی برتر با بیشترین نقش در بهبود عملکرد مدل انتخاب گردیدند. این فرآیند در حقیقت نوعی انتخاب ویژگی مبتنی بر مدل محسوب می‌شود که ضمن کاهش ابعاد داده، موجب ساده‌تر شدن مدل و کاهش احتمال بیش‌برازش می‌گردد. علاوه بر شناسایی ده ویژگی برتر، مقادیر عددی اهمیت هر ویژگی نیز محاسبه گردید. این مقادیر بیانگر سهم نسبی هر ویژگی در کاهش خطا یا افزایش دقت مدل هستند. به عنوان مثال در مدل جنگل تصادفی ویژگی «LogOfAreas» با اهمیت ۰,۰۷۳ و در مدلXGBoost ویژگی «Log_X_Index» با اهمیت ۰,۳۷۳ به عنوان شاخص‌های کلیدی معرفی شدند. وجود این مقادیر عددی، امکان مقایسه عینی نقش هر ویژگی در مدل را فراهم کرده و نشان می‌دهد کدام متغیرها بیشترین تأثیر را در تمایز بین کلاس‌های عیوب فولادی دارند.

پس از شناسایی ویژگی‌های مهم، دو سناریوی مجزا تعریف شد:

۱- آموزش مدل‌ها با تمامی ۲۷ ویژگی اصلی

۲- آموزش مدل‌ها تنها با ۱۰ ویژگی منتخب

برای هر دو حالت، ابتدا داده‌ها به قسمت‌های آموزش (۸۰ درصد) و آزمایش (۲۰ درصد) تقسیم شد و پس از یادگیری مدل‌ها بر روی داده آموزش، عملکرد مدل‌ها بر اساس دقت طبقه‌بندی[1] (نسبت تعداد نمونه‌هایی که به طور صحیح طبقه‌بندی شده‌اند به تعداد کل نمونه‌ها) بر روی داده آزمایش، مقایسه شد. این مقایسه امکان ارزیابی تأثیر انتخاب ویژگی‌ها بر روی کیفیت پیش‌بینی را فراهم ساخت. در جدول ۳ دقت مدل‌ها و در جدول ٤ و ٥ ویژگی‌های برتر به‌دست آمده در هر مرحله به همراه مقادیر عددی که بیانگر میزان تأثیر هر ویژگی است، نشان داده‌شده است.

جدول (۳): دقت مدل‌ها بر روی داده خام

نام مدل	دقت مدل (درصد)
جنگل تصادفی با تمامی ویژگی‌ها	۸۰/۲۱
جنگل تصادفی با ۱۰ ویژگی برتر	۷۷/۳۸
XGBoost با تمامی ویژگی‌ها	۷۹/۱۸
XGBoost با ۱۰ ویژگی برتر	۷٦/۸٦

جدول (٤): ویژگی‌های برتر به‌دست آمده به همراه مقادیر عددی اهمیت هر ویژگی بر روی داده‌های خام با استفاده از مدل جنگل تصادفی

ردیف	نام ویژگی	مقدار عددی اهمیت ویژگی
۱	LogOfAreas	۰/۰۷۳۲٦۳
۲	Length_of_Conveyer	۰/۰٦۱۱٤٥
۳	Pixels_Areas	۰/۰٥٦٥٦۱
٤	Log_X_Index	۰/۰٥۰۰۱٥
٥	Outside_X_Index	۰/۰٤۷۰٦٦
٦	Steel_Plate_Thickness	۰/۰٤٦۹۱۹
۷	X_Minimum	۰/۰٤٤٤۹٦
۸	Sum_of_Luminosity	۰/۰٤۲۹۰۹
۹	Minimum_of_Luminosity	۰/۰٤۱٦۳٤
۱۰	Edges_Index	۰/۰۳۸۱٥۷

جدول (٥): ویژگی‌های برتر به‌دست آمده به همراه مقادیر عددی اهمیت هر ویژگی بر روی داده‌های خام با استفاده از مدل XGBoost

