Ind. Manag. Persp. 2025, 15(3), 170-194; doi: 10.48308/JIMP.15.3.170

Industrial Management Perspective

Journal hompage: https://jimp.sbu.ac.ir/?lang=en

Original Article

Intelligent Anomaly Detection in Unbalanced Industrial Data Using the Xgboost Model and Genetic Algorithm (GA) To Optimize Performance in Identifying Defective Products in the Production Line

Rasoul Nematnia*
Maryam Khademi**
Kiamars Fathi***
Soheila Sardar****

Extended Abstract

Introduction and Objectives: The process of production lines and their sequence is one of the fundamental approaches in planning industrial products in bulk. Lack of proper planning in lines and suitable solutions for optimizing effective systems in the production and assembly process leads to increased time allocated to production, increased machine downtime, and consequently a decrease in the number of products produced in terms of quantity and production rate. Inefficiency of allocated resources results in increased system costs, all of which ultimately lead to low productivity and loss of available resources. Therefore, the main objective of this research is to identify anomalies in the semiconductor wafer production process using machine learning methods. The data used includes various features from produced wafers collected from a major manufacturer in the semiconductor industry, containing information about the status of wafers during the production process. To improve model performance and reduce the negative effects of outlier data, a winsorizing method was used to adjust extreme values in some features. Additionally, to better prepare the data, features were standardized so that the model would not be sensitive to scale differences between features.

Method: In this research, through data preprocessing methods and simulation in Python software, efforts were made to increase the model's accuracy in identifying anomalies. The first step was data preparation and removal or adjustment of outlier data. Since some features contained extreme values that could skew the model, a "winsorizing" method was employed. Winsorizing means limiting very large and very small values of each feature to certain thresholds to reduce their impact on model performance. Another key step in this project was dimensionality reduction; given that this dataset includes 1,558 features, processing and analyzing all these features requires significant computational resources and may complicate the model unnecessarily. Therefore, using Linear Discriminant Analysis (LDA), the dimensionality of the data was reduced to a lower-dimensional space to create better separation between normal and anomalous classes. This dimensionality reduction helps the model classify data more accurately while simplifying computational processing.

Received: Jul. 09, 2024; Revised: Oct. 23, 2024; Accepted: Sep. 06, 2025; Published Online: Sep. 22, 2025.

*Ph.D. Student, Department of Industrial Management, North Tehran Branch, Islamic Azad University, Tehran, Iran.
**Associate Professor, Department of Applied Mathematics, South Tehran Branch, Islamic Azad University, Tehran, Iran.
Corresponding Author: khademi@azad.ac.ir
***Assistant Professor, Department of Industrial Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.
****Assistant Professor, Department of Industrial Management, North Tehran Branch, Islamic Azad University, Tehran, Iran.

Ind. Manag. Persp. 2025, 15(3), 170-194; doi: 10.48308/JIMP.15.3.170

Industrial Management Perspective

Journal hompage: https://jimp.sbu.ac.ir/?lang=en

Original Article

Findings: After data preparation, the standard table of orthogonal arrays in Taguchi method is used to standardize the data. L9(34) orthogonal arrays are selected as the most suitable design for models three to six. Then, the research data is used to identify anomalies using XGBoost model and genetic algorithm and compare the two models. The performance of the model was evaluated using confusion matrix and ROC curve and the efficiency of the algorithm. The results showed that the model has a high ability to identify anomalies and the value under the curve AUC was obtained equal to 0.97.Next, in order to further optimize and manage the challenge of data imbalance, Genetic Algorithm (GA) was used as an evolutionary approach to adjust the feature weights and classification threshold. These results indicate the ability of the model to distinguish healthy and defective samples with high accuracy. This research shows that by using appropriate data preprocessing techniques and machine learning models, successful results can be achieved in identifying production anomalies and identifying defective parts and preventing defective products from entering the market

Conclusion: The results of this study showed that the XGBoost method has a high ability to detect anomalies. Also, the genetic algorithm has been able to improve performance metrics such as precision (92.4%), recall (0.924), and score (0.913) and provide stable convergence over different generations. The combination of XGBoost and genetic algorithm (GA) allows for more accurate detection of anomalies and shows that this approach can be used as a practical framework in improving quality control, reducing waste, and increasing the efficiency of production lines.

Keywords: Intelligent anomaly, XGBoost model, unbalanced industry, production line, defective products, genetic algorithm.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

How to Cite: Nematnia, Rasoul; Khademi, Maryam; Kiamars, Fathi; Sardar, Soheila (2025). Intelligent Anomaly Detection in Unbalanced Industrial Data Using the Xgboost Model and Genetic Algorithm (GA) To Optimize Performance in Identifying Defective Products in the Production Line. *Ind. Manag. Persp.*, 15(3), 170-194 *(In Persian)*.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

چشم‌انداز مدیریت صنعتی، ۱۴۰۴، دوره ۱۵، شماره ۳، صص ۱۷۰–۱۹۴، doi: 10.48308/JIMP.15.3.170

چشم‌انداز مدیریت صنعتی

Journal hompage: https://jimp.sbu.ac.ir/?lang=en

نوع مقاله: پژوهشی

شناسایی هوشمند ناهنجاری در داده‌های صنعتی نامتوازن با استفاده از مدل XGBoost و الگوریتم ژنتیک (GA) جهت بهینه‌سازی عملکرد در شناسایی محصولات معیوب در خط تولید

رسول نعمت‌نیا*

مریم خادمی**

کیامرث فتحی***

سهیلا سردار****

چکیده گسترده

مقدمه و اهداف. فرآیند خطوط تولید و توالی آن یکی از رویکرد های اساسی در برنامه ریزی محصولات صنعتی به صورت انبوه است. عدم برنامه‌ریزی در خطوط و راه کار مناسب برای بهینه‌سازی سیستم‌های موثر در فرآیند تولید و مونتاژ، باعث افزایش زمان تخصیصی به امر تولید و افزایش زمان‌های توقف ماشین‌آلات و در نتیجه کاهش تعداد محصولات تولید از لحاظ تعدادی و نرخ تولید عدم کارایی منابع تخصیصی و در اختیار و در نتیجه افزایش هزینه‌های سیستم می‌شود که همه این عوامل در نهایت باعث بهره وری پایین و از دست دادن منابع موجود است. از این رو در این پژوهش هدف اصلی شناسایی ناهنجاری‌ها در فرآیند تولید ویفرهای نیمه هادی با استفاده از روش‌های یادگیری ماشین است. داده‌های مورد استفاده شامل ویژگی‌های مختلفی از ویفرهای تولیدی است که از یک تولید کننده بزرگ در صنعت نیمه هادی جمع‌آوری شده و حاوی اطلاعاتی از وضعیت ویفرها در فرآیند تولید است. به منظور بهبود عملکرد مدل و کاهش اثرات منفی داده‌های پرت، از روش وینزوریزه کردن برای تعدیل مقادیر بسیار دور از میانگین در برخی از ویژگی‌ها استفاده شد. همچنین، برای آماده‌سازی بهتر داده‌ها، ویژگی‌ها استانداردسازی شدند تا مدل نسبت به تفاوت مقیاس بین ویژگی‌ها حساس نباشد.

روش‌ها. در این پژوهش، با استفاده از روش‌های پیش پردازش داده و همچنین شبیه‌سازی در نرم‌افزار پایتون، سعی شد تا دقت مدل در شناسایی ناهنجاری‌ها افزایش یابد. اولین گام، آماده‌سازی داده‌ها و حذف یا تعدیل داده‌های پرت بود. به دلیل اینکه برخی از ویژگی‌ها شامل مقادیر بسیار و تعداد زیاد دور از میانگین بودند که می‌توانستند مدل را دچار انحراف کنند، از روش "وینزوریزه کردن" استفاده شد. وینزوریزه کردن به این معناست که مقادیر بسیار بزرگ و بسیار کوچک هر ویژگی به آستانه‌های معینی محدود می‌شوند تا تأثیر آن‌ها بر عملکرد مدل کاهش یابد.

تاریخ دریافت: ۱۴۰۳/۰۴/۱۹، تاریخ بازنگری: ۱۴۰۳/۰۸/۰۲، تاریخ پذیرش: ۱۴۰۴/۰۶/۱۵، تاریخ اولین انتشار: ۱۴۰۴/۰۶/۳۱.

*دانشجوی دکتری گروه مدیریت صنعتی، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران.

** دانشیار گروه ریاضی کاربردی، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران.

نویسنده مسئول: khademi@azad.ac.ir

*** استادیار گروه مدیریت صنعتی، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران.

**** استادیار گروه مدیریت صنعتی، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران.

چشم‌انداز مدیریت صنعتی، ۱۴۰۴، دوره ۱۵، شماره ۳، صص ۱۷۰–۱۹۴، doi: 10.48308/JIMP.15.3.170

چشم‌انداز مدیریت صنعتی

Journal hompage: https://jimp.sbu.ac.ir/?lang=en

نوع مقاله: پژوهشی

یکی دیگر از گام‌های کلیدی در این پروژه، کاهش ابعاد داده‌ها بود. با توجه به اینکه این مجموعه داده شامل ۱۵۵۸ ویژگی است، پردازش و تحلیل تمامی این ویژگی‌ها نیازمند منابع محاسباتی قابل توجهی است و ممکن است مدل را پیچیده‌تر از حد لازم کند. از این رو، با بهره‌گیری از روش "تحلیل تفکیک خطی(LDA) ، ابعاد داده‌ها به فضای بُعد پایین‌تری کاهش یافت تا جدایی بهتری بین کلاس‌های نرمال و ناهنجار ایجاد شود. این کاهش ابعاد به مدل کمک می‌کند تا طبقه بندی داده‌ها را با دقت بیشتری انجام دهد و همچنین پردازش محاسباتی ساده‌تر شود.

یافته‌ها. پس از آماده‌سازی داده‌ها، برای استانداردسازی داده‌ها ازجدول استاندارد آرایه‌های متعامد در روش تاگوچی استفاده می‌شود آرایه‌های متعامد L9(34) به عنوان مناسب‌ترین طرح برای مدل‌های سه تا شش انتخاب می‌شوند. سپس داده‌ها مربوط به پژوهش با استفاده از مدل XGBoost و الگوریتم ژنتیک برای شناسایی ناهنجاری‌ها و مقایسه دو مدل استفاده شده است . عملکرد مدل با استفاده از ماتریس در هم ریختگی و منحنی ROC و کارائی الگوریتم ژنتیک مورد ارزیابی قرار گرفت. نتایج نشان داد که مدل توانایی بالایی در شناسایی ناهنجاری‌ها دارد و مقدار زیر منحنیAUC برابر با ۰.۹۷ به دست آمد. در ادامه، به‌منظور بهینه‌سازی بیشتر و مدیریت چالش عدم توازن داده–ها، از الگوریتم ژنتیک (GA) به‌عنوان یک رویکرد تکاملی برای تنظیم وزن ویژگی‌ها و آستانه طبقه‌بندی استفاده شد این نتایج‌نشان‌دهنده توانایی مدل در تفکیک نمونه‌های سالم و معیوب با دقت بالا است. این پژوهش نشان می‌دهد که با استفاده از تکنیک‌های مناسب پیش پردازش داده و مدل‌های یادگیری ماشین، می‌توان در شناسایی ناهنجاری‌های تولید و شناسایی قطعات معیوب به نتایج موفقیت آمیزی دست یافت و از ورود محصولات معیوب به بازار جلوگیری کرد.

نتیجه‌گیری. نتایج به دست آمده از این تحقیق نشان داد که روش XGBoost توانایی بالایی در تشخیص ناهنجاری‌ها دارد.و همچنین الگوریتم ژنتیک توانسته است معیارهای عملکردی مانند دقت (۹۲.۴٪)، ، فراخوانی (۰.۹۲۴) و امتیاز (۰.۹۱۳) را بهبود دهد و همگرایی پایداری در طول نسل‌های مختلف ارائه کند. ترکیب XGBoost و الگوریتم ژنتیک (GA) امکان شناسایی دقیق‌تر ناهنجاری‌ها را فراهم کرده و نشان می‌دهد که این رویکرد می‌تواند به‌عنوان یک چارچوب عملی در بهبود کنترل کیفیت، کاهش ضایعات و افزایش بهره‌وری خطوط تولید مورد استفاده قرار گیرد.

کلمات کلیدی: ناهنجاری هوشمند، مدلXGBoost ، صنعتی نامتعادل، خط تولید، محصولات معیوب، الگوریتم ژنتیک.

استناددهی: نعمت‌نیا، رسول؛ خادمی، مریم؛ فتحی، کیامرث؛ سردار، سهیلا (۱۴۰۴). شناسایی هوشمند ناهنجاری در داده‌های صنعتی نامتوازن با استفاده از مدل XGBoost و الگوریتم ژنتیک (GA) جهت بهینه‌سازی عملکرد در شناسایی محصولات معیوب در خط تولید. چشم‌انداز مدیریت صنعتی، ۱۵(۳)، ۱۷۰–۱۹۴.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

۱. مقدمه

در دنیای صنعتی امروزی، شناسایی ناهنجاری‌ها در فرآیندهای تولید، از اهمیت بالایی برخوردار است. زیرا وجود یک محصول معیوب می‌تواند تأثیرات جدی بر کیفیت و اعتبار برند تولیدکننده داشته باشد. در این تحقیق، از داده‌های یک تولید کننده برتر ویفرهای نیمه هادی در هند استفاده شده است. این داده‌ها شامل اطلاعاتی از فرآیند تولید و پارامترهای مختلفی هستند که توسط دستگاه‌های پیشرفته به طور مداوم و در هر ده میلی ثانیه جمع‌آوری شده است. به دلیل ماهیت پیچیده و دقیق فرآیند تولید نیمه هادی‌ها، سیستم‌های صنعتی باید به دقت تحت نظارت قرار گیرند تا رفتار و کیفیت محصولات نهایی تضمین شود. تکنولوژی به ویژه در توسعه کارخانه‌های کاملاً خود مختار مهم است. بازرسی کیفیت محصول یک جزء حیاتی در تولید صنعتی است. یک سیستم تشخیص ناهنجاری و محلی‌سازی دقیق و قابل اعتماد مبتنی بر هوش مصنوعی برای بازرسی کیفیت محصولات صنعتی در کارخانه‌های تولیدی در دنیای واقعی ضروری است [۲۳ و ۳۳].

جمع‌آوری محصولات غیرعادی عظیم دشوار است زیرا تعداد محصولات غیرعادی محدود و در سناریوی ساخت واقعی نادر است. تشخیص عیوب/ناهنجاری‌ها نقش مهمی در حوزه‌های مختلف تولید صنعتی تولیدی برای حفظ استانداردهای کیفیت محصول ایفا می‌کند تشخیص ناهنجاری و محلی‌سازی معمولاً در مرحله نهایی فرآیند تولید برای بازرسی کیفیت محصول و شناسایی عیوب محصول انجام می‌شود. مقدار و شدت یک عیب به طور قابل توجهی بر قیمت یک محصول تأثیر می‌گذارد و آن را تعیین می‌کند. در سناریوهای بازرسی سنتی و دستی، کارگران پس از تولید محصول در خط تولید، کیفیت محصولات را یکی یکی بررسی می‌کنند. بدیهی است که با افزایش حجم تولید و تقاضاهای رو به رشد، تکیه صرفاً بر بازرسی انسانی چالش برانگیزتر است. علاوه بر این، قضاوت‌های ذهنی و تعصبات کارگران منجر به معیارهای بازرسی کیفیت متناقض می‌شود که این امز منجر به نرخ فرار نقص بالاتر است. در دهه گذشته، هوش مصنوعی (AI)[1] و فن‌آوری تشخیص دید عمیق به سرعت در برنامه‌های کاربردی دنیای واقعی، مانند وسایل نقلیه خودران ، سیستم‌های نظارتی، تصویر برداری پزشکی و غیره در حال توسعه هستند. در عین حال، شبکه‌های عصبی عمیق (DNN)[2] نیز در کارخانه‌ها برای شناسایی و شناسایی عیوب (ناهنجاری) یک محصول به کار گرفته شده اند، زیرا دقت بالاتر و سرعت بازرسی سریع‌تری نسبت به روش‌های بازرسی سنتی ارائه می‌دهند. سیستم بازرسی کیفیت مبتنی بر هوش مصنوعی مزایایی را برای صنعت تولید به ارمغان می‌آورد. اولاً، سیستم بازرسی مبتنی بر هوش مصنوعی با خود کار کردن برخی از فرآیندهای تولید، بار کاری کارگران را کاهش می‌دهد. ثانیاً، یک مدل مبتنی بر هوش مصنوعی منجر به دقت بازرسی دقیق‌تر نقص با زمان بازرسی کمتر می‌شود، بنابراین کارایی کلی فرآیند تولید را برای کمک به تولید کنندگان برای بر آوردن تقاضای بالاتر افزایش می‌دهد. ثالثاً، سیستم بازرسی مبتنی بر هوش مصنوعی معیارهای بازرسی منسجم را در کل خط تولید اعمال می‌کند تا اطمینان حاصل شود که همه محصولات دارای کیفیت ثابت هستند [۱۳].

به طور کلی، سیستم بازرسی مبتنی بر هوش مصنوعی به تولید کنندگان کمک می‌کند تا بهره وری و کارایی را در فرآیندهای تولید افزایش دهند. در این پژوهش ما می‌خواهیم با استفاده از تکنیک‌های مناسب پیش‌پردازش داده و مدل‌های یادگیری ماشین، به شناسایی ناهنجاری‌های تولید به نتایج موفقیت آمیزی دست یابیم و در این سیستم با آنالیز دستگاه در خط تولید به بررسی راندمان کار پرداخته و تولید سالم و معیوب توسط دستگاه را مورد ارزیابی قرار دهیم تا از ورود محصولات معیوب به بازار جلوگیری کنیم.

۲.مبانی نظری و پیشینه پژوهش

مبانی نظری در این پژوهش به بررسی خطاها و اشتباهات در خط تولید پرداخته و در ادامه به بررسی بهینه‌سازی و عملکرد و همچنین ناهنجاری‌ها و مشکلات در خط تولید می‌پردازد و در ادامه به مدل انتخابی ما در این پژوهش یعنی XGBoost اشاره دارد.

خطا در خط تولید: پیش بینی خطاها می‌تواند به طور چشمگیری زمان از کار افتادن ماشین‌ها را در محیط‌های صنعتی کاهش دهد و حتی اجازه می‌دهد تا مدت‌ها قبل از اینکه خطا بر سیستم تولید تأثیر بگذارد، اقدامات متقابل انجام شود. یک سیستم پشتیبان برای پیش‌بینی

1. Artificial Intelligence
2. Deep Neural Networks

بحران‌های آینده برای خطوط تولید تحت شرایط مختلف محیطی همیشه اقداماتی را در نظر می‌گیرد و این سیستم پشتیبان دارای اهمیت بالایی است. همیشه باید بر روی خطاهایی تمرکز کنیم که این خطاها منجر به صدمات زیادی می‌شود که یکی از این خطاها چندین قطعه کار اشتباه در سیستم خط تولید است. این الگوهای خطا نیاز به تصحیح دستی توسط کنترلر ماشین دارند. طبق تجزیه و تحلیل یک سیستم که اطلاعات مربوط به انواع خطاهای قابل مشاهده را جمع‌آوری کرده بود مشاهده شد که ۳۰ درصد خطاها، خطاهای اندازه گیری یا از کار افتادن قطعه‌های معیوب هستند. این خطاها به هیچ نوع عملی از سوی کنترل کننده ماشین نیاز ندارند. ۷۰٪ از خطاهای مشاهده شده انحرافات مداوم سیستم است که منجر به صدمات به بقیه قطعات در یک بازه زمانی می‌شود [۳۰]. در نگاه دیگر به دلیل اثر حرارتی در یک خط تولید، زمانی که ماشین ابزار با سرعت بالا کار می‌کند، دقت ماشین کاری به طور اجتناب ناپذیری در معرض کاهش است، بررسی‌های قبلی در مورد این موضوع نشان داده است که برای ماشین ابزار دقیق، خطاهای ماشین کاری ناشی از تغییر شکل حرارتی ۴۰ درصد تا ۷۰ درصد از کل خطا را تشکیل می‌دهد [۲۶]. چنین نسبتی به اندازه کافی چشمگیر است تا قطعات ماشین کاری شده را تخریب کند و وضعیت به ویژه زمانی بدتر می‌شود که دمای محیط ماشین ابزار در حال سرویس به درستی تنظیم نشده باشد. از آنجایی که اثر حرارتی تأثیر زیادی بر فرآیند ماشین‌کاری ماشین ابزار می‌گذارد، کاهش خطای حرارتی ماشین ابزار برای بهبود دقت ماشین‌کاری اهمیت عملی دارد. در حال حاضر روش‌های کاهش خطای حرارتی ماشین ابزارها به طور عمده به پیشگیری از خطا و جبران خطا تقسیم می‌شوند [۱۲].

بهینه‌سازی تولید: با افزایش ماهیت تقاضای مشتری، تغییرات تولید، محصول و طراحی تولید بیشتر شده است. علاوه بر این، اعتبار سنجی ناکافی در مرحله طراحی ساخت ممکن است منجر به مسائل اضافی مانند طراحی مجدد فرآیند و تخصیص مجدد طرح، در مرحله بهره برداری شود. بنابراین، سیستم‌هایی که بتوانند از قبل اعتبار سنجی کنند و امکان تجزیه و تحلیل دقیق و قابل اعتماد را در مرحله طراحی ساخت فراهم کنند، و همچنین تغییرات را در خطوط تولید در زمان واقعی اعمال و بهینه کنند، نیاز ضروری یک سیستم مدیریتی در خط تولید است [۱۱].

الگوریتم ژنتیک: الگوریتم ژنتیک یک روش بهینه‌سازی الهام‌گرفته از طبیعت است که بر اساس اصول تکامل داروین کار می‌کند. در این الگوریتم، ما با یک جمعیت اولیه از راه حل‌های ممکن (که هر کدام یک فرد نامیده می‌شود) شروع می‌کنیم. هر فرد شامل مجموعه‌ای از پارامترها است، مانند وزن‌ها و ویژگی‌ها و یک آستانه برای طبقه‌بندی. سپس، برای هر فرد، یک تابع تناسب (فیتنس) محاسبه می‌شود که نشان‌دهنده کیفیت آن راه‌حل است. در مسأله ما، این تابع می‌تواند بر اساس دقت طبقه‌بندی یا معیارهایی مانند AUC (مساحت زیر منحنی ROC) باشد، که برای مسائل عدم تعادل کلاس–ها مناسب است.

باید گفت صنعت تولید در حال حاضر با تغییراتی که در بازار اتفاق افتاده است، از جمله تقاضاهای غیر قابل پیش بینی محصول و افزایش تقاضا برای محصولات سفارشی به فکر معرفی سیستم‌های تولید هوشمند و کارخانه‌هایی جهت پاسخگویی به انعطاف پذیری سفارشات و ارائه سیستم هوشمندانه افتاده است [۱۰ و ۱۹]. از آنجایی که کارخانه هوشمند به عنوان «سیستم تولید کاملاً یکپارچه و مشارکتی که می‌تواند در زمان واقعی پاسخگوی نیازها و شرایط پویای کارخانه‌ها، زنجیره‌های تأمین و مشتریان باشد» تعریف می‌شود، تلاش قابل توجهی را انجام داده است که در جهت برنامه ریزی و تحقیق و توسعه است که باعث تولید پیشرفته و بهبود خدمات می‌شود.

علت ناهنجاری‌ها: کنترل کیفیت محصول و پاسخگویی به تقاضای مشتری نقش مهمی در صنعت تولید دارد [۲ و ۱۸]، امروزه نیازهای مشتریان بسیار پیچیده شده است و تنها تعداد کمی از کارخانه‌ها قادر به دست‌یابی به نتایج مناسب خود و رفع نیازهای مشتریان هستند؛ همچنین برخی چالش‌ها و رقابت‌ها مانند فشارهای خارجی کارخانه‌ها را مجبور به کاهش زمان تولید محصولات می‌کند. ازآنجا که بازار در سراسر جهان پراکنده شده است، فعالیت‌های تولیدی نیز در یک محل واحد محدود نمی‌شوند و در سطح جهانی گسترش یافته‌اند. برنامه‌ریزی تولید و زمان‌بندی متمرکز سنتی برای پاسخ به تغییرات سریع بازار به اندازه کافی انعطاف‌پذیر نیست [۵].

در این میان تشخیص مشکلات در عرضه و رفع ناهنجاری‌ها در داده‌های تجهیزات تولید در کارخانه‌های هوشمند بسیار مهم است. از این طریق، تصمیم‌گیری بهتر می‌تواند کارایی تولید را در فرآیندهای تولید بهبود دهد. به عنوان مثال نگوین و همکاران (۲۰۲۱)، یک مدل

رمزگذار خودکار[1] LSTM برای تشخیص ناهنجاری با استفاده از داده‌های سری زمانی چند متغیره پیشنهاد کردند [۱۷]. و یا در دیگر تحقیقات کاربرد عملی را در محیط‌های واقعی کارخانه در نظر نگرفتند. در جدول (۱) مطالعات مرتبط در مورد تشخیص ناهنجاری در خط تولید و سیستم‌های تولید مورد بررسی قرار گرفته است.

جدول ۱. پیشینه مطالعاتی در مورد تشخیص ناهنجاری در خط تولید

نویسندگان / سال	حوزه کاربرد	داده/مجموعه داده	روش / مدل به‌کاررفته	یافته‌های کلیدی	محدودیت‌ها / شکاف پژوهشی
Chen & Guestrin (2016)	علوم داده، طبقه‌بندی	مجموعه داده‌های عمومی (Kaggle, UCI)	معرفی الگوریتم XGBoost	ارائه مدلی بسیار کارا برای داده‌های بزرگ و نامتوازن، کاهش overfitting	تمرکز بیشتر روی داده‌های عمومی، نه صنعتی
Zhang et al. (2019)	صنعت تولید قطعات الکترونیکی	داده‌های سنسور خط تولید	XGBoost + SMOTE	افزایش دقت شناسایی محصولات معیوب در شرایط داده نامتوازن	نیاز به تنظیم پارامترهای زیاد، هزینه محاسباتی بالا
Li et al. (2020)	سیستم‌های هوشمند تولید	داده‌های IoT در کارخانه	XGBoostترکیبی باPCA	بهبود تشخیص ناهنجاری با کاهش ابعاد ویژگی‌ها	عدم بررسی مقیاس‌پذیری برای داده‌های بسیار حجیم
Kumar & Singh (2021)	پیش‌بینی خرابی ماشین‌آلات	داده‌های صنعتی موتور	XGBoost + Random Forest مقایسه	XGBoostعملکرد دقیق‌تری در داده‌های نامتوازن داشت	تمرکز روی داده خرابی ماشین، نه محصول نهایی
Wang et al. (2022)	کنترل کیفیت صنعتی	داده‌های سنسور خط تولید خودرو	XGBoost +روش وزن‌دهی کلاس‌ها	بهبود دقت و F1 در تشخیص قطعات معیوب	حساسیت زیاد به انتخاب هایپرپارامترها
Ahmad et al. (2023)	یادگیری ماشین در صنعت	داده‌های تولید مواد غذایی	XGBoost + Ensemble	ترکیب مدل‌ها برای کاهش نرخ خطای منفی	هزینه زمانی و پردازشی زیاد
پژوهش پیشنهادی	خط تولید صنعتی (پژوهش حاضر)	داده‌های واقعی خط تولید	XGBoostبهینه‌سازی شده (Hyperparameter tuning, Cross-validation)	افزایش دقت شناسایی محصولات معیوب، کاهش خطای نوع دوم	پژوهش در حال توسعه – نیاز به تست در صنایع مختلف

۳. مدل پژوهش:

مدل XGBoost: برنامه ریزی تولید هوشمند بخش مهمی از تولید هوشمند است و زمان و شدت تولید آن به طور معقولی تنظیم شده است که می‌تواند کارایی تولید را به طور کامل بهبود بخشد. بر اساس الگوریتم XGBoost، خط تولید مورد نیاز برای ایجاد ارتباط بین محتوای تولید و خط تولید، و تحقق تطابق خودکار علامت گذاری می‌شود. پس از طبقه بندی، میانگین ساعات کار و سایر اطلاعات بر اساس منطق تعیین شده برآورد می‌شود. در نهایت تولید را بر اساس اصل استفاده حداکثری از خط تولید برنامه ریزی و تنظیم می‌کند. از طریق آزمایش و ارزیابی، نتایج زمان بندی این مدل زمان بندی هوشمند تنها در چند ثانیه تکمیل می‌شود که باعث صرفه‌جویی در زمان و هزینه در فرآیند زمان بندی می‌شود. [۱۶ و ۸]، یافته‌ها و ارزیابی‌های مقایسه‌ای الگوریتم‌های یادگیری ماشین نشان داد که مدل‌های جنگل تصادفی[2] یک الگوریتم مجموعه کیسه‌ای، و XGBoost، یک روش تقویت، از الگوریتم‌های فردی در ارزیابی بهتر عمل می‌کنند. بهترین مدل‌های یادگیری ماشین در این مطالعه با سیستم تولید در کارخانه ادغام شده‌اند [۶].