ردیف	نام ویژگی	مقدار عددی اهمیت ویژگی
۱	Log_X_Index	۰/۳۷۳۲۰٦
۲	TypeOfSteel_A300	۰/۱۳۸٥۹٤
۳	Pixels_Areas	۰/۰٦۹۲۸۲
٤	Steel_Plate_Thickness	۰/۰٥۲۹٥۹
٥	Y_Perimeter	۰/۰٤٦۰٦۸
٦	Orientation_Index	۰/۰٤۰۰٦۰
۷	Length_of_Conveyer	۰/۰۳٤۷٤٤
۸	X_Minimum	۰/۰۲۳۹٥۲
۹	Minimum_of_Luminosity	۰/۰۲۲۲٦۳
۱۰	Square_Index	۰/۰۱۹۹۷۲

نتایج نشان داد که اگرچه دقت مدل‌ها با تمامی ویژگی‌ها اندکی بیشتر بود، اما عملکرد مدل‌ها با ۱۰ ویژگی برتر نیز در سطح مطلوب باقی ماند. این یافته حاکی از آن است که مجموعه‌ای کوچک از ویژگی‌های کلیدی قادر است بخش عمده‌ای از اطلاعات

[1] Accuracy

لازم برای شناسایی عیوب فولادی را فراهم آورد و استفاده از تمامی ویژگی‌ها لزوماً ضروری نیست.

۳-۳ ارزیابی عملکرد مدل‌ها با داده متوازن شده

یکی از چالش‌های اصلی در مجموعه‌داده مورداستفاده، عدم تعادل کلاس‌ها بود؛ به گونه‌ای که برخی از انواع عیوب دارای تعداد نمونه بسیار بیشتری در مقایسه با سایر دسته‌ها بودند. این موضوع می‌تواند منجر به ایجاد سوگیری مدل‌های یادگیری ماشین به سمت کلاس‌های پرتکرار شود و در نتیجه دقت پیش‌بینی برای کلاس‌های کم‌تکرار کاهش یابد. برای رفع این مشکل، در این پژوهش از روش SMOTE استفاده گردید. در این روش، داده‌های مصنوعی برای کلاس‌های اقلیت ایجاد می‌شوند و بدین ترتیب توزیع داده‌ها متوازن می‌شود.

پس از متوازن‌سازی داده‌ها با SMOTE، فرآیند مشابه مرحله اول تکرار گردید. بدین صورت که ابتدا مدل‌های جنگل تصادفی و XGBoost بر روی داده‌های متوازن‌شده آموزش داده شدند. سپس برای شناسایی مهم‌ترین متغیرهای مؤثر در پیش‌بینی نوع عیب، از معیارهای اهمیت ویژگی هر مدل استفاده شد. در فرآیند متوازن‌سازی، داده‌های تولیدشده توسط SMOTE صرفاً در مجموعه آموزش استفاده شده‌اند تا از ایجاد دقت غیرواقعی در ارزیابی جلوگیری شود.

جدول (٦): دقت مدل‌ها بر روی داده متوازن شده

نام مدل	دقت مدل (درصد)
جنگل تصادفی با تمامی ویژگی‌ها	۹۰/۹۹
جنگل تصادفی با ۱۰ ویژگی برتر	۸۹/۵۰
XGBoost با تمامی ویژگی‌ها	۹۱/۷۳
XGBoost با ۱۰ ویژگی برتر	۸۹/٦۱

اهمیت ویژگی‌ها در هر مدل به صورت عددی بیان می‌شود؛ این مقادیر عددی نشان‌دهنده سهم نسبی هر ویژگی در عملکرد مدل هستند و امکان رتبه‌بندی آن‌ها را فراهم می‌کنند. درنهایت، ۱۰ ویژگی برتر با بالاترین اهمیت در هر مدل انتخاب‌شده و نتایج دقت مدل‌ها با همه ویژگی‌ها و با این ۱۰ ویژگی برتر مورد مقایسه قرار گرفت. در جدول ٦ دقت مدل‌ها و در جدول ۷ و ۸ ویژگی‌های برتر به‌دست آمده در هر مرحله به همراه مقادیر عددی که بیانگر میزان تأثیر هر ویژگی است، نشان داده‌شده است.