مدل XGBoost یک پیشرفت نسبتاً جدید در یاد گیری استقرایی است. این یاد گیرنده به طور گسترده با توافق استفاده شده است که تمایل دارد از نظر محاسباتی کار آمدتر باشد و به طور کلی برای طیف گسترده‌ای از انواع داده‌ها، از مشکلات طبقه بندی رایج تا تشخیص الگو در ویژگی‌های وابسته به زمان قابل استفاده باشد. این الگوریتم بر اساس الگوریتم تقویت گرادیان با معرفی یک پارامتر منظم‌سازی است که حساسیت درخت رگرسیون فردی را نسبت به نقاط پرت در مجموعه داده کاهش می‌دهد. به این ترتیب، این الگوریتم منجر به مدلی خواهد شد که واریانس کمتری را نسبت به مدلی که از تقویت گرادیان به تنهایی آموخته شده است را داشته باشد. استفاده از این یاد

1.Long Short Term Memory

2.Random Forest

گیرنده در ترکیب با جنگل تصادفی دو مدل ارائه می‌دهد که به ترتیب از اصول مجموعه‌های بوست شده و کیسه‌ای درختان تصمیم آموخته شده اند. روش‌های اضافی مانند ماشین‌های بردار پشتیبان با استفاده از هسته‌های مختلف یا شبکه‌های عصبی مصنوعی نیز می‌توانند برای مقایسه در صورت عدم رضایت بخش بودن نتایج این یاد گیرندگان استفاده شوند [۲۴].

در تعریفی دیگر مدل XGBoost، یک روش یادگیری نظارت شده بهینه معرفی شده است در مقایسه با سایر مدل‌های یادگیری تحت نظارت احتمالی، مانند رگرسیون لجستیک، K- نزدیک‌ترین همسایه‌ها، و درخت تصمیم، XGBoost دارای مزایای متمایز بسیاری است:

۱. **ماهیت مجموعه‌ای** XGBoost با ترکیب طبقه‌بندهای ضعیف در چارچوب تقویت گرادیان، یک طبقه‌بندی قدرتمند و دقیق ایجاد می‌کند که توانایی بالایی در پیش‌بینی دارد.
۲. **انعطاف‌پذیری بالا:** این مدل از توابع هدف و معیارهای ارزیابی تعریف شده توسط کاربر پشتیبانی می‌کند، مشروط بر اینکه تابع هدف دارای مشتق درجه دوم باشد.
۳. **کارایی محاسباتی** XGBoost : به طور ذاتی از پردازش موازی بهره می‌برد و در نتیجه زمان محاسبات به شکل چشمگیری کاهش می‌یابد.
۴. **کنترل پیچیدگی مدل:** با افزودن عبارت منظم‌سازی به تابع هزینه، مدل قادر است پیچیدگی را مهار کرده، واریانس را کاهش دهد و از بروز برازش بیش‌ازحد[1] جلوگیری کند.
۵. **مدیریت داده‌های ناقص:** این الگوریتم به‌صورت درونی قابلیت پردازش داده‌های گمشده را دارد، بدون آنکه نیاز به پیش‌پردازش اضافی باشد [۲۵].

تقویت گرادیان شدید (XGBoost) این الگوریتم روش محاسبه تابع هدف را بر اساس تقویت گرادیان بهبود می‌بخشد و زمان محاسبه را کاهش می‌دهد در طول دوره آموزش، محاسبات موازی به طور خود کار برای حل سریع و دقیق مسائل علم داده‌های بزرگ تحقق می‌یابد مفهوم اصلی XGBoost یادگیری ویژگی‌های جدید با افزودن ساختار درختی، برازش باقی مانده‌های پیش بینی نهایی و سپس به دست آوردن امتیاز نمونه است. با جمع امتیازهای هر درخت می‌توان امتیاز پیش بینی نهایی نمونه را به دست آورد. برای n نمونه با m ویژگی، فرمول پیش بینی امتیازات با توابع جمع K به شرح زیر است:

مسأله بهینه‌سازی تابع هدف را به مسأله یافتن حداقل مقدار تابع درجه دوم تبدیل می‌کند و از اطلاعات مشتق دوم تابع ضرر برای آموزش مدل درختی استفاده می‌کند. در همان زمان، پیچیدگی درخت به عنوان یک عبارت منظم به تابع هدف اضافه می‌شود تا از مشکل بیش از حد برازش جلوگیری شود. تابع هدف XGBoost به شرح زیر است [۹].

رابطه (۱)

$$l = \sum_{i=1}^{n} L\ (Y_i\ , \hat{y}_i\ ) + \sum_{k=1}^{k} \Omega\ (f_k),$$

پس از هر تکرار، XGBoost وزن گره برگ را در یک ضریب ضرب می‌کند تا تأثیر هر درخت را ضعیف کند تا فضای یادگیری بیشتری در مراحل بعدی وجود داشته باشد. ابزار XGBoost از موازی‌سازی پشتیبانی می‌کند و یکی از برترین زمان مراحل در یادگیری درخت تصمیم، مرتب کردن مقادیر ویژگی‌ها است. XGBoost داده‌ها را از قبل مرتب می‌کند و آنها را به عنوان یک ساختار بلوکی ذخیره می‌کند که به طور مکرر در تکرارهای بعدی استفاده شود تا به میزان قابل توجهی میزان محاسبه را کاهش دهد. این ساختار بلوکی موازی‌سازی را نیز ممکن می‌سازد. هنگام تقسیم گره‌ها، بهره در هر ویژگی باید محاسبه شود و می‌توان با محاسبه موازی به دست آورد. در نهایت، ویژگی با بیشترین بهره برای تقسیم انتخاب می‌شود [۲۷].

1.Overfitting

رابطه (۲)

$$\hat{y} = \sum_{k=1}^{k} f_k\ (x_i), f_k \in F,$$

$$F = \{f(x) = w_{q(x)}\ (q: R^m \to T, w \in R^T)\}$$

صادقی و همکاران[1] (۲۰۲۳) در پژوهشی تحت عنوان برنامه‌ریزی سیستم تولید اقتصادی با در نظر گرفتن تقاضای متغیر و خرابی تصادفی ماشین، به بررسی مسأله تولید اقتصادی با در نظر گرفتن محصولات معیوب حین تولید و همچنین خرابی ماشین و تقاضای متغیر اشاره کرد. در این پژوهش فرض شده است که محصولات با نرخ ثابتی تولید می‌شوند؛ ولی ماشین حین تولید ممکن است دچار خرابی شود. خراب شدن ماشین در حین تولید، یک متغیر تصادفی است که از توزیع نمایی با پارامتر مشخص پیروی می‌کند. به این منظور ابتدا یک مدل ریاضی ارائه می‌شود، سپس مقدار متوسط هزینه در واحد زمان تعیین شده و بر اساس مفاهیم بهینه‌سازی سراسری، مقادیر بهینه مشخص خواهد شد. در نهایت با حل یک مثال عددی به تجزیه و تحلیل مسأله بیان شده پرداخته شده است [۲۲].

مسلم‌پور و غدیرپور[2] (۲۰۲۱) در مقاله ای با موضوع طراحی هوشمند استقرار پویای تسهیلات در محیط تصادفی سیستم‌های تولید انعطاف‌پذیر با در نظر گرفتن انعطاف‌پذیری مسیر تولید یک مدل ریاضی جدید مبتنی بر مدل تخصیص درجه دوم برای طراحی استقرار بهینه تسهیلات در هر دوره از افق برنامه‌ریزی زمانی چند دوره ای مسأله استقرار پویا و تصادفی تسهیلات را ارائه کردند و همچنین برای حل مدل ریاضی پیشنهادی یک الگوریتم ترکیبی فراابتکاری [3]جدید با استفاده از الگوریتم‌های کرافت و شبیه‌سازی تبرید ارائه شد که نتایج آنان نشان می‌دهد که الگوریتم ترکیبی پیشنهادی از نظر کیفیت جواب و زمان محاسبه نسبت به الگوریتم تبرید شبیه‌سازی شده دارای عملکرد بهتری است [۱۵].

اوسوجا و همکاران[4] (۲۰۲۰) یک SLR را بر روی برنامه‌ریزی و کنترل تولید به کمک یادگیری ماشین انجام دادند. آنها فعالیت‌های درگیر در این مقالات (به عنوان مثال، استخراج ویژگی و آموزش مدل)، تکنیک‌ها (به عنوان مثال، ANN)، منابع داده (مانند، مدیریت، تجهیزات، کاربر، محصول، عمومی، مصنوعی)، موارد استفاده، و ویژگی‌های صنعت را بررسی کردند. با توجه به مطالعه SLR آنها، برنامه‌ریزی و زمان‌بندی هوشمند کاربردی‌ترین مورد استفاده است و ANN کاربردی ترین الگوریتم است. که استفاده از فن‌آوری‌های اینترنت اشیاء برای جمع‌آوری داده‌ها پیچیده است و اصلاح مدل یادگیری ماشین بر اساس تغییرات سیستم تولید آسان نیست [۳۱].

صبحی شجاع و سمویی[5] (۱۳۹۷) یک مدل ثابت برای مسأله بالانس خط مونتاژ رباتیک با چیدمان سلولی با در نظر گرفتن سه عامل الف. شروع عملیات ب. نصب و راه اندازی هر عملیات و ج. زمان‌های توقف ماشین‌ها برای هر نوع عملیات تعمیر و نگهداری، راه‌اندازی مجدد در جهت کاهش زمان فرآیند برای تعداد مشخصی از فعالیت‌ها در هر ایستگاه کاری با سیستم هوشمند و به صورت رباتیک ارایه می‌دهد. که با استفاده از نرم افزار گمز[6] بررسی شده است. که این نتیجه ضرورت در نظر گرفتن عدم قطعیت در برنامه ریزی و زمان بندی برای خطوط تولید با سیستم رباتیک است [۲۸]، دلیس و همکاران[7] (۲۰۱۷) یک الگوریتم ازدحام ذرات اصلاح شده برای بالانس خط مونتاژ دو طرفه با مدل ترکیبی ارائه داده‌اند. رویکرد پیشنهادی شامل روش های جدیدی مانند پروسه تولید که بر اساس مکانیزم انتخابی ترکیبی و رمز گشایی پروسه است. این روش های جدید ظرفیت الگوریتم راه‌حل را بالا می‌برد و آن را قادر به جستجو در نقاط مختلف فضای حل می‌نماید. الگوریتم برای حل مسأله با نرم افزار متلب مبنی بر الگوریتم بهینه‌سازی انبوه ذرات[8] با دانش منفی طراحی شده است. هدف این

1. Sadeghi
2. Moslemipour
3. Metaheuristic Algorithms
4. Usuga et al
5. Sobhi Shoje et al
6. Gams
7. Delice et al
8. PSO

مقاله، کمینه کردن تعداد ایستگاه های کاری برای یک زمان چرخه داده شده است. مکانیزم انتخابی ترکیبی از گیر افتادن در نقاط مینیمم محلی جلوگیری می‌کند [۴]، ژانگ و همکاران[1] (۲۰۱۷) یک الگوریتم تکاملی ترکیبی برای مسأله بالانس خط مونتاژ چند هدفه با عدم قطعیت ارائه داده‌اند. در این مقاله هدف بهینه‌سازی و کاهش زمان جابجایی و هزینه و نصب برای تعدادی از ایستگاه‌های همسان است. الگوریتم تکاملی ترکیبی معرفی شده یک روش ساده برای انتخاب بهینه پارتو بین پارتو مغلوب و رابطه مغلوب مبتنی بر تابع مناسب و الگوریتم ژنتیک ارزیابی شده برای افزایش همگرایی و عملکرد توزیع ارائه می‌دهد. نتایج بدست آمده از این پژوهش نشان داده است که مدل پیشنهادی و ارزیابی تکاملی انجام شده عملکرد بهتری در مقایسه با دو الگوریتم ژنتیک چند هدفه معمولی و الگوریتم ژنتیک غیر مغلوب و الگوریتم تکاملی پارتو در همگرایی و کارایی برای توزیع منابع در زمان تعریف شده دارد [۳۵]، لی و همکاران[2] (۲۰۱۷) یک بررسی جامع بر روش‌های هیوریستیک و متاهیوریستیک برای مسأله بالانس خط مونتاژ دو طرفه انجام داده‌اند. سهم این مقاله در دانش، مقایسه روش های تحقیقات، دسته بندی آنها به صورت شبیه سازی شناخته شده تا جستجوی محلی تکرار شونده و ارزیابی ۶ طرح رمز گذاری، ۳۰ فرآیند رمز گشایی و ۵ تابع هدف است. رویکرد طراحی تجربی برای کسب نتایج معتبر با استفاده از الگوریتم‌های تست تحت ۴ معیار به کار گرفته شده است. نتایج محاسباتی نشان داده است که انتخاب مناسب طرح رمز گذاری، فرآیند رمز گشایی و توابع هدف، عملکرد الگوریتم‌ها را به وسیله یک تفاوت قابل توجه بهبود داده است [۱۴].

همانطور که در بررسی‌ها نشان داده شده است، اکثر پژوهش‌های در حوزه بالانس خطوط مونتاژ و خط تولید و زمان‌بندی و هزینه است و همچنین در روش کار خود الگوریتم‌های پرکاربرد را ارائه دادند و کمتر به بررسی بهینه‌سازی و شناسایی محصولات معیوب در خط تولید و ناهنجاری‌ها در تولید پرداخته اند و از متد XGBoost که یک متد هوشمند و نوین است استفاده نشده است.

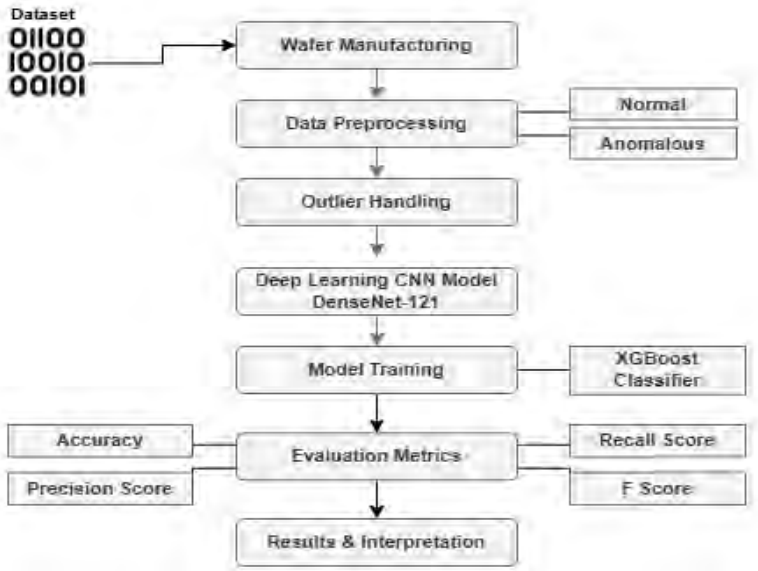
لازم به ذکر است نوآوری به کاررفته در این مدل درنظرگرفتن ناهنجاری‌ها در مسیر تولید محصولات است که به عنوان شناسایی ناهنجاری‌ها در تولید ویفرهای نیمه هادی و قطعات تولیدی است که با استفاده از مدل XGBoost بررسی و مقایسه مدل و روش مورد نظر با الگوریتم ژنتیک (GA) انجام می‌شود نتایج حاصل از هر دو روش مورد تجزیه و تحلیل قرار گرفته و خروجی هر کدام نشان داده می‌شود تا با استفاده از مقایسه حاصل از خروجی بهترین کارایی از خروجی حاصل نشان داده می‌شود این آنالیز هوشمند بوده و قطعات معیوب و سالم را به صورت دقیق تشخص می‌دهد که در هیچ یک از پژوهش‌های پیشین به آن پرداخته نشده است

که این نوآوری پژوهش محسوب می‌شود و دیتاست استفاده شده در کار بروز رسانی شده و خروجی کار دقیق است که در ادامه در این پژوهش بدان پرداخته خواهد شد.

۳. روش‌شناسی پژوهش

در این بخش، ابتدا به فلوچارت کلی پژوهش می‌پردازیم که روند شبیه‌سازی کار پژوهش را از ابتدا تا انتها نمایش می‌دهد (شکل ۱) و در ادامه به مفروضات مدل، سپس به معرفی دیتاست، مدل‌سازی و در نهایت به تست و ارزیابی مدل پیشنهادی با استفاده از مدل XGBoost و الگوریتم ژنتیک (GA) جهت بهینه‌سازی عملکرد در شناسایی محصولات معیوب در خط تولید ارائه شده است. برای استانداردسازی داده‌ها ازجدول استاندارد آرایه‌های متعامد در روش تاگوچی استفاده می‌شود. سپس تمامی روند کار در نرم افزار پایتون اتفاق می‌افتد. هدف از انتخاب این نرم‌افزار به این است که پایتون یک برنامه‌نویسی چند منظوره و سطح بالا است که به طور گسترده در جهان استفاده می‌شود و در دنیای واقعی پایتون یک ابزار سطح قدرتمند، و یک زبان برنامه‌نویسی شی‌گرا است. از این برنامه‌نویسی می‌توان در زمینه‌های مختلف هوش مصنوعی مانند یادگیری ماشین یا داده کاوی نیز استفاده کرد. در سال‌های اخیر استفاده از این زبان برنامه‌نویسی به منظور پردازش تصاویر افزایش یافته است [۲۱ و ۲۹].

3. Zhang et al
2. Li et al



شکل ۱. فلوچارت روند کار

مفروضات مدل پیشنهادی به شرح زیر است:

۱. پردازش داده‌ها و بهبود مدل برای شناسایی ناهنجاری‌ها در این پژوهش صورت گرفت.
۲. پس از آماده‌سازی داده‌ها، از مدل برای طبقه بندی[1] داده‌ها و شناسایی ناهنجاری‌ها استفاده شد.
۳. عملکرد مدل با استفاده از ماتریس در هم ریختگی و منحنی ROC مورد ارزیابی قرار گرفت.
۴. توانایی مدل در تفکیک نمونه‌های سالم و معیوب با دقت بالا است.
۵. استفاده از تکنیک‌های مناسب پیش پردازش داده و مدل‌های یادگیری ماشین، می‌توان در شناسایی ناهنجاری‌های تولید به‌نتایج موفقیت‌آمیزی دست یافت و از ورود محصولات معیوب به بازار جلوگیری کرد.
۶. در خط تولید در صد خطا و تلفات در تولید قطعات بالا است.
۷. راندمان کاری با توجه به استفاده از تکنولوژی و هوش مصنوعی در خط تولید بالا است.
۸. استفاده از دستگاه‌های قدیمی و بدون سرویس در خط تولید عامل بی کیفیتی قطعات است.
۹. بررسی ناهماهنگی‌ها و داده‌های پرت مهمترین روش در حل یک مشکل در خط تولید است.
۱۰. جهت کاهش اثر داده‌های پرت، تکنیک وینزوریزه کردن گزینه مناسبی است و انتخاب شد.
۱۱. در پیاده‌سازی برای محصولات خوب (صفر) و برای محصولات معیوب (یک) نشان دهنده وضعیت تولید کار دستگاه است.

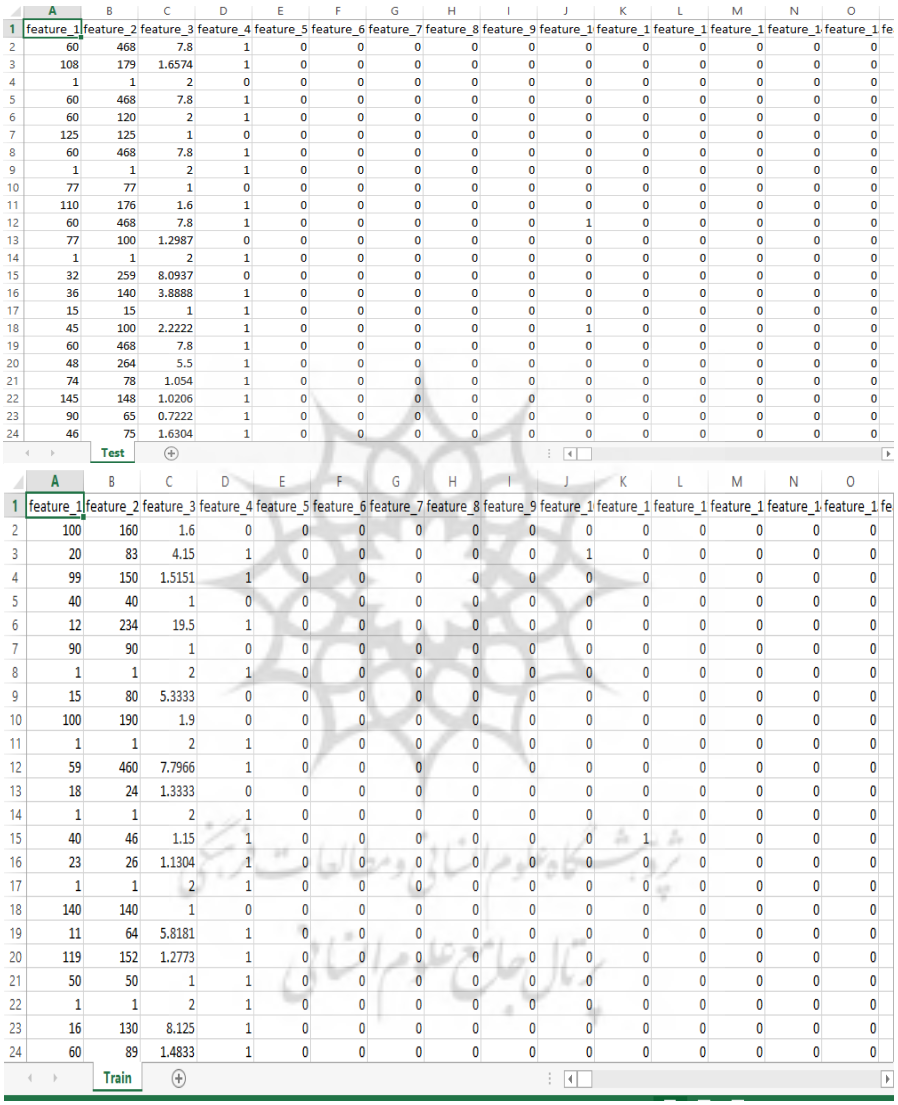
دیتاست و پیش‌پردازش:

بیان مسأله: این مجموعه داده شامل ۱۷۶۳ نمونه در فایل آموزشی (Train.csv) و ۷۵۶ نمونه در فایل تست (Test.csv) است[2] و هر کدام دارای ۱۵۵۸ ویژگی می‌باشند.

و کدهای داده‌ها برای نرم‌افزار پایتون در فهرست فایل‌ها است. ستون ویژگی‌های هر نمونه بیانگر پارامترهای خاصی از دستگاه تولیدی است و شامل اطلاعاتی در مورد شرایط و مشخصات محصول است. همچنین یک ستون[3] در داده‌های آموزشی وجود دارد که نمایانگر برچسب کیفیت محصول است، مقدار (۰ صفر) نشان دهنده محصولی بدون نقص و مقدار (۱ یک) نشان‌دهنده محصولی دارای نقص و

1. Classifier
2. anomaly-detection
3 Clas

ناهنجاری است. با این حال، شناسایی ناهنجاری‌ها به دلیل نادر بودن آن‌ها و نامتوازن بودن داده‌ها، به چالشی مهم تبدیل شده است، این عدم توازن در داده‌ها، مدل یادگیری ماشین را با چالشی جدی مواجه می‌سازد، زیرا تعداد کمی از نمونه‌ها به عنوان ناهنجار بر چسب‌گذاری شده‌اند، و این موضوع می‌تواند باعث بایاس مدل به سمت طبقه نرمال شود. از آنجا که در فرآیند تولید، ایجاد محصول معیوب به ندرت رخ می‌دهد، ناهنجاری‌ها شبیه به "یافتن سوزنی در انبار کاه" هستند که نیاز به استفاده از تکنیک‌های پیشرفته یادگیری ماشین و پردازش داده دارد تا بتوان آن‌ها را به درستی شناسایی کرد.



feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10	feature_11	feature_12	feature_13	feature_14	feature_15
60	468	7.8	1	0	0	0	0	0	0	0	0	0	0	0
108	179	1.6574	1	0	0	0	0	0	0	0	0	0	0	0
1	1	2	0	0	0	0	0	0	0	0	0	0	0	0
60	468	7.8	1	0	0	0	0	0	0	0	0	0	0	0
60	120	2	1	0	0	0	0	0	0	0	0	0	0	0
125	125	1	0	0	0	0	0	0	0	0	0	0	0	0
60	468	7.8	1	0	0	0	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
77	77	1	0	0	0	0	0	0	0	0	0	0	0	0
110	176	1.6	1	0	0	0	0	0	0	0	0	0	0	0
60	468	7.8	1	0	0	0	0	0	1	0	0	0	0	0
77	100	1.2987	0	0	0	0	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
32	259	8.0937	0	0	0	0	0	0	0	0	0	0	0	0
36	140	3.8888	1	0	0	0	0	0	0	0	0	0	0	0
15	15	1	1	0	0	0	0	0	0	0	0	0	0	0
45	100	2.2222	1	0	0	0	0	0	1	0	0	0	0	0
60	468	7.8	1	0	0	0	0	0	0	0	0	0	0	0
48	264	5.5	1	0	0	0	0	0	0	0	0	0	0	0
74	78	1.054	1	0	0	0	0	0	0	0	0	0	0	0
145	148	1.0206	1	0	0	0	0	0	0	0	0	0	0	0
90	65	0.7222	1	0	0	0	0	0	0	0	0	0	0	0
46	75	1.6304	1	0	0	0	0	0	0	0	0	0	0	0

Test

feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10	feature_11	feature_12	feature_13	feature_14	feature_15
100	160	1.6	0	0	0	0	0	0	0	0	0	0	0	0
20	83	4.15	1	0	0	0	0	0	1	0	0	0	0	0
99	150	1.5151	1	0	0	0	0	0	0	0	0	0	0	0
40	40	1	0	0	0	0	0	0	0	0	0	0	0	0
12	234	19.5	1	0	0	0	0	0	0	0	0	0	0	0
90	90	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
15	80	5.3333	0	0	0	0	0	0	0	0	0	0	0	0
100	190	1.9	0	0	0	0	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
59	460	7.7966	1	0	0	0	0	0	0	0	0	0	0	0
18	24	1.3333	0	0	0	0	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
40	46	1.15	1	0	0	0	0	0	0	1	0	0	0	0
23	26	1.1304	1	0	0	0	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
140	140	1	0	0	0	0	0	0	0	0	0	0	0	0
11	64	5.8181	1	0	0	0	0	0	0	0	0	0	0	0
119	152	1.2773	1	0	0	0	0	0	0	0	0	0	0	0
50	50	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
16	130	8.125	1	0	0	0	0	0	0	0	0	0	0	0
60	89	1.4833	1	0	0	0	0	0	0	0	0	0	0	0

Train

شکل ۲. دیتاست تحقیق

دیتاست مورد استفاده در این تحقیق از یکی از تولیدکنندگان بزرگ ویفرهای نیمه‌هادی در هند به دست آمده است (شکل ۲). این داده‌ها به صورت ناشناس ارائه شده‌اند تا از افشای اطلاعات حساس جلوگیری شود. داده‌ها شامل دو فایل داده‌های مقاله و کدهای مربوط به پایتون و تست هستند:

- Train.csv : شامل ۱۷۶۳ نمونه با ۱۵۵۸ ویژگی و یک ستون برای برچسب‌گذاری (صفر برای محصولات خوب و یک برای محصولات معیوب).