جدول (۷): ویژگی‌های برتر به‌دست آمده به همراه مقادیر عددی اهمیت هر ویژگی بر روی داده‌های متوازن شده با استفاده از مدل جنگل تصادفی

ردیف	نام ویژگی	مقدار عددی اهمیت ویژگی
۱	Length_of_Conveyer	۰/۰۷۸۱۸۷
۲	LogOfAreas	۰/۰٦۷۸۱۹
۳	Pixels_Areas	۰/۰۵۷۸۷۵
٤	Edges_Index	۰/۰٤۸۷۷۷
۵	Log_X_Index	۰/۰٤۷۳۰۹
٦	Outside_X_Index	۰/۰٤٦۹٤۷
۷	X_Minimum	۰/۰٤۵۵۹۹
۸	Orientation_Index	۰/۰٤۲۸۸۷
۹	Sum_of_Luminosity	۰/۰٤۲۷۲۰
۱۰	Steel_Plate_Thickness	۰/۰٤۱۱٦۳

جدول (۸): ویژگی‌های برتر به‌دست آمده به همراه مقادیر عددی اهمیت هر ویژگی بر روی داده‌های متوازن شده با استفاده از مدل XGBoost

ردیف	نام ویژگی	مقدار عددی اهمیت ویژگی
۱	TypeOfSteel_A400	۰/۲۵۰۳۲۲
۲	LogOfAreas	۰/۲۰۱۲۸۲
۳	Outside_X_Index	۰/۰۹۸۸۲٤
٤	Orientation_Index	۰/۰۵۲۸۱۲
۵	Length_of_Conveyer	۰/۰۵۱۰۷٦
٦	Square_Index	۰/۰۳۵۳۹٦
۷	Steel_Plate_Thickness	۰/۰۳۰۱۱٤
۸	X_Minimum	۰/۰۲۵۲۸۰
۹	Minimum_of_Luminosity	۰/۰۲۳٦۰۸
۱۰	X_Maximum	۰/۰۲۳۱۱۷

برای اطمینان از پایداری مدل‌ها، روش اعتبارسنجی متقابل پنج‌تایی[1] نیز به‌عنوان مبنای ارزیابی تئوریک در نظر گرفته شد و نتایج حاصل میانگین این عملکرد مدل‌ها است.

نتایج این مرحله نشان داد که اعمال روش SMOTE موجب افزایش قابل توجه دقت مدل‌ها گردیده و علاوه بر آن، ویژگی‌های کلیدی‌تر با وضوح بیشتری شناسایی شدند. این امر تأیید می‌کند که متوازن‌سازی داده‌ها پیش‌نیاز مهمی برای تحلیل‌های یادگیری ماشین در داده‌های صنعتی با کلاس‌های نامتوازن است.

[1] 5-Fold Cross Validation

۴-۳ بحث و نتایج

یافته‌های حاصل از مراحل مختلف این پژوهش نشان داد که انتخاب ویژگی و متوازن‌سازی داده‌ها نقش تعیین‌کننده‌ای در بهبود عملکرد مدل‌های یادگیری ماشین برای طبقه‌بندی عیوب صفحات فولادی دارند. در مرحله اول، آموزش مدل‌ها بر روی داده‌های خام صورت گرفت. نتایج نشان داد که با وجود دستیابی به دقت نسبتاً مناسب، همچنان برخی کلاس‌ها به دلیل عدم تعادل داده‌ها با دقت پایینی شناسایی شدند. در این مرحله ویژگی‌هایی همچون LogOfAreas، Length_of_Conveyer، Pixels_Areas، Log_X_Index و Steel_Plate_Thickness به طور مکرر در میان متغیرهای مهم مشاهده شدند که نشان‌دهنده نقش اساسی آن‌ها در شناسایی الگوهای مربوط به عیوب است.

در مرحله دوم، با استفاده از روش SMOTE داده‌ها متوازن شدند. نتایج حاکی از آن بود که دقت مدل‌ها به شکل محسوسی افزایش یافت (بیش از ده درصد) و نشان‌دهنده اهمیت متوازن‌سازی داده‌ها در کاربردهای صنعتی است. علاوه بر بهبود دقت کلی، اهمیت ویژگی‌ها نیز با وضوح بیشتری مشخص گردید. ویژگی‌هایی نظیر LogOfAreas، TypeOfSteel_A400، Length_of_Conveyer، Orientation_Index و Outside_X_Index در این مرحله اهمیت بالاتری یافتند.