- **Test.csv:** شامل ۷۵۶ نمونه بدون ستون برچسب.

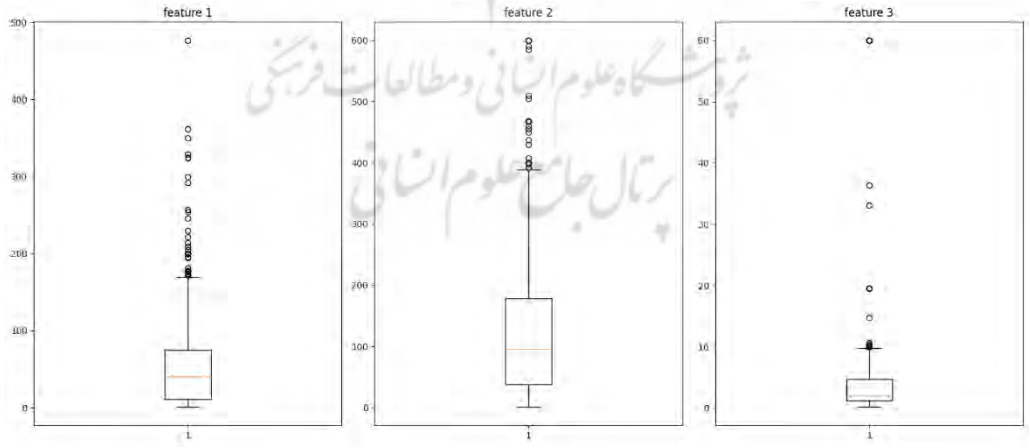
این دیتاست حاوی ویژگی‌های مختلف از ویفرهای تولیدی است که در هر نمونه، پارامترهای مختلفی از فرآیند تولید ثبت شده اند. با توجه به ناشناس بودن ویژگی‌ها، تفسیر دقیق آن‌ها بدون داشتن دانش تخصصی دشوار است.

بررسی ناهماهنگی‌ها و داده‌های پرت: در مرحله‌ی اول، داده‌ها از لحاظ وجود مقادیر ناهماهنگ و پرت مورد بررسی قرار گرفتند. این مقادیر ممکن است به دلیل خطاهای دستگاهی یا عوامل محیطی به وجود آمده باشند و در صورت عدم مدیریت می‌توانند بر عملکرد مدل تأثیر منفی بگذارند.

وینزوریزه کردن: به منظور کاهش اثر داده‌های پرت، از تکنیک وینزوریزه کردن استفاده شد. وینزوریزه کردن مقادیر دور افتاده در ویژگی‌های خاص را به یک محدوده خاص محدود می‌کند تا تأثیر آن‌ها بر داده‌های دیگر کاهش یابد. این تکنیک با تنظیم مقادیر به یک آستانه از پیش تعیین شده، توزیع داده‌ها را بهبود می‌بخشد و مدل را از انحراف در برابر داده‌های غیر طبیعی محافظت می‌کند.

استانداردسازی ویژگی‌ها: به دلیل تفاوت در مقیاس ویژگی‌ها، داده‌ها استانداردسازی شدند تا به مقیاس یکسانی برسند. این مرحله با استفاده از تابع استاندارد[1] از کتابخانه یاد گیری برای ماشین[2] انجام شد که ویژگی‌ها را به میانگین صفر و انحراف معیار یک مقیاس می‌دهد. این کار باعث می‌شود که مدل به اختلافات در مقیاس ویژگی‌ها حساس نباشد و عملکرد بهتری داشته باشد، دلیل نامگذاری این تابع به مقدار واریانس و میانگین آن مربوط می‌شود که برابر با تابع توزیع استاندارد هستند.

مدل‌سازی: مدل‌سازی و ارزیابی (تجسم ویژگی‌های کاهش یافته): برای ارزیابی میزان تفکیک پذیری داده‌ها پس از کاهش بُعد، داده‌ها در فضای بُعد کاهش یافته به صورت گرافیکی تجسم شده است. این تجسم به ما کمک می‌کند که ببینیم آیا کلاس‌های نرمال و ناهنجار به خوبی از یکدیگر جدا شده‌اند یا خیر. نمودار هیستوگرام نیز برای نمایش این جدایی استفاده شده است با توجه به نوع داده‌ها و نیاز به شناسایی ناهنجاری‌ها، مدل XGBoost Classifier برای این پروژه انتخاب شد XGBoost یک مدل یادگیری جمعی[3] است که بر اساس درخت‌های تصمیم گیری کار می‌کند و به دلیل قابلیت بالا در مدیریت داده‌های نامتوازن و شناسایی الگوهای پیچیده، به عنوان انتخاب مناسبی برای این دیتاست در نظر گرفته شد. مدل XGBoost بر روی داده‌ها، بارگزاری شد. به دلیل نامتوازن بودن کلاس‌ها، از تکنیک‌های تنظیم وزن برای نمونه‌های ناهنجار استفاده شد تا مدل بتواند به درستی این نمونه‌ها را تشخیص دهد. این تنظیمات به مدل کمک کرد تا حساسیت بیشتری نسبت به کلاس ناهنجار داشته باشد و دقت شناسایی این کلاس افزایش یابد.



شکل ۳. نمودار جعبه‌ای ویژگی‌های اصلی برای بررسی داده‌های پرت

1. StandardScaler
2. sklearn
3. Ensemble

شکل (۳) شامل سه نمودار جعبه‌ای[1] است که برای بررسی توزیع داده‌ها و شناسایی مقادیر پرت در سه ویژگی مختلف از مجموعه داده‌ها استفاده شده است. هر نمودار نشان دهنده یک ویژگی خاص از مجموعه داده است و اطلاعات زیر را به نمایش می‌گذارد:

- خط وسط جعبه نشان‌دهنده میانه[2] داده‌ها است که وضعیت مرکزی داده را مشخص می‌کند.

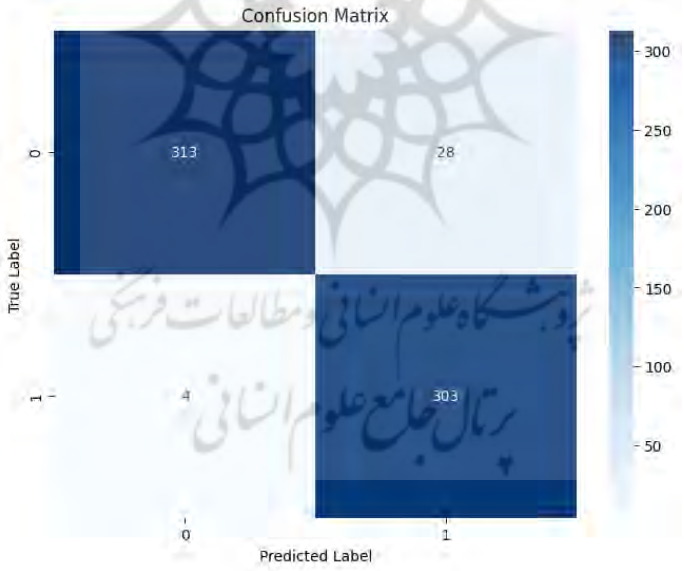
بخش‌های جعبه بین چارک اول (Q1) و چارک سوم (Q3) قرار دارند و نشان‌دهنده گستره ۵۰ درصد میانی داده‌ها هستند.

- خطوط یا "سبیل‌ها" که از جعبه بیرون زده‌اند، محدوده مقادیر غیرپرت داده‌ها را نشان می‌دهند.
- نقاط بیرون از سبیل‌ها نشان‌دهنده مقادیر پرت[3] هستند که به وضوح خارج از محدوده طبیعی داده‌ها قرار دارند.

در تحلیل این بخش می‌توان گفت در هر سه نمودار، حضور مقادیر پرت به خوبی مشخص است. به ویژه در ویژگی‌های ۱ و ۲، تعداد زیادی از نقاط خارج از سبیل‌ها وجود دارد که نشان دهنده مقادیر بسیار بالاتر از سایر داده‌ها هستند. این مقادیر پرت می‌توانند باعث انحراف مدل‌های یادگیری ماشین شوند و باید با دقت مورد بررسی و پردازش قرار گیرند. در ویژگی ۳، هرچند مقادیر پرت کمتری نسبت به ویژگی‌های ۱ و ۲ دیده می‌شود، اما همچنان حضور برخی مقادیر پرت قابل توجه است. این نمودارها به متخصصان کمک می‌کنند تا تصمیمات بهتری در خصوص حذف یا تعدیل مقادیر پرت برای بهبود کیفیت داده‌ها و عملکرد مدل بگیرند.

۳ . تحلیل داده و یافته‌های پژوهش با استفاده از مدل XGBoost

ارزیابی مدل با ماتریس در هم ریختگی[4]: یکی از ابزارهای اصلی برای ارزیابی عملکرد مدل، استفاده از ماتریس در هم ریختگی است. ماتریس در هم ریختگی، تعداد پیش‌بینی‌های صحیح و نا صحیح مدل را برای هر کلاس نشان می‌دهد و اطلاعات دقیقی در مورد تعداد ناهنجاری‌های درست[5] تشخیص داده شده ، ناهنجاری‌های نادرست[6] تشخیص داده شده ، و سایر موارد ارائه می‌دهد. این ماتریس به ما امکان می‌دهد که دقت، حساسیت و ویژگی مدل را محاسبه و تحلیل کنیم.



شکل ۴. ماتریس در هم ریختگی برای ارزیابی عملکرد مدل در تشخیص ناهنجاری‌های تولید ویفر

1. Box Plot
2. Median
3. Outliers
4. Confusion Matrix
5. True Positives
6. False Positives

شکل (۴) ماتریس در هم ریختگی مدل را نشان می‌دهد که برای ارزیابی دقت و عملکرد مدل در طبقه بندی نمونه‌های ناهنجار و سالم در فرآیند تولید ویفر نیمه هادی استفاده شده است. در این ماتریس:

مقدار ۳۱۳ در بالا سمت چپ نشان دهنده تعداد نمونه‌های کلاس نرمال (صفر) است که به درستی به عنوان نرمال پیش‌بینی شده‌اند،(TN[1])

مقدار ۲۸ در بالا سمت راست، تعداد نمونه‌های نرمالی را نشان می‌دهد که به اشتباه به عنوان ناهنجار طبقه بندی شده‌اند، (FP)

مقدار ۴ در پایین سمت چپ، تعداد نمونه‌های ناهنجاری را نشان می‌دهد که به اشتباه به عنوان نرمال تشخیص داده شده‌اند، (FN[2])

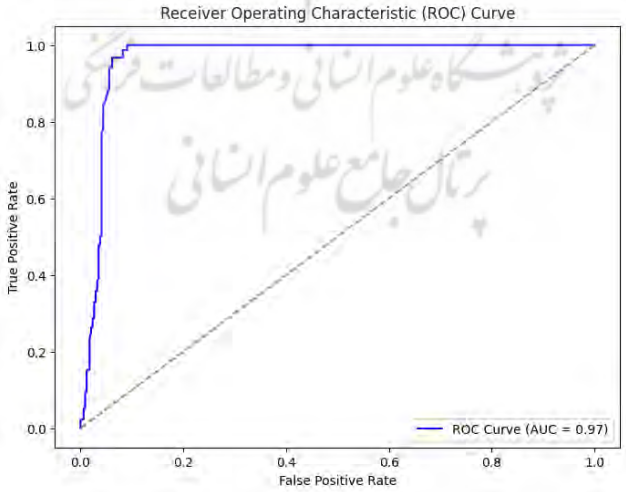
مقدار ۳۰۳ در پایین سمت راست نشان‌دهنده تعداد نمونه‌های ناهنجار است که به درستی به عنوان ناهنجار طبقه بندی شده‌اند،(TP)

در تحلیل کلی این ماتریس نشان می‌دهد که مدل در شناسایی هر دو کلاس عملکرد خوبی داشته است. تعداد کم موارد منفی‌های کاذب و مثبت‌های کاذب نشان می‌دهد که مدل توانسته با دقت بالایی نمونه‌های ناهنجار و سالم را تشخیص دهد. با این حال، ۲۸ مورد منفی کاذب وجود دارد که نشان دهنده نمونه‌های سالمی است که به اشتباه ناهنجار تشخیص داده شده‌اند.

در نگاه دیگر ماتریس سردرگمی، ماتریسی از اعداد است که به ما می‌گوید کجا یک مدل گیج و سردرگم می‌شود. این یک توزیع طبقاتی از عملکرد پیش‌بینی یک مدل طبقه‌بندی است. یعنی ماتریس سردرگمی روشی سازمان‌یافته برای نگاشت پیش‌بینی‌ها به کلاس‌های اصلی است که داده‌ها به آنها تعلق دارند. همچنین نشان می‌دهد که ماتریس‌های سردرگمی تنها زمانی می‌توانند استفاده شوند که توزیع خروجی مشخص باشد، یعنی در چارچوب‌های یادگیری نظارت شده باشد. ماتریس سردرگمی نه تنها امکان محاسبه دقت یک طبقه‌بندی‌کننده را می‌دهد، چه دقت کلی و چه صحت کلاسی، بلکه به محاسبه سایر معیارهای مهمی که توسعه‌دهندگان اغلب برای ارزیابی مدل‌های خود استفاده می‌کنند، کمک می‌کند.

منحنی ویژگی عملیاتی گیرنده (ROC) و معیار AUC :

برای ارزیابی کلی توانایی مدل در شناسایی ناهنجاری‌ها، از منحنی ویژگی عملیاتی گیرنده (ROC) و معیار (AUC) استفاده شد ، ROC ،نموداری است که توانایی مدل را در تفکیک کلاس‌های نرمال و ناهنجار در سطوح مختلف آستانه نشان می‌دهد. معیار AUC (مساحت زیر منحنی) نشان دهنده میزان موفقیت مدل در تفکیک دو کلاس است. مقدار AUC نزدیک به یک (۱) نشان‌دهنده عملکرد عالی مدل است. در این پروژه، مقدار AUC برابر با ۰.۹۷ به دست آمد که این نشان می‌دهد مدل توانایی بسیار خوبی در تشخیص ناهنجاری‌ها دارد.



شکل ۵. منحنی ROC برای ارزیابی دقت مدل در تشخیص ناهنجاری‌های تولید ویفر

1. True Negatives
2. False Negatives

طبق شکل (۵) (ROC)[1] را برای مدل یادگیری ماشین نشان می‌دهد که به ارزیابی عملکرد آن در طبقه بندی نمونه‌ها به ناهنجار و غیر ناهیجاری می‌پردازد. محور افقی (FPR)[2] یا نرخ مثبت کاذب و محور عمودی (TPR)[3] یا نرخ مثبت صحیح را نمایش می‌دهد.

خط آبی نشان دهنده عملکرد مدل در سطوح مختلف آستانه‌های تصمیم‌گیری است. هرچه این منحنی به گوشه بالا–چپ نمودار نزدیک‌تر باشد، مدل عملکرد بهتری در شناسایی کلاس‌های مختلف دارد. مقدار AUC (مساحت زیر منحنی) که در این تصویر برابر با ۰.۹۷ است، معیار کلی از توانایی مدل در تمایز بین کلاس‌ها ارائه می‌دهد. مقدار AUC نزدیک به یک (۱) نشان‌دهنده این است که مدل توانایی بسیار خوبی در تشخیص ناهنجاری‌ها دارد. خط خاکستری نقطه‌چین نشان دهنده عملکرد یک مدل تصادفی است که عملکردی معادل حدس زدن دارد. اینکه منحنی آبی به طور قابل ملاحظه‌ای بالاتر از این خط قرار دارد، تأیید می‌کند که مدل استفاده شده به خوبی توانسته است نمونه‌های سالم و ناهنجار را تفکیک کند.

٤ . تحلیل داده و یافته‌های پژوهش با استفاده از الگوریتم ژنتیک

شناسایی ناهنجاری‌ها در داده‌ها یکی از چالش‌های مهم در حوزه‌های مختلف مانند صنعتی، و تولیدی است. مشکل اصلی تشخیص نمونه‌های سالم از نمونه‌های آسیب دیده در مجموعه داده‌هایی است که از فرآیندهای تولیدی مانند ساخت ویفرهای الکترونیکی به دست آمده‌اند. این مشکل به این دلیل مهم است که در صنایع تولیدی، وجود ناهنجاری‌ها می‌تواند منجر به کاهش کیفیت محصولات، افزایش هزینه‌ها و حتی خطرات ایمنی شود.

روش‌های سنتی تشخیص ناهنجاری، مانند روش‌های آماری ساده، در مواجهه با داده‌های پیچیده و پر ابعاد کارایی خود را از دست می‌دهند. ناهنجاری‌ها ممکن است الگوهای پنهانی داشته باشند که با روش‌های خطی معمولی قابل شناسایی نباشند. همچنین، در محیط‌های واقعی، داده‌ها ممکن است نویزی باشند یا شامل مقادیر گم شده، که این امر پیچیدگی را افزایش می‌دهد. برای حل این مشکل، باید از روش‌های هوشمند یادگیری ماشین استفاده کنیم که بتوانند الگوهای پیچیده را یاد بگیرند و بهینه‌سازی شوند. یکی از رویکردهای موثر، استفاده از الگوریتم‌های تکاملی مانند الگوریتم ژنتیک است، که می‌تواند فضای جستجو را به طور هوشمند کاوش کند و راه حل‌های بهینه پیدا کند. این الگوریتم با شبیه‌سازی فرآیندهای طبیعی تکامل، می‌تواند وزن‌های ویژگی‌ها را تنظیم کند و آستانه‌ای برای طبقه‌بندی تعیین نماید، تا دقت تشخیص را افزایش دهد.

در این الگوریتم، ما با یک جمعیت اولیه از راه‌حل‌های ممکن (که هر کدام یک فرد نامیده می‌شود) شروع می‌کنیم. هر فرد شامل مجموعه‌ای از پارامترها است، مانند وزن‌های ویژگی‌ها و یک آستانه برای طبقه‌بندی. سپس، برای هر فرد، یک تابع تناسب (فیتنس) محاسبه می‌شود که نشان دهنده کیفیت آن راه حل است. در مسأله ما، این تابع می‌تواند بر اساس دقت طبقه‌بندی یا معیارهایی مانند AUC (مساحت زیر منحنی ROC) باشد، که برای مسائل عدم تعادل کلاس‌ها مناسب است.

فرآیند الگوریتم ژنتیک به این صورت است: ابتدا افراد برتر را انتخاب می‌کنیم (انتخاب)، سپس با ترکیب آن‌ها (تقاطع) افراد جدید تولید می‌کنیم، و در نهایت با تغییرات کوچک (جهش) تنوع را حفظ می‌نماییم. این فرآیند در چندین نسل تکرار می‌شود تا به راه حل بهینه برسیم. مزیت الگوریتم ژنتیک این است که می‌تواند در فضای جستجوی بزرگ و پیچیده، بدون نیاز به فرضیات خاص، راه‌حل‌های خوبی پیدا کند. در مسأله تشخیص ناهنجاری، این الگوریتم می‌تواند وزن‌های مناسبی برای ویژگی‌ها پیدا کند تا امتیاز ناهنجاری را محاسبه نماید و با مقایسه با آستانه، نمونه‌ها را به سالم یا آسیب دیده طبقه‌بندی کند.

در الگوریتم ژنتیک این پژوهش ، پارامترهایی مانند اندازه جمعیت (۵۰)، تعداد نسل‌ها (۱۰۰)، نرخ جهش (۰.۰۵) و نرخ تقاطع (۰.۸) تنظیم شده‌اند تا تعادل بین کاوش و بهره‌برداری برقرار شود. نخبه‌گرایی (۵ فرد برتر) تضمین می‌کند بهترین راه‌حل‌ها حفظ شوند. در کد

1. Receiver Operating Characteristic
2. False Positive Rate
3. True Positive Rate

اول، تابع ایجاد فرد وزن‌های تصادفی بین ۱ تا ۱ و آستانه بین ۰ تا ۱ تولید می‌کند. محاسبه امتیاز ناهنجاری با ضرب داخلی انجام می‌شود و نرمال‌سازی به [۰,۱] کمک می‌کند آستانه معنادار باشد.

روش پیاده‌سازی الگوریتم ژنتیک

بخش اول- نتایج جامع

نتایج جامع را بر اساس نمودارهای تشخیصی را ایجاد و ذخیره و برای توزیع اطمینان: هیستوگرام ، خط میانگین، عنوان و برچسب‌ها. رنگ‌آمیزی می‌شوند (سبز برای بالا، نارنجی متوسط، قرمز پایین). توزیع اطمینان را بیان و سریع شناسایی می‌شوند، خط میانگین نشان‌دهنده میانگین کلی اطمینان است.

بخش دوم - توزیع امتیاز ناهنجاری

این بخش هیستوگرام توزیع امتیاز ناهنجاری برای کلاس‌های سالم و آسیب دیده را ایجاد می‌کند. و اگر کلاسی خالی باشد مدیریت می‌شود. هیستوگرام‌ها با شفافیت ۰.۷ و لبه سیاه نمایش داده می‌شوند و عنوان، برچسب‌ها به شبکه اضافه می‌شوند. این نمودار کمک می‌کند آیا آستانه مناسب انتخاب شده است یا خیر. برای مسائل عدم تعادل، مدل چقدر در شناسایی ناهنجاری‌ها موفق است.

بخش سوم - اطمینان بر اساس درستی

هیستوگرام اطمینان برای پیش‌بینی‌های درست و غلط را مقایسه می‌کند. و هیستوگرام‌ها با ۱۵ bin و شفافیت ۰.۸ نمایش داده می‌شوند. این نمودار نشان می‌دهد مدل در پیش‌بینی‌های با اطمینان بالا کمتر اشتباه می‌کند، که برای اعتماد به مدل مفید است. اگر پیش‌بینی‌های غلط عمدتاً در اطمینان پایین باشند، مدل خوب عمل کرده، اما اگر در بالا باشند، نیاز به بهبود دارد

بخش چهارم - دقت بر اساس سطح اطمینان

اطمینان‌ها به سطوح پایین (۰–۰.۴)، متوسط (۰.۴–۰.۷) و بالا (۰.۷–۱) تقسیم می‌شوند . سپس، دقت میانگین برای هر سطح محاسبه می‌شود، دقت در سطوح بالا نزدیک به ۱ است، که کمک می‌کند مدل قابل اعتماد است و می‌توان از اطمینان برای فیلتر پیش‌بینی‌ها استفاده کرد.

بخش پنجم - خلاصه اعتبارسنجی

برای اعتبار سنجی دقت، فراخوانی و F1 محاسبه می‌شوند. توزیع واقعی و پیش‌بینی‌شده کلاس‌ها استخراج می‌شود. متن خلاصه شامل تعداد نمونه‌ها، توزیع‌ها، عملکرد و اطمینان ساخته می‌شود. این بخش خلاصه‌ای متنی و تصویری ارائه می‌دهد که برای گزارش‌گیری سریع مفید است، و کمک می‌کند نقاط قوت و ضعف مدل را در یک نگاه ببینیم.

بخش ششم - نمونه‌های تصادفی آزمون

نمونه‌های تصادفی از اعتبارسنجی انتخاب می‌شوند حداکثر پیش‌بینی، امتیاز و اطمینان برای آن‌ها محاسبه می‌شود. شاخص، برچسب واقعی، پیش‌بینی، امتیاز، اطمینان، سطح و درستی سپس دقت کلی، دقت اطمینان بالا، میانگین اطمینان و تعداد اطمینان بالا محاسبه می‌شود. این متد برای ارزیابی کیفی نمونه‌های خاص مفید است

بخش هفتم – گزارش مدل عمومی

دقت کلی محاسبه می‌شود. میانگین اطمینان، تعداد اطمینان بالا و تعداد نمونه‌های اعتبارسنجی ساخته می‌شود. این بخش خلاصه‌ای فشرده برای استفاده برنامه‌ای ارائه می‌دهد، که برای گزارش‌گیری خودکار و کارآمد برای ادغام با سیستم‌های بزرگ‌تر استفاده کرد.

در این پژوهش ابتدا برای دسترسی بهتر و طبقه بندی داده‌ها از روش تاگوچی برا طبقه بندی داده‌ها استفاده می‌کنیم

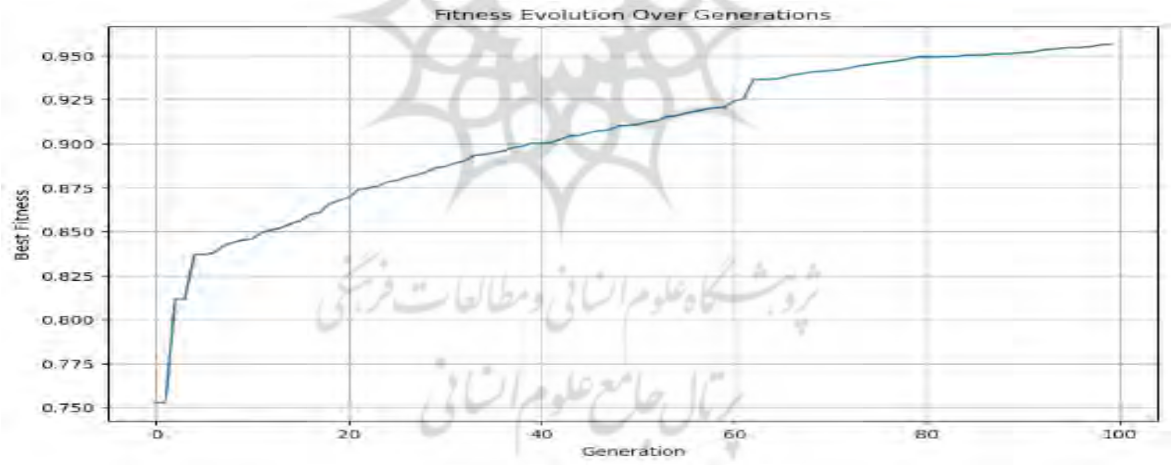
با مراجعه به جدول استاندارد آرایه‌های متعامد در روش تاگوچی و با استفاده از آرایه‌های متعامد L9(34) به عنوان مناسب ترین طرح برای مدل‌های سه تا شش انتخاب می‌شود. آرایه‌های متعامد این طرح در جدول ۲ نشان داده شده است.

جدول ۲ آرایه‌های متعامد $L9(33)$ برای الگوریتم ژنتیک

شماره آزمایش	nPop	Pc	Pm
۱	۱	۱	۱
۲	۱	۲	۲
۳	۱	۳	۳
۴	۲	۱	۲
۵	۲	۲	۳
۶	۲	۳	۱
۷	۳	۱	۳
۸	۳	۲	۱
۹	۳	۳	۲

پس از اجرای کد، نتایج مختلفی به دست می‌آید که نشان‌دهنده عملکرد مدل هستند. ابتدا، تاریخچه تناسب را بررسی می‌کنیم. این نمودار نشان می‌دهد که چگونه بهترین تناسب در هر نسل افزایش می‌یابد. در نسل‌های اولیه، تناسب پایین است، اما با پیشرفت نسل‌ها، به دلیل انتخاب افراد برتر و ایجاد تنوع، تناسب بهبود می‌یابد.

طبق شکل (۶) نمودار تکامل تناسب[1] عملکرد الگوریتم ژنتیک را در طول ۱۰۰ نسل به تصویر می‌کشد. این نمودار با یک افزایش سریع اولیه از تناسب بهترین فرد حدود ۰.۷۵ در نسل صفر آغاز و به سرعت تا حدود نسل ۲۰ به ۰.۸۵ می‌رسد، که نشان‌دهنده انتخاب موثر افراد برتر و بهبود سریع در مراحل اولیه تکامل است.