تعدادی از ویژگی‌ها تقریباً در تمام نتایج به‌عنوان ده ویژگی برتر ظاهر شدند:

- LogOfAreas (لگاریتم مساحت‌ها)
- Length_of_Conveyer (طول نوار نقاله)
- Log_X_Index (لگاریتم شاخص X)
- Steel_Plate_Thickness (ضخامت صفحه فولادی)
- Luminosity (روشنایی)
- Pixels_Areas (تعداد پیکسل‌های ناحیه)
- Outside_X_Index (شاخص بیرونی X)
- X_Minimum (حداقل مختصات X)

ازآنجاکه این ویژگی‌ها در چندین مدل و شرایط اهمیت بالایی دارند، بنابراین این ویژگی‌ها به احتمال زیاد واقعاً یک عامل بنیادی در ایجاد یا شناسایی عیوب فولادی هستند و می‌توانند به‌عنوان متغیرهای کلیدی برای کنترل کیفیت هوشمند به‌کار گرفته شوند. این تکرار بیانگر پایداری و اهمیت بنیادی این ویژگی‌ها در شناسایی الگوهای عیوب فولادی است، بنابراین به جای استفاده از همه ویژگی‌ها، می‌توان فقط این ویژگی‌های مشترک را مبنای مدل‌سازی قرار داد. این کار باعث کاهش هزینه محاسباتی و افزایش قابلیت تفسیر مدل می‌شود. ویژگی‌های مرتبط با ابعاد هندسی شامل LogOfAreas، Steel_Plate_Thickness، Length_of_Conveyer و X_Minumium نشان می‌دهد که اندازه و ضخامت صفحه فولادی تأثیر مستقیم در احتمال بروز عیب دارد. همچنین ویژگی‌های مرتبط با شاخص‌های مکانی و روشنایی شامل Outside_X_Index، Minimum_of_Luminosity، Log_X_Index، Sum_of_Luminosity و Pixels_Areas بیانگر آن است که مکان قرارگیری عیب و شدت روشنایی تصویر نقش مهمی در شناسایی دقیق عیوب دارند.

یکی از یافته‌های مهم این پژوهش آن است که روش‌های مختلف انتخاب ویژگی، حتی در شرایط مشابه داده‌ای، مجموعه‌های متفاوتی از ویژگی‌های مهم را معرفی می‌کنند. این اختلاف ناشی از تفاوت در ماهیت مدل‌ها و معیارهای ارزیابی اهمیت ویژگی‌هاست؛ به‌طور مثال، جنگل تصادفی بر اساس کاهش خطای جینی یا آنتروپی تصمیم‌گیری می‌کند، در حالی که XGBoost با بهره‌گیری از تقویت گرادیانی، سهم هر ویژگی در بهبود تابع هدف را به‌عنوان معیار اهمیت در نظر می‌گیرد. پیامد این موضوع آن است که انتخاب ویژگی‌های برتر، نسبی بوده و کاملاً وابسته به الگوریتم مورداستفاده است. بنابراین، تکیه بر یک روش منفرد ممکن است باعث از دست رفتن برخی ابعاد مهم داده شود. ازاین‌رو، رویکرد ترکیبی یا مقایسه‌ای میان چند الگوریتم می‌تواند دید جامع‌تری نسبت به عوامل کلیدی مؤثر بر کیفیت سطح فولاد ارائه دهد. این نتیجه همچنین بیانگر آن است که تصمیم‌گیری نهایی در کاربردهای صنعتی (مانند پایش کیفیت و تشخیص عیوب) نباید صرفاً بر اساس خروجی یک مدل صورت گیرد، بلکه نیازمند تحلیل چندجانبه و هم‌پوشانی میان ویژگی‌های منتخب الگوریتم‌های مختلف است.

٤- نتیجه‌گیری و پیشنهادها

در این پژوهش، یک چارچوب سه‌مرحله‌ای برای شناسایی ویژگی‌های کلیدی و بهینه‌سازی مدل‌های طبقه‌بندی در تشخیص عیوب صفحات فولادی ارائه گردید. در گام نخست، با استفاده از الگوریتم جنگل تصادفی و تقویت گرادیان حداکثری بر روی داده‌های خام، ۱۰ ویژگی برتر به ترتیب اهمیت شناسایی و اهمیت عددی هر یک محاسبه شد. نتایج نشان داد که با استفاده از تمامی ویژگی‌ها، دقت مدل‌ها به‌ترتیب ۸۰/۲۱٪ و ۷۹/۱۸٪ بوده و با کاهش ویژگی‌ها به ۱۰ مورد برتر، دقت به ۷۷/۳۸٪ و ۷٦/۸٦٪ کاهش یافت. در گام دوم، اثر متعادل‌سازی داده‌ها با روش SMOTE بر انتخاب ویژگی‌ها و عملکرد مدل‌ها بررسی گردید. در این مرحله، با استفاده از الگوریتم XGBoost + SMOTE دقت مدل با تمام ویژگی‌ها به ۹۱/۷۳٪ و با ۱۰ ویژگی برتر به ۸۹/٦۱٪ رسید. همچنین، الگوریتم Random Forest + SMOTE نیز عملکردی مشابه داشته و دقت ۹۰/۹۹٪ با تمام ویژگی‌ها و ۸۹/۵۰٪ با ۱۰ ویژگی برتر را به دست آورد. مقایسه سه مرحله نشان داد که لیست ویژگی‌های برتر وابسته به نوع الگوریتم و فرآیند پیش‌پردازش داده‌ها است و این مسئله اهمیت تحلیل چندجانبه را در مسائل صنعتی مشابه برجسته می‌کند. همچنین، استفاده از SMOTE باعث بهبود چشمگیر دقت پیش‌بینی در هر دو مدل گردید که بیانگر اهمیت توازن داده در مسائل طبقه‌بندی با توزیع نامتوازن است. در این پژوهش، ویژگی‌های مرتبط با ابعاد هندسی، شاخص‌های مکانی و روشنایی به عنوان ویژگی‌های کلیدی در شناسایی عیوب سطحی صفحات فولاد، انتخاب گردید.

به عنوان پیشنهادی برای کارهای آینده، می‌توان از روش‌های پیشرفته و ترکیبی برای انتخاب ویژگی استفاده کرد و تأثیر نرمال‌سازی و استانداردسازی پیشرفته داده‌ها در شرایط مختلف را بر بهبود عملکرد مدل‌ها در این حوزه بررسی کرد. همچنین پیاده‌سازی چارچوب توسعه‌یافته بر روی داده‌های صنعتی واقعی و بررسی قابلیت تعمیم‌پذیری مدل، از دیگر پیشنهادها است.

References

[1] Y. Zhang, S. Shen, and S. Xu, “Strip steel surface defect detection based on lightweight YOLOv5,” Frontiers in Neurorobotics, vol. 17, Oct. 2023, doi: https://doi.org/10.3389/fnbot.2023.1263739.

[2] Y. Zhan and F. Feng, "Detection and Identification of Strip Surface Defects Based on Deep Learning," 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Guangzhou, China, 2022, pp. 402-405, doi: 10.1109/MLISE57402.2022.00086.

[3] X. Lv, F. Duan, J. Jiang, X. Fu, and L. Gan, “Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network,” Sensors, vol. 20, no. 6, p. 1562, Mar. 2020, doi: https://doi.org/10.3390/s20061562.

[4] V. Vasan, N. V. Sridharan, Sugumaran Vaithiyanathan, and Mohammadreza Aghaei, “Detection and Classification of Surface Defects on Hot-Rolled Steel using Vision Transformers,” Heliyon, vol. 10, no. 19, pp. e38498–e38498, Oct. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e38498.

[5] Y. Wang et al., “A steel defect detection method based on edge feature extraction via the Sobel operator,” Scientific Reports, vol. 14, no. 1, Nov. 2024, doi: https://doi.org/10.1038/s41598-024-79205-5.

[6] O. C. Han and U. Kutbay, “Detection of Defects on Metal Surfaces Based on Deep Learning,” Applied Sciences, vol. 15, no. 3, p. 1406, Jan. 2025, doi: https://doi.org/10.3390/app15031406.

[7] Y. Jiang, “Surface defect detection of steel based on improved YOLOv5 algorithm,” Mathematical Biosciences and Engineering, vol. 20, no. 11, pp. 19858–19870, Jan. 2023, doi: https://doi.org/10.3934/mbe.2023879.

[8] A. Ashwin, Edmond, B. Gao, and Wai Lok Woo, “A Review and Benchmark on State-of-the-Art Steel Defects Detection,” SN Computer Science, vol. 5, no. 1, Dec. 2023, doi: https://doi.org/10.1007/s42979-023-02436-2.

[9] E. C. ÖZKAT, “A method to classify steel plate faults based on ensemble learning,” Journal of Materials and Mechatronics: A, vol. 3, no. 2, Oct. 2022, doi: https://doi.org/10.55546/jmm.1161542.

[10] A. Kharal, "Explainable artificial intelligence based fault diagnosis and insight harvesting for steel plates manufacturing," arXiv preprint arXiv:2008.04448, 2020.

[11] M. Buscema, S. Terzi, and W. Tastle, "Steel Plates Faults," UCI Machine Learning Repository, 2010. [Online]. Available: https://doi.org/10.24432/C5J88N. [Accessed: Nov. 13, 2025].

[12] A. Dorbane, F. Harrou, and Y. Sun, "Detecting Faulty Steel Plates Using Machine Learning," in Advances in Computing and Data Sciences, S. Silakari, P. P, S. S, and B. Panda, Eds., Cham, Switzerland: Springer Nature, 2024, pp. 321–333. doi: 10.1007/978-3-031-70906-7_27. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-70906-7_27.

[13] M. Fernandes, J. M. Corchado, and G. Marreiros, “Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review,” Applied Intelligence, vol. 52, no. 12, Mar. 2022, doi: https://doi.org/10.1007/s10489-022-03344-3.

[14] Sadegh Mosharrafzadeh, Bahman Ravaei, and Ehsanollah Koozegar, “Diagnosis of Diabetes Using a Random Forest Algorithm,” JOURNAL OF DIABETES AND METABOLIC DISORDERS, vol. 21, no. 2, pp. 92–100, 2021, [Online]. Available: https://sid.ir/paper/415460/en.

[15] N. Venkatesan and G. Priya, "A Study of Random Forest Algorithm with Implementation Using WEKA," International Journal of Innovative Research in Computer Science and Engineering, vol. 1, no. 6, pp. 156–162, 2015. [Online]. Available:https://www.ioirp.com/Doc/IJIRCSE/i6/JCSE242.pdf.

[16] S. Bakhtiari, " Malware Detection Using Data Mining and Xgboost and Random Forest Algorithms," Journal of Information and communication Technology in policing (JICTP), vol. 3, no. 1, pp. 55-68, Apr. 2020, doi:https://doi.org/10.22034/pitc.2022.1267935.1130.

[17] T. Chen and C. Guestrin, “XGBoost: a Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, vol. 1, no. 1, pp. 785–794, Aug. 2016, doi: https://doi.org/10.1145/2939672.2939785.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, no. 16, pp. 321–357, Jun. 2002, doi: https://doi.org/10.1613/jair.953.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 3, No. 2, Autumn & Winter 2025, pp. 175 - 186

Analysis and Identification of Key Features for Classifying Surface Defects in Steel Plates Using Machine Learning Models and Data Balancing Techniques

Habib Khodadadi[1*], Seyed Ali Mehri[2], Mahsa Khodadadi[3]

[1]Assistant Professor, Department of Electrical and Computer Engineering, University of Hormozgan, Bandar Abbas, Hormozgan, Iran

[2] B.Sc. Graduate, Department of Computer, Islamic Azad University, Bandar Abbas, Iran

[3] Lecturer, Department of Materials and Metallurgy, Islamic Azad University, Minab, Iran

Article Information

Original Research Paper

Received:
2025 August 20

Accepted:
2025 November 05

Keywords:
Data Balancing, Key Feature Identification, Machine Learning, Steel Industry, Surface Defects

Corresponding Author*:
khodadadi@hormozgan.ac.ir

Abstract

In this study, the main challenges in detecting surface defects of steel sheets—including data imbalance, feature overlap, and the difficulty of identifying influential features—were investigated. The primary objective of the research was to identify the most effective features for improving the accuracy of defect classification models such as scratch, stain, and bump detection. To achieve this, a three-stage analysis was conducted. The first and second stages involved applying the Random Forest and Extreme Gradient Boosting (XGBoost) algorithms to the raw data, while the third stage applied the same algorithms to data balanced using the SMOTE technique. The results showed that employing SMOTE increased the accuracy of the XGBoost model from 79.18% to 91.7%, and that of the Random Forest model from 80.21% to 91.0%. The main innovation of this study lies in integrating machine learning methods with data balancing and numerical feature-importance analysis. Furthermore, common features such as geometric dimension parameters, spatial indices, and brightness characteristics were identified as stable indicators that can serve as the foundation for designing intelligent quality-control systems in the steel industry. The considerable variation in the list of important features across algorithms and data conditions demonstrated that feature selection in this problem is not fixed but depends on the learning method and data preprocessing approach. This methodology not only provides deeper scientific insights into steel defect detection but also supports more precise decision-making in selecting key features for similar industrial applications.

: 10.22034/ABMIR.2025.23555.1155

E-ISSN: 2821-2037