شکل ۶. ماتریس سردرگمی

سپس، نتایج اعتبارسنجی شامل دقت، گزارش طبقه بندی و ماتریس سردرگمی است. دقت کلی ممکن است بالای ۸۵ درصد باشد، با توجه به اینکه کلاس آسیب‌دیده کمتر است، معیارهایی مانند دقت و فراخوانی برای کلاس ۱ مهم هستند. ماتریس سردرگمی نشان می‌دهد چند نمونه سالم به اشتباه آسیب‌دیده تشخیص داده شده‌اند و بالعکس، که کمک می‌کند نقاط ضعف مدل را شناسایی کنیم.

1. Fitness Evolution Over Generations

در این میان اهمیت ویژگی‌ها نیز استخراج می‌شود که نشان می‌دهد کدام ویژگی‌ها (مانند ویژگی‌های ۱ تا ۱۰) وزن بیشتری دارند و بنابراین در تشخیص ناهنجاری نقش کلیدی ایفا می‌کنند. این لیست به صورت جدول مرتب شده بر اساس اهمیت مطلق وزن‌ها نمایش داده می‌شود.

در بخش تحلیلگر نتایج، چندین نمودار تولید می‌شود. نمودار توزیع اطمینان، هیستوگرامی است که اطمینان پیش‌بینی‌ها را نشان می‌دهد. اطمینان بر اساس فاصله امتیاز از آستانه محاسبه می‌شود و به [۰,۱] نرمال‌سازی می‌گردد. میانگین اطمینان ممکن است حدود ۰.۷ باشد، و نمودار نشان می‌دهد بیشتر پیش‌بینی‌ها در سطوح بالا یا متوسط قرار دارند، که نشانه خوبی از اعتماد مدل است.

نمودار توزیع امتیاز ناهنجاری برای کلاس‌های سالم و آسیب دیده، هیستوگرام‌های جداگانه‌ای را نشان می‌دهد. امتیازهای کلاس آسیب دیده معمولاً بالاتر از آستانه هستند، در حالی که امتیازهای کلاس سالم پایین‌تر. خط آستانه در نمودار مشخص است و کمک می‌کند ببینیم چقدر جداسازی خوب انجام شده.

نمودار اطمینان بر اساس درستی پیش‌بینی، هیستوگرام اطمینان برای پیش‌بینی‌های درست و غلط را مقایسه می‌کند. پیش‌بینی‌های غلط اغلب اطمینان کمتری دارند، که نشان می‌دهد مدل در موارد نامطمئن اشتباه می‌کند.

نمودار دقت بر اساس سطح اطمینان (پایین، متوسط، بالا)، میله‌هایی را نشان می‌دهد که دقت در سطوح بالا نزدیک به ۱ است، در حالی که در سطوح پایین کمتر. این کمک می‌کند بفهمیم مدل در پیش‌بینی‌های با اطمینان بالا قابل اعتماد است.

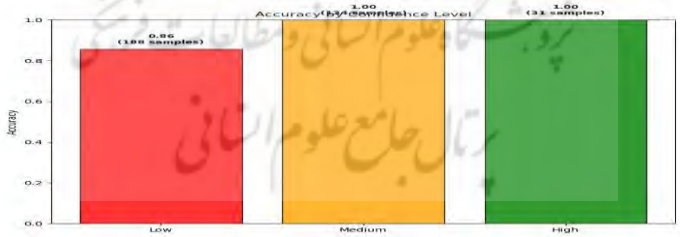
در نهایت، خلاصه اعتبارسنجی به صورت متن و نمودار نمایش داده می‌شود، شامل توزیع واقعی و پیش‌بینی‌شده کلاس‌ها، دقت، دقت، فراخوانی و امتیاز F1. برای مثال، اگر ۸۰ درصد نمونه‌ها سالم باشند، مدل ممکن است ۷۵ درصد را درست پیش‌بینی کند، با میانگین اطمینان ۰.۶۵ می‌باشد.

این نتایج نشان می‌دهند که الگوریتم ژنتیک توانسته است مدل موثری برای تشخیص ناهنجاری بسازد، و استفاده از کدهای جداگانه برای آموزش و آزمایش، کارایی را افزایش می‌دهد.

در تحلیل نتایج، با توجه به خلاصه اعتبارسنجی ارائه شده، تعداد کل نمونه‌ها ۳۵۳ است.

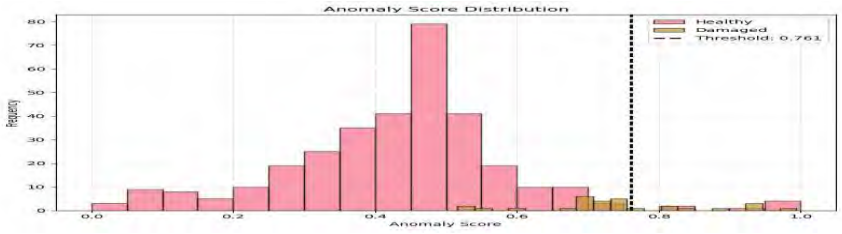
توزیع واقعی: سالم ۳۲۴ (۹۱.۸٪)، آسیب‌دیده ۲۹ (۸.۲٪). توزیع پیش‌بینی‌شده: سالم ۳۳۷ (۹۵.۵٪)، آسیب‌دیده ۱۶ (۴.۵٪). عملکرد: دقت ۰.۹۲۴ (۹۲.۴٪)،دقت ۰.۹۱۰، فراخوانی ۰.۹۲۴، امتیاز F1- ۰.۹۱۳ است.

اطمینان: میانگین ۰.۴۱۳، تعداد اطمینان بالا (>۰.۷): ۳۱ نمونه است.



شکل ۷. دقت بر اساس سطوح

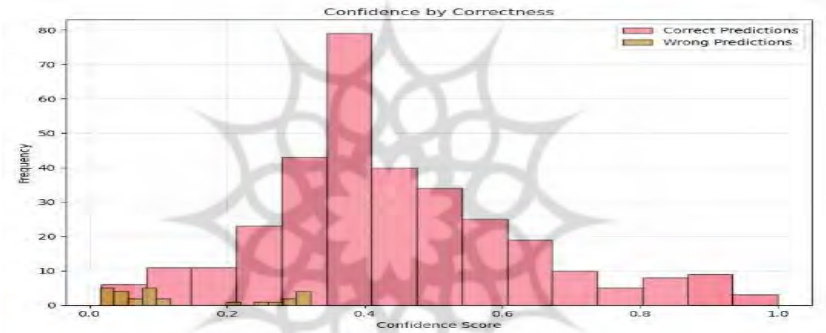
این آمار نشان دهنده عملکرد قوی مدل در برابر عدم تعادل کلاس‌ها است، هر چند پیش‌بینی آسیب دیده کمتر از واقعی است که ممکن است به دلیل تمرکز بر کاهش خطاهای مثبت کاذب باشد.



شکل ۸. توزیع امتیاز ناهنجاری برای نمونه‌های (سالم صورتی و آسیب دیده قهوه ای)

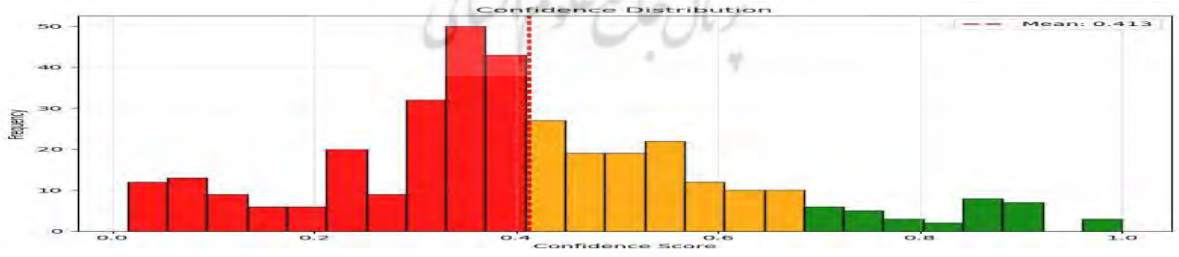
نمودار (شکل ۷) دقت را بر اساس سطوح اطمینان نشان می‌دهد. در سطح پایین (۰.۸۶ با ۱۸۸ نمونه)، متوسط (۱.۰۰ با ۱۳۱ نمونه) و بالا (۱.۰۰ با ۳۱ نمونه). این نشان می‌دهد که مدل در پیش‌بینی‌های با اطمینان متوسط و بالا کاملاً دقیق است، اما در سطوح پایین کمی کاهش دارد، که منطقی است و می‌توان از آن برای فیلتر کردن پیش‌بینی‌های نامطمئن استفاده کرد.

توزیع امتیاز ناهنجاری برای نمونه‌های سالم (صورتی) و آسیب دیده (قهوه‌ای) با آستانه ۰.۷۶۱. بیشتر نمونه‌های سالم در امتیازهای پایین‌تر متمرکز هستند، در حالی که آسیب دیده‌ها بالاتر از آستانه قرار دارند. این جداسازی خوب نشان دهنده کارایی آستانه در طبقه‌بندی است، هر چند برخی هم‌پوشانی در نواحی میانی وجود دارد که ممکن است منبع خطاها باشد.



شکل ۹. توزیع اطمینان برای پیش‌بینی‌های درست (صورتی) و غلط (قهوه‌ای)

توزیع اطمینان برای پیش‌بینی‌های درست (صورتی) و غلط (قهوه‌ای). پیش‌بینی‌های درست عمدتاً در اطمینان‌های بالاتر هستند، در حالی که غلط‌ها در پایین‌تر متمرکز هستند. این الگو مثبت است و نشان می‌دهد مدل اشتباهات را عمدتاً در موارد نامطمئن مرتکب می‌شود، که می‌توان با تنظیم آستانه اطمینان بهبود بخشید.



شکل ۱۰ توزیع اطمینان کلی

توزیع اطمینان کلی با میانگین ۰.۴۱۳ و رنگ بندی سطوح (قرمز پایین، نارنجی متوسط، سبز بالا). توزیع به سمت اطمینان‌های پایین‌تر تمایل دارد، اما وجود پیک در نواحی متوسط و بالا نشان دهنده اعتماد کلی مدل است. میانگین پایین ممکن است به دلیل طبیعت محافظه‌کارانه مدل در برابر ناهنجاری‌های نادر باشد.

جدول ۳. خلاصه اعتبارسنجی با استفاده از الگوریتم ژنتیک

معیار	مقدار
تعداد کل نمونه‌ها	۳۵۳
توزیع واقعی – سالم	۳۲۴ (۹۱.۸٪)
توزیع واقعی – آسیب‌دیده	۲۹ (۸.۲٪)
توزیع پیش‌بینی‌شده – سالم	۳۳۷ (۹۵.۵٪)
توزیع پیش‌بینی‌شده – آسیب‌دیده	۱۶ (۴.۵٪)
دقت(Accuracy)	۰.۹۲۴ (۹۲.۴٪)
Precision	۰.۹۱۰
Recall	۰.۹۲۴
F1-Score	۰.۹۱۳
میانگین اطمینان	۰.۴۱۳
تعداد اطمینان بالا(>0.7)	۳۱

این جدول خلاصه‌ای از عملکرد مدل ارائه می‌دهد و نشان دهنده تعادل خوب بین معیارها است، هرچند تمرکز بیشتر بر کلاس غالب (سالم) مشاهده می‌شود.

۵. نتیجه‌گیری و پیشنهادها

نتایج نشان داد که مدل XGBoost به خوبی قادر به شناسایی ناهنجاری‌ها در تولید با دقت بالااست و مقدار بالای AUC تأیید می‌کند که مدل توانسته است با دقت بالا محصولات معیوب را شناسایی کرده و از ورود آن‌ها به چرخه تولید جلوگیری کند. این روش می‌تواند به صنعت تولید کمک کند تا با استفاده از تکنیک‌های یادگیری ماشین و داده‌کاوی، کیفیت محصولات خود را بهبود بخشد و از تولید محصولات معیوب جلوگیری کند.

هدف اصلی شناسایی ناهنجاری‌ها در فرآیند تولید با استفاده از مدلXGBoos و الگوریتم ژنتیک که عنوان یک رویکرد تکاملی برای تنظیم وزن ویژگی‌ها و آستانه طبقه‌بندی استفاده شد است. پس از آماده‌سازی داده‌ها و حذف یا تعدیل مقادیر پرت، مدل XGBoost برای طبقه بندی داده‌ها به دو کلاس ناهنجار و سالم به کار گرفته شده است.

برای ارزیابی عملکرد مدل، از معیارهایی مانند ماتریس در هم ریختگی و منحنی ROC استفاده شد. نتایج نشان داد که مدل با دقت بالایی قادر به شناسایی ناهنجاری‌ها است و مقدار AUC برابر با ۰.۹۷ به دست آمد. این نتیجه نشان‌دهنده توانایی بالای مدل در تشخیص صحیح نمونه‌های ناهنجار و سالم است و بیانگر این است که مدل پیشنهادی می‌تواند به صنعت تولید نیمه هادی کمک کند تا محصولات معیوب را شناسایی و از ورود آن‌ها به بازار جلوگیری نماید.

باید گفت در زمینه تولید، ناهنجاری‌ها صرفاً نشانه‌هایی هستند که نشان می‌دهد داده‌ها متفاوت از آنچه معمولاً انجام می‌دهند شروع به عمل کرده‌اند. ناهنجاری‌ها می‌توانند پیامدهای مثبت و منفی داشته باشند. مدل‌های تشخیص ناهنجاری نیز کاملاً منعطف و سازگار هستند. آنها می‌توانند تغییرات ظریف یا تدریجی در داده‌ها را با دقت بیشتری نسبت به روش‌های سنتی مانند SPC مشخص کنند. مدل‌های تشخیص ناهنجاری به جای تکیه بر محدودیت‌های از پیش تعیین شده مانند SPC، می‌توانند خطاهای تعریف نشده قبلی را کشف کنند. به این دلایل، تشخیص ناهنجاری زمانی مورد استفاده قرار می‌گیرد که روش‌های دیگر برای حل مسائل چالش برانگیز تولید ناموفق بوده‌اند.

مدل‌های تشخیص ناهنجاری در تولید به سه دسته تشخیص ناهنجاری تحت نظارت و تشخیص ناهنجاری بدون نظارت و ناهنجاری در داده‌های سری زمانی تقسیم می‌شوند.

تشخیص ناهنجاری تحت نظارت شامل آموزش مدلی بر روی یک مجموعه داده است که در آن نمونه‌ها به‌عنوان «عادی» یا «غیر عادی» برچسب‌گذاری می‌شوند. هدف این است که مدل ویژگی‌های نقاط داده عادی و غیرعادی را بیاموزد تا بتواند نمونه‌های دیده نشده

را به طور دقیق طبقه بندی کند اما در تشخیص ناهنجاری بدون نظارت نوعی یادگیری ماشینی است که شامل آموزش مدلی بر روی داده‌ها بدون برچسب یا نتایج از پیش تعریف شده است.

تحلیل ماتریس در هم ریختگی نیز نشان داد که مدل در شناسایی ناهنجاری‌ها موفق عمل کرده و تعداد کمی از نمونه‌ها به اشتباه طبقه بندی شده‌اند. این موفقیت به دلیل استفاده از تکنیک‌های پردازش داده‌های پرت و استانداردسازی است.

با وجود نتایج مثبت، هنوز امکان بهبود و پیشرفت در این پژوهش وجود دارد. چند پیشنهاد برای پژوهش‌های آینده به شرح زیر است:

- **استفاده از روش‌های پیچیده‌تر یادگیری** عمیق[1] استفاده از شبکه‌های عصبی عمیق می‌تواند به شناسایی بهتر الگوهای پیچیده در داده‌های با ابعاد زیاد کمک کند. شبکه‌های عصبی کانولوشن (CNN) و شبکه‌های عصبی بازگشتی (RNN) می‌توانند در این زمینه کارایی بالاتری از خود نشان دهند.
- **افزایش تعداد داده‌های آموزشی**: از آنجا که ناهنجاری‌ها در مجموعه داده کم هستند، جمع‌آوری داده‌های بیشتر و یا تولید داده‌های مصنوعی با استفاده از تکنیک‌های مختلف می‌تواند دقت مدل را بهبود بخشد.
- **استفاده از تکنیک‌های ترکیبی یادگیری** ماشین[2] ترکیب مدل‌های مختلف یادگیری ماشین مانند جنگل تصادفی[3] یا تولید می‌تواند به بهبود عملکرد و کاهش خطاهای مدل کمک کند و توانایی مدل را در تشخیص ناهنجاری‌ها افزایش دهد.
- **تحلیل دقیق‌تر نتایج با معیارهای مختلف ارزیابی**: استفاده از معیارهای ارزیابی دیگر مانند دقت کلی، حساسیت و ویژگی، به تحلیل عمیق‌تر عملکرد مدل و یافتن نقاط قوت و ضعف آن کمک می‌کند.
- **استفاده از تکنیک‌های مدرن‌تر**، استفاده از روش مدرن دقت و کارایی مدل را بهبود داده و قابلیت استفاده از آن در صنایع دیگر راامکان پذیر می‌سازد

در این پژوهش استفاده از روش XGBoost توانایی بالایی در تشخیص ناهنجاری‌ها دارد. استفاده از این مدل در شرایطی که داده‌ها به شدت نامتوازن هستند و ناهنجاری‌ها بسیار کمتر از نمونه‌های سالم‌اند، کمک شایانی به افزایش دقت شناسایی کرده است. استفاده از الگوریتم ژنتیک که عنوان یک رویکرد تکاملی برای تنظیم وزن ویژگی‌ها و آستانه طبقه‌بندی در این پژوهش استفاده شد، پس از اجرای کد، نتایج مختلفی به دست آمد که نشان‌دهنده عملکرد مدل هستند مدل توانسته است با دقت بالا و با نرخ پایین خطا، نمونه‌های ناهنجار را از سالم تفکیک کند.

در نهایت، این پژوهش نشان داد که با استفاده از تکنیک‌های مناسب پردازش داده و یادگیری ماشین، می‌توان به نتایج قابل توجهی در شناسایی ناهنجاری‌های تولید دست یافت. در این میان استفاده از الگوریتم ژنتیک که عنوان یک رویکرد تکاملی برای تنظیم وزن ویژگی‌ها و آستانه طبقه‌بندی در این پژوهش استفاده شد، پس از اجرای کد، نتایج مختلفی به دست آمد که نشان‌دهنده عملکرد مدل هستند. و نمودارها نشان دادند که چگونه بهترین تناسب در هر نسل افزایش می‌یابد. در نسل‌های اولیه، تناسب پایین بوده، اما با پیشرفت نسل‌ها، به دلیل انتخاب افراد برتر و ایجاد تنوع، تناسب بهبود می‌یابد. این روند کار بیانگر کارایی الگوریتم در کاوش فضای جستجو و دست یابی به بهینه‌سازی پایدار است، جایی که بهبودهای اولیه بزرگ و ناشی از تنوع جمعیت هستند، در حالی که مراحل پایانی بر تنظیم دقیق وزن‌ها و آستانه تمرکز دارند و از خطر محلی ماندن جلوگیری می‌کنند و در نهایت نتایج اعتبارسنجی روش ژنتیک با دقت کلی ۸۵ درصد را نشان داد و نتایج مدلXGBoost نشان داد که مدل فوق با دقت بالایی قادر به شناسایی ناهنجاری‌ها است و مقدار AUC برابر با ۰.۹۷ به دست آمد پس بنابر این مدل فوق کارائی بالاتری دارد.

1.Deep Learning
2. Ensemble Methods
3. Random Forest

تعارض منافع. برای ارائه مطالب و نگارش این مقاله هیچ گونه کمک مالی از هیچ فرد، نهاد و سازمانی دریافت نشده است و نتایج حاصل این مقاله به نفع یا ضرر سازمان یا فردی خاص نخواهد بود. حضور نویسندگان در این پژوهش به عنوان شاهدی بی‌طرف ولی متخصص بوده است و نویسندگان هیچ گونه تعارض منافعی ندارند.

منابع

1. Antonini, M., Pincheira, M., Vecchio, M., & Antonelli, F. (2023). An adaptable and unsupervised TinyML anomaly detection system for extreme industrial environments. *Sensors*, 23(4), 2344. MDPI AG. https://doi.org/10.3390/s23042344
2. Choi, H., Kim, D., Kim, J., Kim, J., & Kang, P. (2022). Explainable anomaly detection framework for predictive maintenance in manufacturing systems. *Applied Soft Computing*, 125, 109147. Elsevier BV. https://doi.org/10.1016/j.asoc.2022.109147
3. Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey (Version 2). *arXiv*. https://doi.org/10.48550/arXiv.1901.03407
4. Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017). A particle swarm optimization algorithm to mixed-model two-sided assembly line balancing. *Journal of Intelligent Manufacturing*, 28(1), 23–36.
5. Kamran, K., & Behnamian, J. (2023). Unrelated parallel machine scheduling with sequence-dependent setup times in multi-factory production network: Modeling and algorithm. *Industrial Management Perspective*, 13(3), 223–248. https://doi.org/10.48308/JIMP.13.3.223 (In Persian)
6. Lee, K. S., Kim, S. B., & Kim, H.-W. (2023). Enhanced anomaly detection in manufacturing processes through hybrid deep learning techniques. *IEEE Access*, 11, 93368–93380. https://doi.org/10.1109/ACCESS.2023.3308698
7. Kiangala, S. K., & Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting–XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications*, 4, 100024. Elsevier BV. https://doi.org/10.1016/j.mlwa.2021.100024
8. Lu, L., Zhang, Y., Si, Z., & Dou, Z. (2024). Research on anomaly detection of parts in workshop production line based on BO-XGBoostLSS. *Lecture Notes in Computer Science*, 1291. Springer Nature Singapore.
9. Liu, J. (2024). Predicting Chinese stock market using XGBoost multi-objective optimization with optimal weighting. *PeerJ Computer Science*, 10, e1931. PeerJ. https://doi.org/10.7717/peerj-cs.1931
10. Leng, J., Chen, Z., Sha, W., Lin, Z., Lin, J., & Liu, Q. (2022). Digital twins-based flexible operating of open architecture production line for individualized manufacturing. *Advanced Engineering Informatics*, 53, 101676. Elsevier BV. https://doi.org/10.1016/j.aei.2022.101676
11. Lee, D., Kim, C.-K., Yang, J., Cho, K.-Y., Choi, J., Noh, S.-D., & Nam, S. (2022). Digital twin-based analysis and optimization for design and planning of production lines. *Machines*, 10(12), 1147. MDPI AG. https://doi.org/10.3390/machines10121147
12. Liu, C., He, Y., Wang, Y., Li, Y., Wang, S., Wang, L., & Wang, Y. (2020). Effects of process parameters on cutting temperature in dry machining of ball screw. *ISA Transactions*, 101.
13. Loh, C.-H., Chen, Y.-C., & Su, C.-T. (2024). Using transfer learning and radial basis function deep neural network feature extraction to upgrade existing product fault detection systems for Industry 4.0. *Electronics*, 14(7), 2913. https://doi.org/10.3390/electronics14072913
14. Li, Z., Kucukkoc, I., & Tang, Q. (2017). New MILP model and station-oriented ant colony optimization algorithm for balancing U-type assembly lines. *Computers & Industrial Engineering*.
15. Moslemipour, G., & Ghadirpour, S. M. (2021). Intelligent design of a dynamic facility layout in the stochastic environment of flexible manufacturing systems considering routing flexibility. *Journal of Industrial Management Perspective*, 11(1), 175–209. https://doi.org/10.52547/jimp.11.1.175 (In Persian)
16. Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Artificial intelligence applications for Industry 4.0: A literature-based study. *Journal of Industrial Integration and Management*, 7(1), 83–111.
17. Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. (2021). Forecasting and anomaly detection approaches using LSTM and LSTM autoencoder techniques with the applications in supply chain

management. *International Journal of Information Management*, 57, 102282. Elsevier BV. https://doi.org/10.1016/j.ijinfomgt.2020.102282
18. Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2021). Deep learning for anomaly detection. *ACM Computing Surveys*, 54(2), 1–38. ACM. https://doi.org/10.1145/3439950
19. Park, K. T., Yang, J., & Noh, S. D. (2020). VREDI: Virtual representation for a digital twin application in a work-center-level asset administration shell. *Journal of Intelligent Manufacturing*, 32(2), 501–544. https://doi.org/10.1007/s10845-020-01586-x
20. Rousopoulou, V., Vafeiadis, T., Nizamis, A., Iakovidis, I., Samaras, L., Kirtsoglou, A., Georgiadis, K., Ioannidis, D., & Tzovaras, D. (2022). Cognitive analytics platform with AI solutions for anomaly detection. *Computers in Industry*, 134, 103555. Elsevier BV. https://doi.org/10.1016/j.compind.2021.103555
21. Rawat, A. (2020). A review on Python programming. *International Journal of Research in Engineering, Science and Management*, 3(12), 8–11.
22. Sadeghi, H., Farughi, H., Kalevandi, F., & Solgi, M. (2023). Production planning system with variable demand and stochastic machine breakdown. *Industrial Management Perspective*, 13(3), 93–126. https://doi.org/10.48308/JIMP.13.3.93 (In Persian)
23. Shi, H., Cao, G., Ma, G., Duan, J., Bai, J., & Meng, X. (2022). New progress in artificial intelligence algorithm research based on big data processing of IoT systems on intelligent production lines. *Computational Intelligence and Neuroscience*, 2022, 1–12. Hindawi. https://doi.org/10.1155/2022/3283165
24. Phuyal, S., Bista, D., & Bista, R. (2020). Challenges, opportunities and future directions of smart manufacturing: A state-of-the-art review. *Sustainable Futures*.
25. Yang, S., Feng, M., & Guan, D. (2022). Intelligent scheduling system for production line automatic matching based on DSSM-XGBoost. *Journal of Physics: Conference Series*, 2203, 012072. https://doi.org/10.1088/1742-6596/2203/1/012072
26. Shi, X., Xiao, Y., Mei, X., Tao, T., & Wang, H. (2023). Thermal error modeling of machine tool based on dimensional error of machined parts in automatic production line. *ISA Transactions*, 135, 575–584. https://doi.org/10.1016/j.isatra.2022.09.043
27. Sankhye, S., & Hu, G. (2020). Machine learning methods for quality prediction in production. *Logistics*, 4(4), 35. https://doi.org/10.3390/logistics4040035
28. Sobhi Shoje, M., & Smuee, P. (2015). Presenting a stable approach for the robotic assembly line sequence balance problem considering robot failures. *15th International Industrial Engineering Conference*, Yazd University, Yazd.
29. Srinath, K. R. (2017). Python—the fastest growing programming language. *International Research Journal of Engineering and Technology*, 4(12), 354–357.
30. Soller, S., Kranz, M., & Hoelzl, G. (2020). Adaptive error prediction for production lines with unknown dependencies. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics* (pp. 227–234). ACM. https://doi.org/10.1145/3405962.3405994
31. Usuga Cadavid, J. P., Lamouri, S., Grabot, B., Pellerin, R., & Fortin, A. (2020). Machine learning applied in production planning and control: A state-of-the-art in the era of Industry 4.0. *Journal of Intelligent Manufacturing*, 31(6), 1531–1558. https://doi.org/10.1007/s10845-019-01531-7
32. Wang, Y., Perry, M., Whitlock, D., & Sutherland, J. W. (2022). Detecting anomalies in time series data from a manufacturing system using recurrent neural networks. *Journal of Manufacturing Systems*, 62, 823–834. Elsevier BV. https://doi.org/10.1016/j.jmsy.2020.12.007
33. Wagner, R., Fischer, J., Gauder, D., Haefner, B., & Lanza, G. (2020). Virtual in-line inspection for function verification in serial production by means of artificial intelligence. *Procedia CIRP*, 92, 63–68. Elsevier BV. https://doi.org/10.1016/j.procir.2020.03.126
34. Yang, S., Feng, M., & Guan, D. (2022). Intelligent scheduling system for production line automatic matching based on DSSM-XGBoost. *Journal of Physics: Conference Series*, 2203(1), 012072. https://doi.org/10.1088/1742-6596/2203/1/012072

35. Zhang, W., Xu, W., Liu, G., & Gen, M. (2017). An effective hybrid evolutionary algorithm for stochastic multiobjective assembly line balancing problem. *Journal of Intelligent Manufacturing*, 28(3), 783–791.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی