INNOVATION IN SCIENCE AND TECHNOLOGY

ISSN 2788-7030 MAR. 2026 VOL. 5, NO. 1

CONTENTS

Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
MAR. 2026 VOL.5, NO.1

Quench and Partitioning Heat Treatment to Improve the Ductility of Ultra High Strength Steel

J. N. Mohapatra[1] & D. Satish Kumar[1]

[1] JSW Steel Ltd., Toranagallu, Bellary, Karnataka 583275, India

Correspondence: J. N. Mohapatra, JSW Steel Ltd., Toranagallu, Bellary, Karnataka 583275, India.

doi:10.63593/IST.2788-7030.2026.03.001

Abstract

Quench & Partitioning (Q&P) heat treatment was carried out on an ultra high strength steel with the Q&P temperatures below the Ms-Temperature. It was found that by increasing the single stage Q&P temperature resulted in decrease in yield strength, ultimate tensile strength and hardness with marginal increase in the total elongation of the steel where no retained austenite is found. With the double stage Q&P heat treatment a very marginal decrease in the ultimate tensile strength found with a significant increase in its total elongation due to presence of 3-5% retained austenite detected by XRD. Hence, double stage Q&P heat treatment can be an effective method to improve the ductility of ultra high strength steels by stabilizing the retained austenite which give the TRIP effect while deformation.

Keywords: quench and partitioning, ultra high strength steel, improved ductility, microstructure

1. Introduction

Quench and partitioning (Q&P) steels have similar chemical composition to TRIP assisted steels with the excellent combination of strength and elongation suitable for automotive applications. In this process the steel is subjected to quenching in a temperature range below the martensite start temperature after austenitization above A_3 temperature or after inter critical annealing, the steel may be either held at the same quenching temperature called as single stage Q&P or the steel temperature is raised above the quench temperature either in the martensite regime or above the martensite regime to the bainitic regime followed by water quenching or air cooling called as second stage Q&P to achieve attractive level of strength and elongations in the third generation AHSS regime suitable for automotive applications (David K Matlock et al., 2003; Hana Jirkova et al., 2012; Wang Li et al., 2013; G. A. Thomas & J. G. Speer, 2014; Yuki Toji et al., 2014; Emmanuel De Moor et al., 2017; Li, Z. et al., 2021; Y.Y. Cheng et al., 2022; Xu, Y. et al., 2022; Yuki Toji et al., 2023; Christian Illgen et al., 2023; Evgeniy Tkachev et al., 2023; Pengsheng Hu et al., 2023; Roman Mishnev et al., 2023). In the process of partitioning the carbon from the martensite is rejected to the surrounding to be become relatively soft martensite whereas the rejected carbon is captured by the retained austenite to stabilize at room temperature. Various alloying elements play a vital role in the carbon partitioning and stabilizing the retained austenite in addition to the grain refinement and morphological changes in the microstructure to give the unique characteristics microstructure and mechanical properties to the steel. Effect of C & Mn-Carbon and Mn stabilizes the retained austenite in the Q&P steel also imparts the morphology of the microstructure in the steel (Huan Xiao et al., 2022; Emmanuel De Moor et al., 2011). Effect of Si-Si suppresses the cementite and carbide formation leading to availability of more rejected/partitioned carbon from the martensite to the retained austenite to stabilize it (S Jenicek et al., 2017). Effect of Cr & Mo and V-Q&P heat treatments in Cr-Mo steel with less Si and Al elements presents an obvious characteristic of tempered martensite, while high-Silicon and high-aluminum steels with more Si or Al contain more residual austenite and Cr, Mo, V further gives more stable austenite (Wen-hua Xu et al., 2023; Roman A. Kussa et al., 2022). Effect of Nb & Ti-grain boundary pinning by precipitates and Nb solute

drag effects refine the austenite grain size during the hot-rolling process to give refined in the final microstructure of Q&P steel. The remaining supersaturated Nb suppresses the bainite formation and decreases the final bainite fraction formed in the Q&P process. Ti too helps in grain refinement by the precipitation of fine carbides and carbo nitrides (Zhisong Chai et al., 2021; Ji Dong et al., 2017).

Effect of Cu-Combination of nanosized Cu-rich precipitates and ultrafine microstructure through the addition of Cu can be a highly potential method to improve the mechanical properties of Q&P steel (Xu Wang et al., 2022).

2. Experimental

The steel used in the present study was designed and induction melted followed by hot deformation to 70mm*70mm billet size to further processing. The chemical composition of the steel was obtained using SPECTRO optical emission spectroscopy. The hot deformed billet was then sliced to 2mm thick sheets in an EDM machine for the further experiments. JMat Pro software was used to construct the CCT diagram and the critical temperatures such as A_3, A_1, B_s and M_s and found to be 843, 743, 498 and 328^{o}C respectively. The experiment involved holding the samples above A_3 temperature for 5 min followed by quenching at 150, 200, 250^{o}C for 5min for the single stage Q&P and the 200 and 250^{o}C samples were again partitioned at 300^{o}C for 5min for the double stage Q&P and then water quenched. The heat-treated samples were then subjected to standard metallography for microstructural observation using a Hitachi scanning electron microscope (SEM). Subsize standard (ASTM E8) tensile specimens were made from the as-hot-deformed and heat-treated samples for the tensile test on a Zwick/Roell make 250 kN universal tensile testing machine at a standard strain rate of 0.008/s for the evaluation of mechanical properties.

3. Results and Discussion

Chemical composition of the steel obtained through OES is shown in Table 1. The steel is having (Wt. %) 0.2C-2Mn-1.7Si-0.9Cr-0.18Mo-0.07Nb-0.04Ti-0.2Cu-0.05Ni-0.025Al-0.002B-0.04S-0.06P. The steel is a low carbon low alloy steel suitable for automotive applications. C and Mn stabilizes the retained austenite in the steel whereas Si helps in retarding the cementite formation leading to additional austenite stabilization. Cr, Mo and V give strength to the steel and Nb, Ti with the fine precipitates helps in grain refinement in the steel. The Cu with the nano size precipitate further improves the strength of the steel.

Table 1. Chemical composition of the steel (wt. %)

C	Mn	S	P	Si	Al	Cr	Ni	Cu	Nb	V	Ti	Mo	B
0.213	2.07	0.037	0.056	1.71	0.025	0.899	0.045	0.213	0.069	0.005	0.038	0.176	0.0017

The CCT diagram of steel with the critical temperatures and heat treatment cycle is shown in Figure 1.

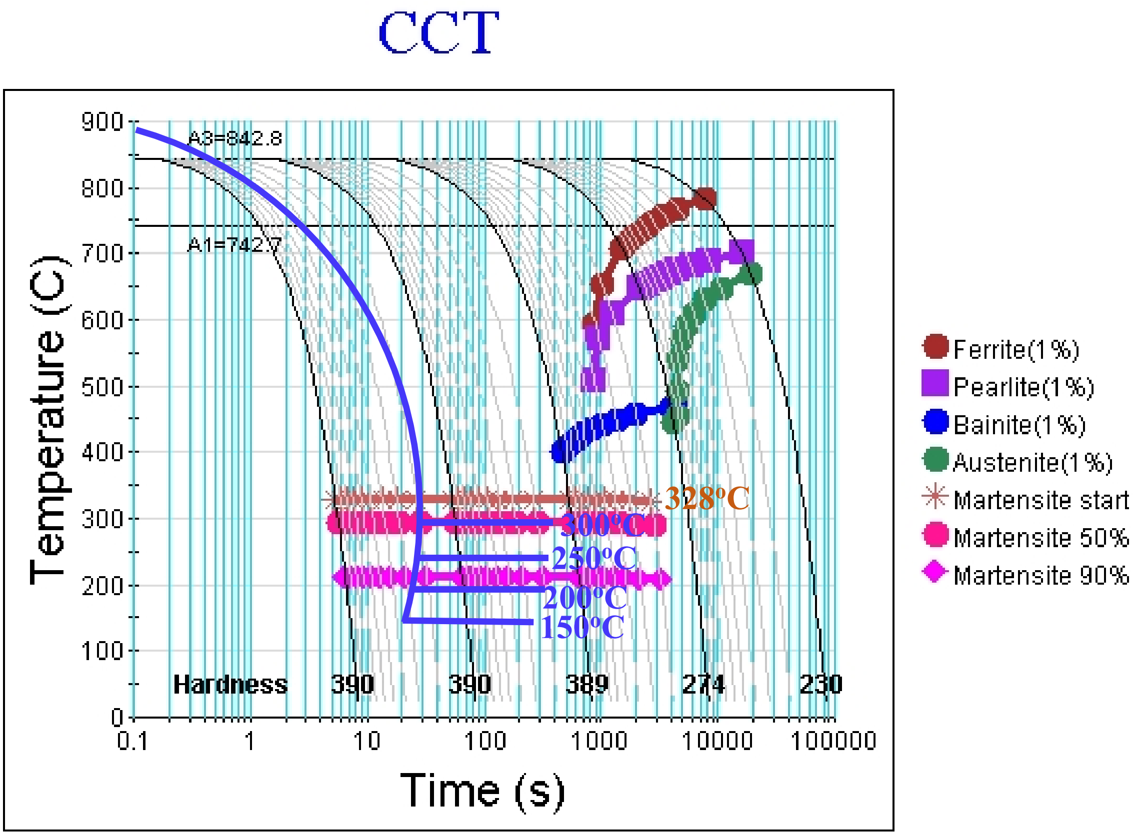

Figure 1. CCT diagram of the steel with the critical temperature and heat treatment cycle

The low and high magnification SEM micrograph of the steel austenitized at 880ºC followed by salt bath quenching and partitioning at 150, 200 and 250ºC and then water quenched is shown in Figure 2. The volume fraction of martensite at each quenched temperature can be evaluated through Koistinen-Marburger empirical equation.

$$f=1-[-0.011\,(Ms-T)] \quad \ldots\ldots (1)$$

For the present steel Ms =327.9ºC. Hence, the volume fraction of martensite

For 150ºC, f = 0.86

For 200ºC, f = 0.75

For 250ºC, f = 0.58

Hence, the volume fraction of martensite at the temperature of 150, 200 and 250ºC are 86%, 75% and 58% respectively.

The stress-strain diagram of the single stage and double stage quench and partitioning steel is shown in Figure 3. The change in mechanical properties of the steel with increase in quench and partitioning temperature is shown in Figure 4. The yield strength (YS), ultimate tensile strength (UTS) and hardness (HV) were found to decrease with a marginal increase in the total elongation of the steel with the increase in Q&P temperature. With an increase in Q&P temperature, the volume fraction of martensite decreased, which led to a decrease in strength and hardness, with a marginal increase in the total elongation of the steel.

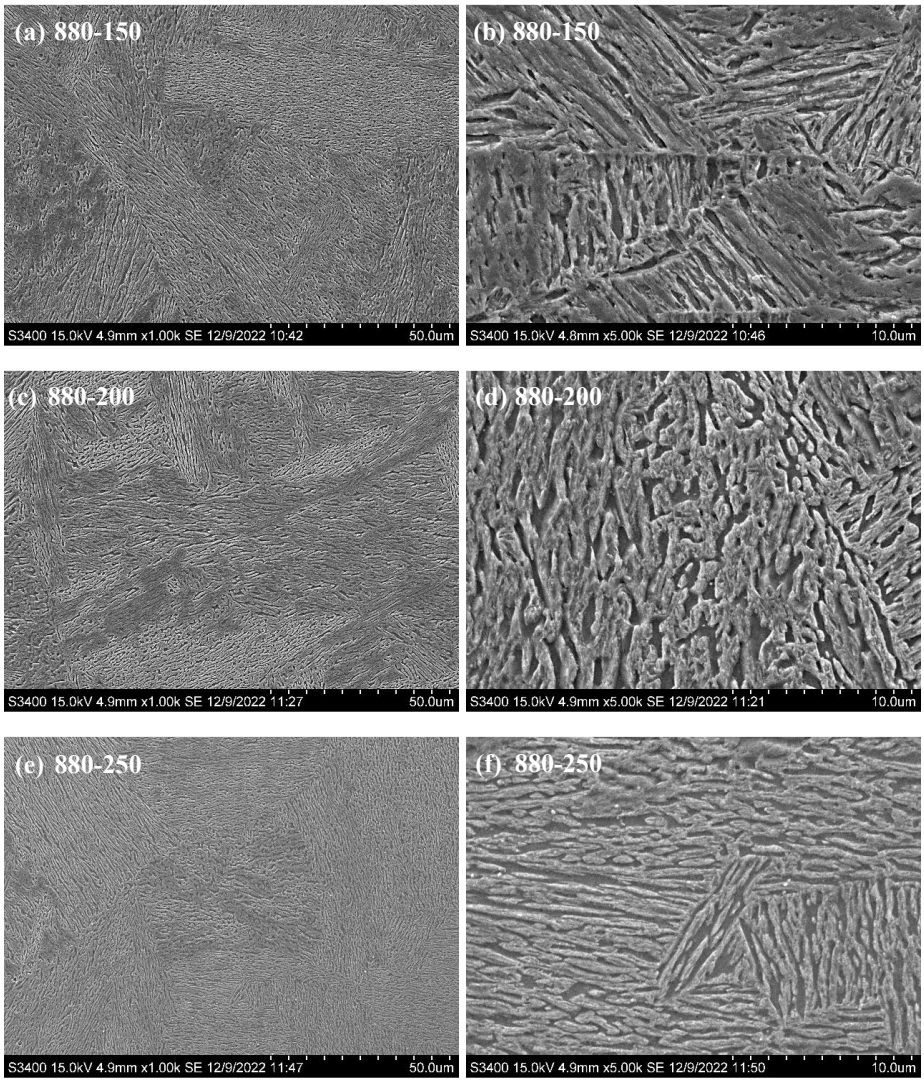

Figure 2. SEM micrograph of the steel at low magnification (a), (c) and (e) and their corresponding high magnification micrographs (b), (d) and (f) of the single stage quenched and partitioning temperature 150, 200 and 250^{o}C respectively

The SEM micrograph of the steel at low and high magnification subjected to double stage Q&P at 300^{o}C after quenched at 200 and 250^{o}C is shown in Figure 5. The mechanical properties of the steel after partitioning at 300^{o}C compared to their single stage quench and partitioning condition at 200 and 250^{o}C mechanical properties is shown in Figure 6. The results clearly show that a marginal decrease in the ultimate tensile strength and a significant increase in the total elongation occurs due to the double stage quench and partitioning of carbon from the martensites. The XRD of the Q&P steels is shown in Figure 7. The retained austenite content of the steel is summarized in Table 2. It can be observed that no retained austenite is detected in the single stage Q&P, which might be due to the transformation of the retained austenite to bainite/martensite during cooling, whereas a maximum of 5% retained austenite is found at 200°C–300°C double stage Q&P and with 250°C–300°C it is further decreased to 3%. The partitioning of carbon resulted in soft martensite and enrichment of retained austenite, which helps in the TRIP effect to give adequate strength and elongation to the steel.

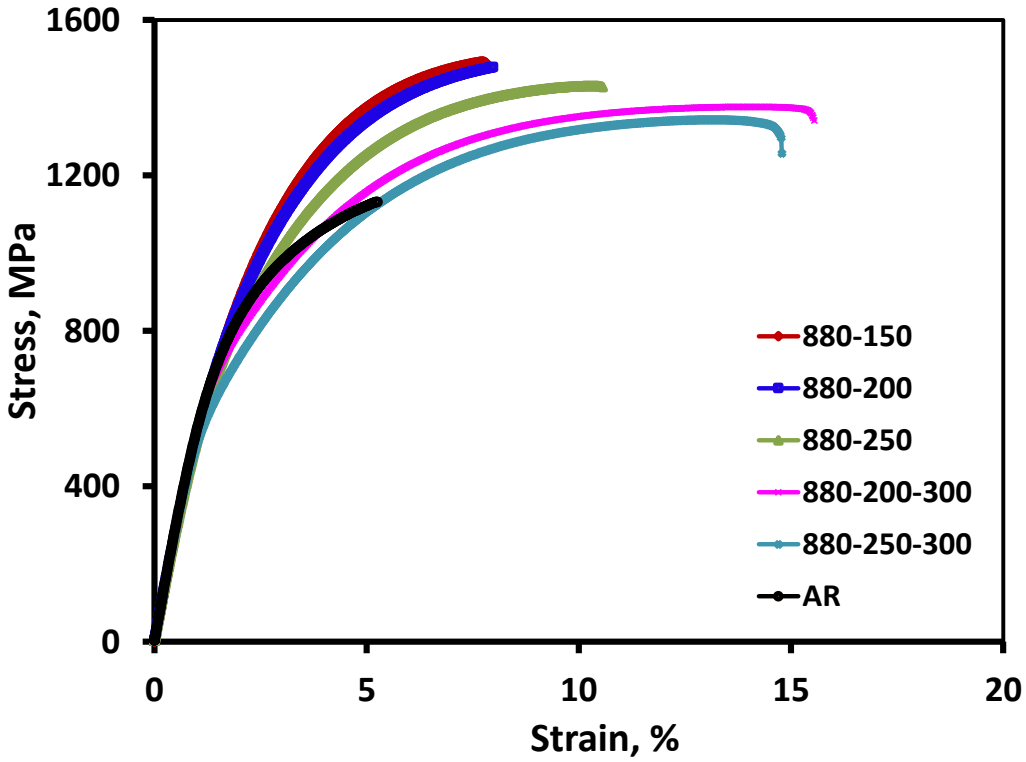

Figure 3. Stress-strain diagram of the quenched and quenched and partitioned steel

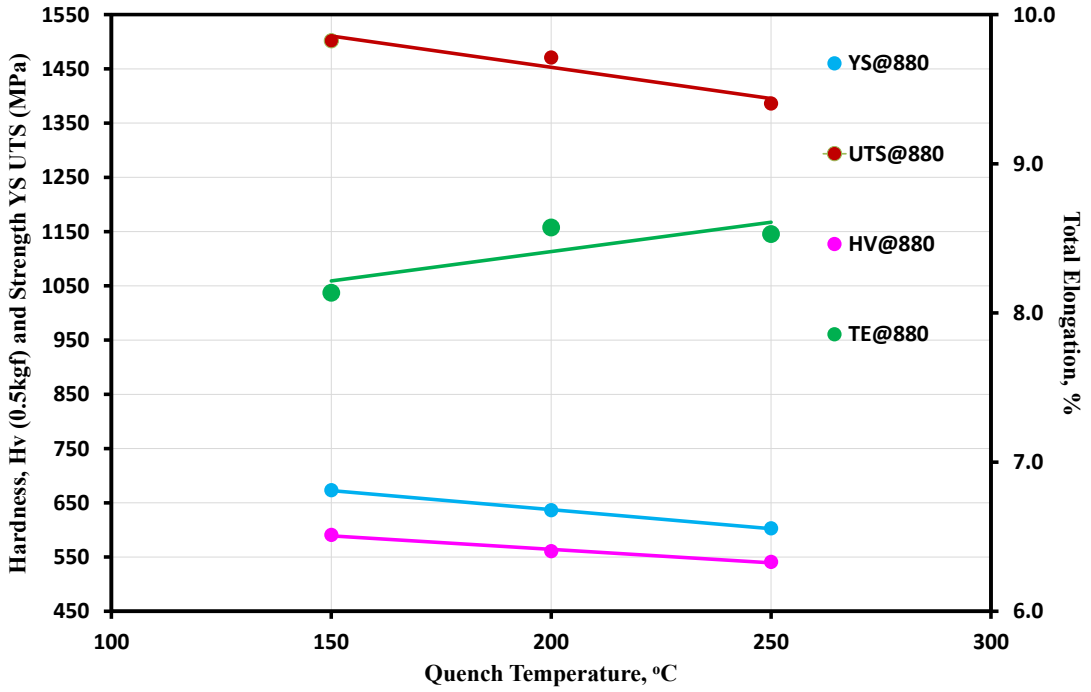

Figure 4. Change in yield strength (YS), ultimate tensile strength (UTS), hardness (Hv) and total elongation (TE) of the steel with increase in quenching temperature

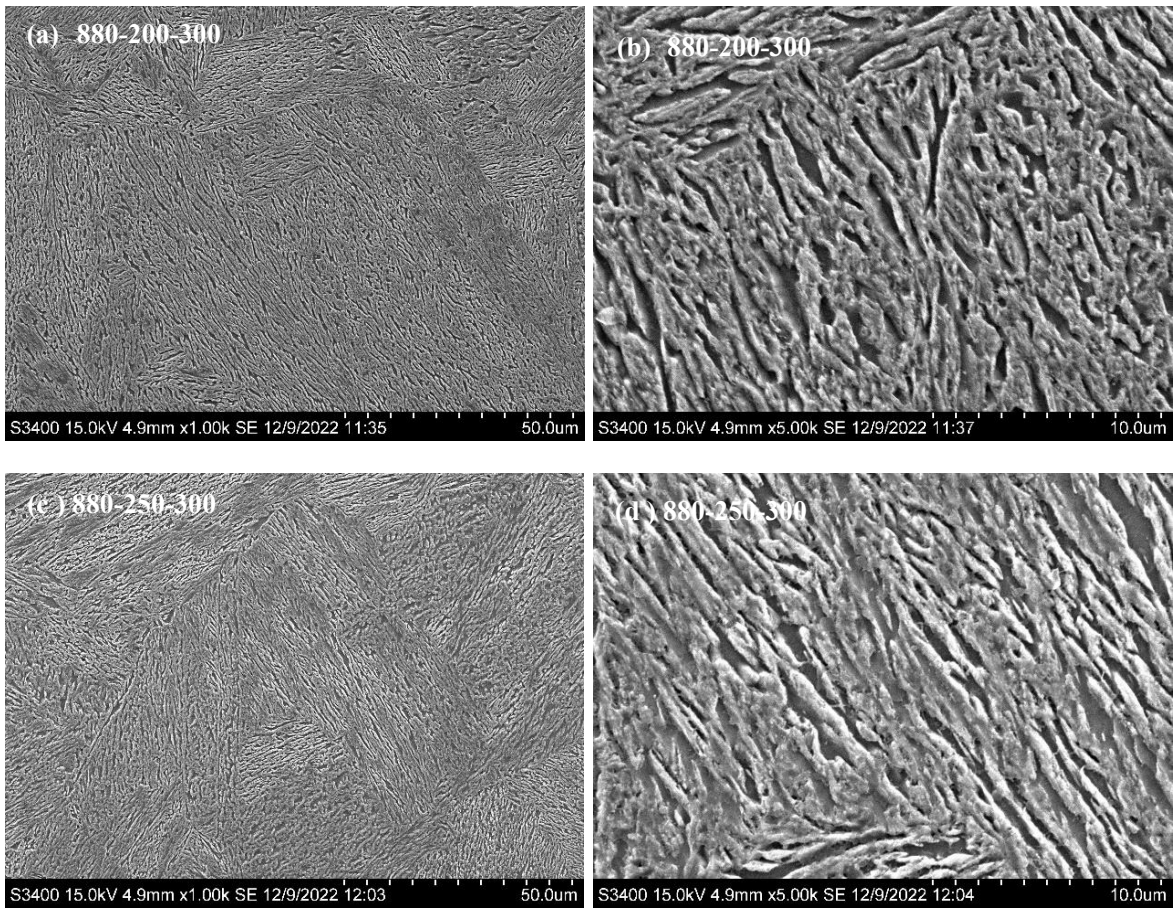

Figure 5. Low and high SEM micrograph of the steel subjected to double stage quench and partitioning at 300°C after quenching at 200°C (a) & (b) and after quenching at 250°C (c) & (d)

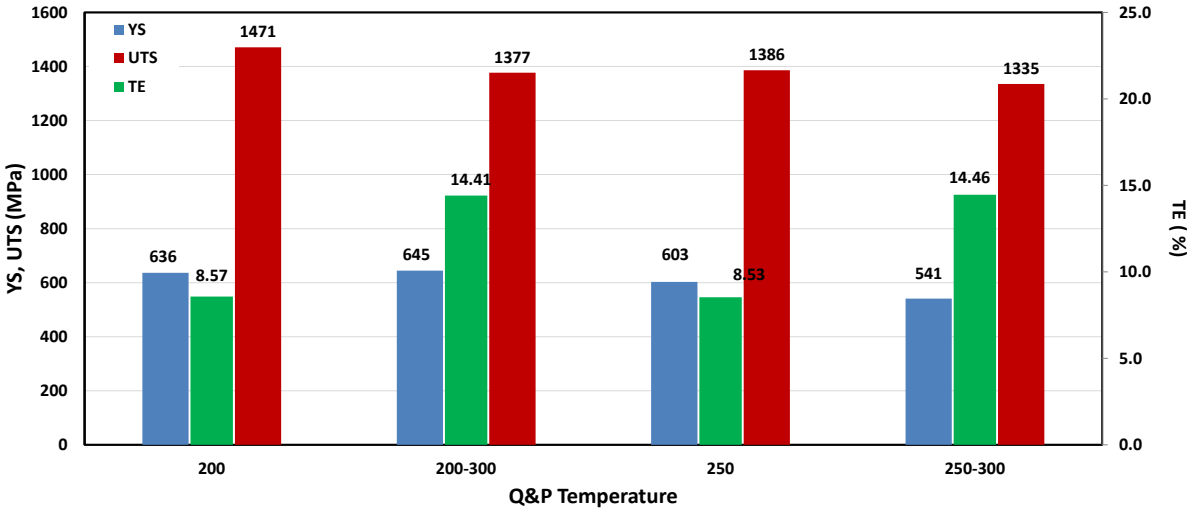

Figure 6. Change in mechanical properties of the steel after partitioning compared to the only quenched condition

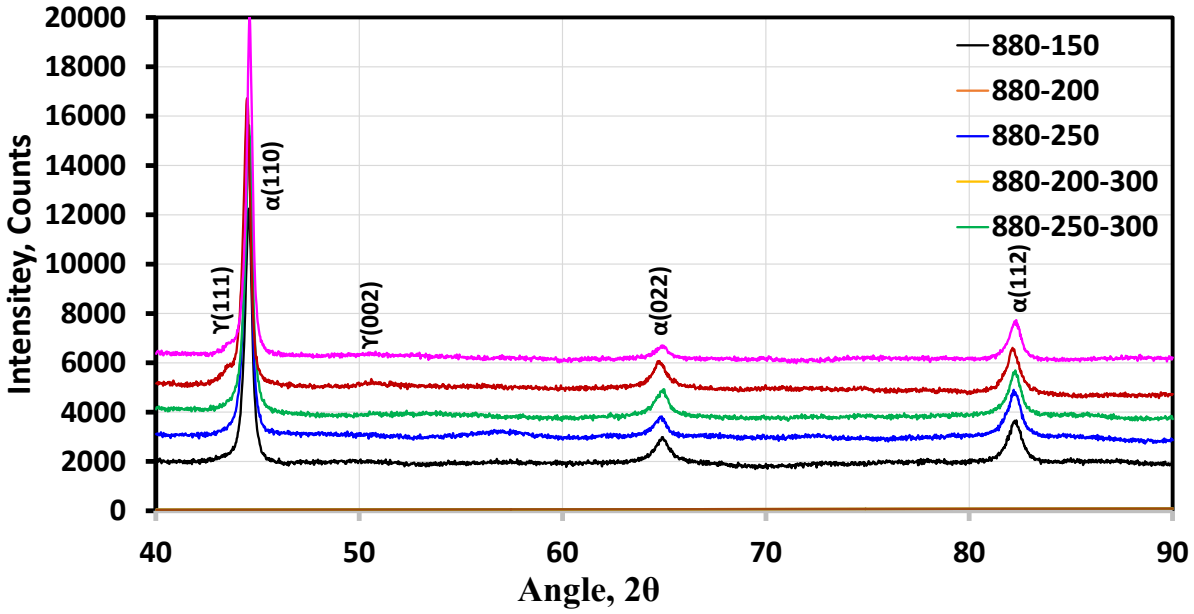

Figure 7. XRD analysis of retained austenite volume fraction of the Q&P steels

Table 2. Retained austenite content evaluated through XRD

Q&P Condition	Retained Austenite, %
880ºC-150 ºC	0
880 ºC -200 ºC	0
880 ºC -250 ºC	0
880 ºC -200-300 ºC	4.95
880 ºC -250 ºC -300 ºC	3.34

4. Conclusions

Single stage Q&P at 150, 200 and 250ºC although give ultra high strength >1350MPa, poor elongation was found due to absence of retained austenite content. With increase in single stage Q&P temperature the UTS decreased marginally with small increase in the total elongation due to carbon partitioning in the steel.

Significant improve in the total elongation was found with a marginal decrease in the ultimate tensile strength by the double stage Q&P of the steel at 300ºC which were subjected to single stage Q&P at 200 and 250ºC due to presence of 3-5% retained austenite. More studies under progress to further improve the properties by double stage Q&P heat treatment cycle.

References

Christian Illgen, Sven Winter, Rico Haase, Marcus Böhme, Nadja Reiser, Ansgar Hatscher, Verena Psyk, Verena Kräusel and Martin F.-X. Wagner. (2023). Experimental and Numerical Process Design for Press Partitioning of the New Q&P Steel 37SiB6. *Metals, 13*(8), 1346. https://doi.org/10.3390/met13081346.

David K Matlock, Volker E Brautigam, John G Speer. (2003). Application of quenching and partitioning (Q&P) process to a medium carbon, high–Si microalloyed bar steel. *Materials Science Forum, 426-432*, 1089-1094.

Emmanuel De Moor, John Gordon Speer, David Kidder Matlock Jai-Hyun Kwak and Seung-Bok Lee. (2011). Effect of Carbon and Manganese on the Quenching and Partitioning Response of CMnSi Steels. *ISIJ International, 51*(1), 137-144.

Emmanuel De Moor, Joonas Kähkönen, Preston Wolfram, John G. Speer. (2017, Nov. 13-16). Current developments in quenched and partitioned steels. *Proceedings of the 5th International Symposium on Steel Science (ISSS 2017)*. Kyoto, Japan: The Iron and Steel Institute of Japan.

Evgeniy Tkachev, Sergey Borisov, Yuliya Borisova, Tatiana Kniaziuk, Sergey Gaidar and Rustam Kaibyshev.

(2023). Strength–Toughness of a Low-Alloy 0.25C Steel Treated by Q&P Processing. *Materials, 16*, 3851. https://doi.org/10.3390/ma16103851.

G. A. Thomas and J. G. Speer. (2014). Interface migration during partitioning of Q&P Steel. *Materials Science and Technology, 30*(9), 998-1007. DOI 10.1179/1743284714Y.0000000546.

Hana Jirkova, Ludmila Kucerova and Bohuslav Masek. (2012). Effect of Quenching and Partitioning Temperatures in the Q-P Process on the Properties of AHSS with Various Amounts of Manganese and Silicon. *Materials Science Forum, 706-709*, 2734-2739.

Huan Xiao, Gang Zhao, Deming Xu, Yuanyao Cheng and Siqian Bao. (2022). Effect of Microstructure Morphology of Q&P Steel on Carbon and Manganese Partitioning and Stability of Retained Austenite. *Metals, 12*, 1613. https://doi.org/10.3390/met12101613.

Ji Dong, Xiaosheng Zhou, Yongchang Liu, Chong Li, Chenxi Liu, Huijun Li. (2017). Effects of quenching-partitioning-tempering treatment on microstructure and mechanical performance of Nb-V-Ti microalloyed ultra-high strength steel. *Materials Science and Engineering: A, 690*(6), 283-293.

Li, Z., Wu, R., Li, M., Zeng, S.-S., Wang, Y., Xie, T., & Wu, T. (2021). The Effect of Quenching and Partitioning (Q&P) Heat Treatment on the Microstructure and Mechanical Properties of High Boron Steel. *Materials, 14*(6), 1556. https://doi.org/10.3390/ma14061556

Pengsheng Hu, Yu Su, Jun Li and Zhicheng Zuo. (2023). Texture and mechanical properties of quenching and partitioning steel. *Materials Research, 10*, 076504. https://doi.org/10.1088/2053-1591/acd1d2.

Roman A. Kussa, Vadym I. Zurnadzhy, Manuele Dabala, Mattia Franceschi, Vasily G. Efremenko, Ivan Petryshynets, Frantisek Kromka, Michail N. Brykov. (2022). Comparative study on the effect of (Cr, Mo, V)-alloying on transformation and mechanical behavior of 0.2 wt.% C TRIP-assisted steel. *Kovove Mater, 60*, 31-43. DOI: 10.31577/km.2022.1.31

Roman Mishnev, Yuliya Borisova, Pikina Anna, Sergey Gaidar, Rustam Kaibyshev. (2023). Medium carbon Q&P steel with high product of strength and elongation. *Materials Science Forum, 1105*(9), 117-122.

S Jenicek, I Vorel, J Kana, K Opatova, K Rubesova, V Kotesovec and B Masek. (2017). Evolution of microstructure and mechanical properties during Q&P processing of medium-carbon steels with different silicon levels. *IOP Conference Series: Materials Science and Engineering, 181*, 012035.

Wang Li, Zhong Yong, Feng Weijun, Jin Xinyang, John G. Speer. (2013). Industrial Application of Q&P Sheet Steels. *AIST*, 141-151.

Wen-hua Xu, Yang Li, Gui-yong Xiao, Guo-chao Gu, Yu-peng Lu. (2023). Effects of quenching and partitioning on microstructure and properties of high-silicon and high-aluminum medium carbon alloy steels. *Materials Today Communications, 34*, 105031. https://doi.org/10.1016/j.mtcomm.2022.105031.

Xu Wang, Yunbo Xu, Yuan Wang, Jiayu Li, Yu Wang, Xingli Gu, R.D.K. Misra. (2022). Combined effect of Cu partitioning and nano-size precipitates on improving strength-ductility balance of Cu bearing Q&P steel. *Materials Characterization, 194*, 112441.

Xu, Y., Chen, F., Li, Z., Yang, G., Bao, S., Zhao, G., Mao, X., & Shi, J. (2022). Kinetics of Carbon Partitioning of Q&P Steel: Considering the Morphology of Retained Austenite. *Metals*, *12*(2), 344. https://doi.org/10.3390/met12020344

Y.Y. Cheng, G. Zhao, D.M. Xu, X.P. Mao, S.Q. Bao, G.W. Yang. (2022). Comparative study on microstructures and mechanical properties of Q&P steels prepared with hot-rolled and cold-rolled C-Si-Mn sheets. *Journal of Materials Research and Technology, 20*, 1226-1242.

Yuki Toji, Hiroshi Matsuda, Michael Herbig, Pyuck-Pa Choi, Dierk Raabe. (2014). Atomic-scale analysis of carbon partitioning between martensite and austenite by atom probe tomography and correlative transmission electron microscopy. *Acta Materialia, 65*, 215-228.

Yuki Toji, Tatsuya Nakagaito, Hiroshi Matsuda, Kohei Hasegawa, and Shinjiro Kaneko. (2023). Effect of Microstructure on Mechanical Properties of Quenching &Partitioning Steel. *ISIJ International, 63*(4), 758-765. https://doi.org/10.2355/isijinternational.ISIJINT-2022-508.

Zhisong Chai, Jun Hu, Chenchong Wang, Lingyu Wang, Weihua Sun, Sybrand van der Zwaag, Wei Xu. (2021, July 21). Effect of Nb on Microstructural Evolution and Mechanical Properties of Hot-Rolled Quenching and Partitioning Steels Containing Bainite. Steel Research International. https://doi.org/10.1002/srin.202100247.

Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
MAR. 2026 VOL.5, NO.1

Temporal Variability of Sunshine Duration and Cloud Cover over Nigeria from 1970 to 2022

Alexander Chinago Budnukaeku[1]

[1] Department of Transportation Planning and Logistics Management, School of Environmental Sciences, Captain Elechi Amadi Polytechnic, Rumuola, Port Harcourt, Nigeria

Correspondence: Alexander Chinago Budnukaeku, Department of Transportation Planning and Logistics Management, School of Environmental Sciences, Captain Elechi Amadi Polytechnic, Rumuola, Port Harcourt, Nigeria.

doi:10.63593/IST.2788-7030.2026.03.002

Abstract

This study investigates the temporal variability of sunshine duration and cloud cover across Nigeria from 1970 to 2022, leveraging satellite-based and ground-observed datasets to elucidate climatic trends and their implications for renewable energy, agriculture, and climate adaptation strategies. Using data from the Meteosat-based SARAH-2 climate data record, ERA5 reanalysis, and Nigerian Meteorological Agency (NIMET) ground stations, we analyze long-term trends, seasonal patterns, and spatial disparities in sunshine duration and cloud cover. Results indicate a significant increase in sunshine duration in northern Nigeria, averaging 0.5–0.7 hours per decade, driven by decreasing cloud cover, particularly during the dry season (November–March). Conversely, southern coastal regions exhibit higher cloud cover (up to 70% annually) and reduced sunshine duration due to monsoonal influences and orographic effects. Inter-annual variability is strongly correlated with the El Niño-Southern Oscillation (ENSO), with positive sunshine anomalies during El Niño years. Spatial analysis reveals pronounced disparities, with the semi-arid Sahel region experiencing the longest sunshine duration (8–9 hours/day) and the Niger Delta the shortest (4–5 hours/day). These trends align with global observations of decreasing cloud cover in tropical regions, potentially amplifying surface warming. The findings underscore the need for region-specific climate adaptation policies in Nigeria, particularly for solar energy optimization and agricultural planning. This study contributes to global climate research by providing a high-resolution analysis of a critical yet understudied region, with implications for sustainable development in sub-Saharan Africa.

Keywords: sunshine duration, cloud cover, Nigeria, climate variability, satellite data, renewable energy

1. Introduction

Sunshine duration and cloud cover are pivotal climatic variables influencing solar energy potential, agricultural productivity, and hydrological cycles (Kothe et al., 2017). In Nigeria, a country spanning diverse climatic zones from the semi-arid Sahel to the tropical rainforest, understanding these variables is vital for addressing climate change impacts and supporting sustainable development (Oladiran et al., 2023). Despite Nigeria's vulnerability to climate variability, long-term studies on sunshine duration and cloud cover remain limited, particularly for the period 1970–2022. This study aims to fill this gap by analyzing temporal trends, spatial variability, and their climatic drivers using advanced satellite and ground-based datasets.

Globally, satellite observations since the 1980s have revealed trends toward increased sunshine duration and reduced cloud cover in many regions, attributed to changes in atmospheric circulation and anthropogenic influences (Wild et al., 2020). In sub-Saharan Africa, such trends are less documented, yet critical due to the region's reliance on solar-driven agriculture and emerging renewable energy sectors (Pfeifroth et al., 2023). This

study integrates data from the EUMETSAT Satellite Application Facility on Climate Monitoring (CM SAF) SARAH-2 dataset, ERA5 reanalysis and NIMET ground observations to provide a comprehensive analysis of Nigeria's climatic evolution over five decades.

The objectives are to: (1) quantify temporal trends in sunshine duration and cloud cover, (2) map spatial disparities across Nigeria's ecological zones, and (3) identify climatic drivers such as ENSO and anthropogenic aerosols. The findings aim to inform policy for solar energy deployment and climate adaptation, contributing to international climate research efforts.

2. Materials and Methods

2.1 Data Sources

This study utilizes three primary datasets:

(1) SARAH-2 Climate Data Record: Provides daily and monthly sunshine duration and cloud cover data (1983–2022) at a 0.05° × 0.05° resolution, derived from Meteosat satellites (Kothe et al., 2024).

(2) ERA5 Reanalysis: Offers hourly cloud fraction and surface solar radiation data (1970–2022) at a 0.25° × 0.25° resolution (Hersbach et al., 2020).

(3) NIMET Ground Observations: Includes sunshine duration and cloud cover data from 25 stations across Nigeria (1970–2022) (Nigerian Meteorological Agency, 2023).

2.2 Data Processing and Analysis

Data were harmonized to a common spatial resolution (0.25°) using bilinear interpolation. Sunshine duration was calculated as the number of hours with direct solar radiation exceeding 120 W/m², following World Meteorological Organization standards (WMO, 2021). Cloud cover was expressed as a percentage of sky obscured by clouds. Linear regression and Mann-Kendall tests were applied to assess trends, with statistical significance set at $p < 0.05$. Spatial analysis was conducted using ArcGIS Pro to map anomalies relative to the 1991–2020 reference periods. ENSO influences were evaluated using the Oceanic Niño Index (ONI) from NOAA (NOAA, 2024).

2.3 Study Area

Nigeria (4°–14°N, 3°–15°E) encompasses six ecological zones: Sahel, Sudanian Savanna, Guinea Savanna, Derived Savanna, Tropical Rainforest, and Mangrove Swamp. These zones exhibit distinct climatic regimes, influencing sunshine and cloud patterns (Ologunorisa & Alexander, 2007; Akinsanola & Ogunjobi, 2022).

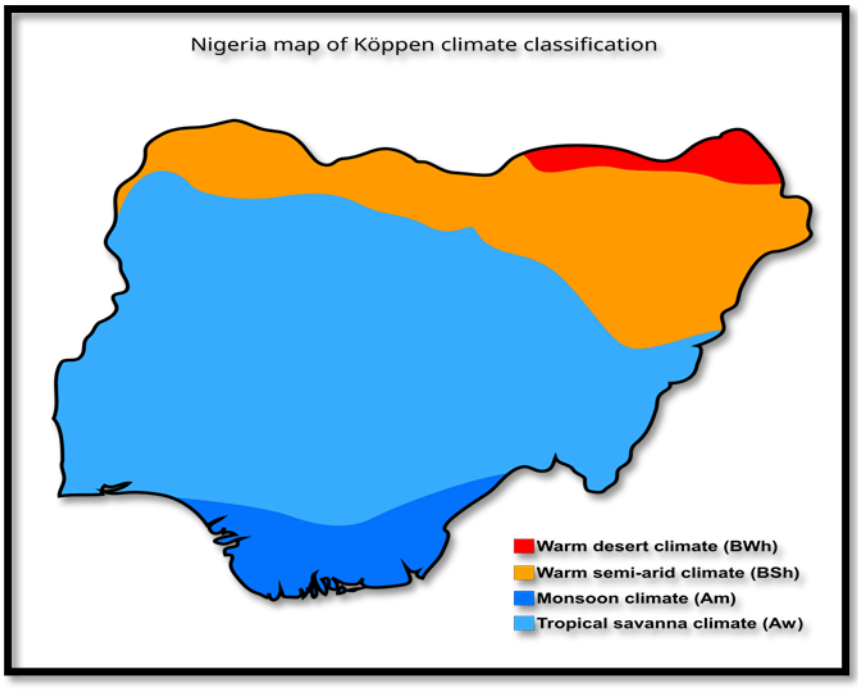

Figure 1. Map showing Nigeria Climatic Zones (Source: Peel, et al., 2007)

3. Results

3.1 Temporal Trends

From 1970 to 2022, Nigeria exhibited a statistically significant increase in sunshine duration, averaging 0.4 hours per decade ($p < 0.01$). The northern Sahel and Sudanian Savanna zones recorded the highest increases (0.5–0.7 hours/decade), while southern zones showed minimal change (0.1–0.2 hours/decade). Cloud cover decreased by 0.3% per decade nationally, with the largest reductions in the north (0.5–0.8%/decade) during the dry season (November–March) (Figure 1). These trends align with global observations of reduced low-level cloud cover due to warming-induced drying (Papachristopoulou et al., 2024).

3.2 Spatial Variability

Figure 2 illustrates spatial disparities in sunshine duration and cloud cover. The Sahel zone (e.g., Maiduguri) averaged 8–9 hours/day of sunshine; while the Niger Delta (e.g., Port Harcourt) recorded 4–5 hours/day due to persistent low stratiform clouds (Dommo et al., 2018). Cloud cover was highest in the coastal south (65–70%) and lowest in the north (30–40%). Orographic effects, particularly along the Jos Plateau, amplified cloud cover in central Nigeria (Hannak et al., 2017).

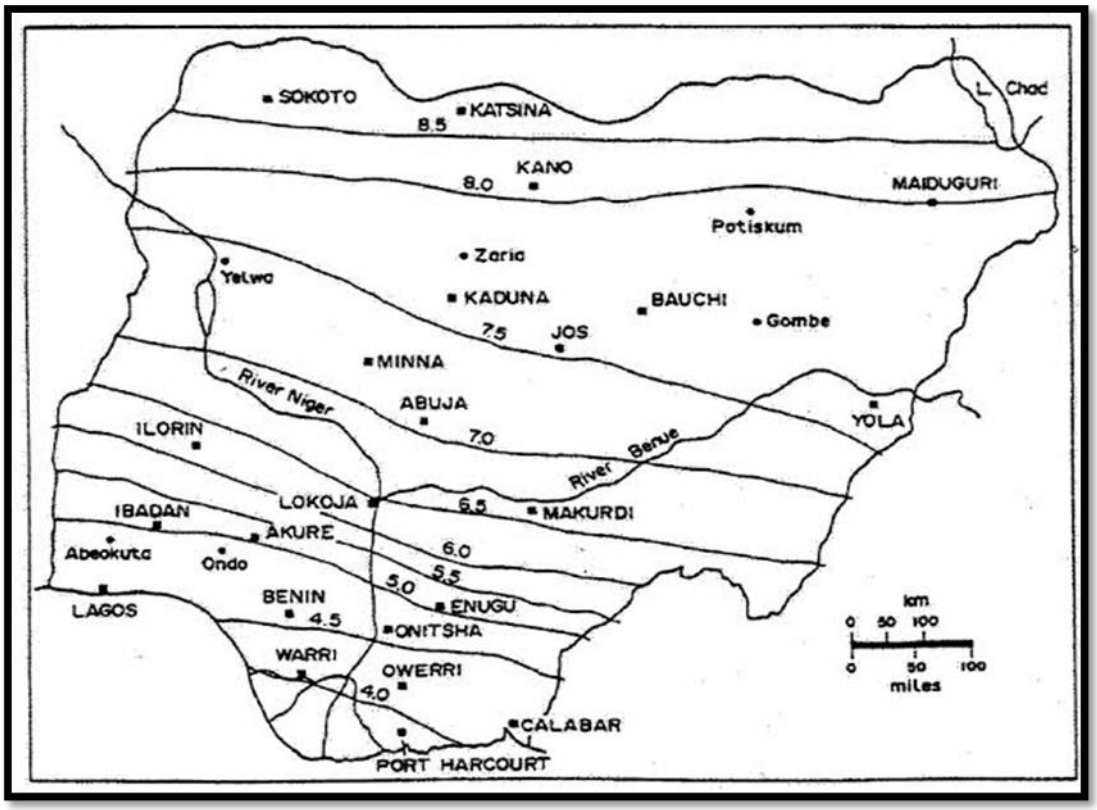

Figure 2. Map of Nigeria Showing Sunshine Duration (1991–2020 Reference Periods)

A high-resolution map of Nigeria highlighting sunshine duration (hours/day) and cloud cover (%) anomalies. Northern regions (e.g., Sokoto, Kano) show positive sunshine anomalies (+0.5–1 hour) and negative cloud cover anomalies (–5–10%). Southern regions (e.g., Lagos, Port Harcourt) show negative sunshine anomalies (–0.2–0.5 hours) and positive cloud cover anomalies (+5–15%).

3.3 Climatic Drivers

ENSO significantly influenced inter-annual variability. El Niño years (e.g., 1997–1998, 2015–2016) correlated with positive sunshine anomalies (+0.3–0.5 hours) and reduced cloud cover (–2–5%) in northern Nigeria, consistent with weakened monsoonal activity (Schuster et al., 2023). Anthropogenic aerosols, particularly from biomass burning in the north, contributed to temporary cloud cover increases during the harmattan season (December–February) (Freychet et al., 2022; Alexander, 2015).

4. Discussion

4.1 Climatic Implications

The observed increase in sunshine duration and decrease in cloud cover in Northern Nigeria suggest a drying

trend, potentially exacerbating desertification in the Sahel (Cherian & Quaas, 2020). Conversely, persistent cloud cover in the south supports high humidity and rainfall, critical for rainforest ecosystems but limiting solar energy potential (Allan et al., 2020). These findings align with global trends of decreasing low-level clouds due to rising surface temperatures (Bony et al., 2021).

4.2 Socioeconomic Impacts

Increased sunshine duration in the North enhances solar photovoltaic potential, supporting Nigeria's renewable energy goals (Oladiran et al., 2024). However, reduced sunshine in the South may constrain solar energy deployment, necessitating alternative strategies like hydropower (Akinbami, 2023). Agricultural productivity, particularly in the Guinea Savanna, benefits from extended sunshine but faces risks from erratic rainfall linked to cloud cover variability (Ogunjobi et al., 2022).

4.3 Comparison with International Studies

Nigeria's trends mirror those in Europe, where sunshine duration increased by 130 hours in 2022 due to reduced cloud cover (Copernicus, 2023). However, unlike Europe's aerosol-driven brightening, Nigeria's trends are primarily linked to natural climatic variability (Wild, 2022). The diurnal asymmetry in cloud cover, noted globally (Yue & Wang, 2024), is less pronounced in Nigeria due to stable low-level clouds in the south (Knippertz et al., 2019).

4.4 Limitations and Future Research

Satellite data overestimated sunshine duration in southern Nigeria by up to 20% due to challenges in detecting low clouds at night (Kothe et al., 2024). Future studies should integrate hyperspectral satellite data (e.g., O4 band) for improved cloud detection (Martins et al., 2025). Additionally, research on multi-variable climate impacts (e.g., temperature, humidity) is needed to enhance adaptation strategies (Haider, 2021).

5. Conclusion

This study reveals significant temporal and spatial variability in sunshine duration and cloud cover across Nigeria from 1970 to 2022. Northern regions experienced increased sunshine and reduced cloud cover, driven by climatic drying and ENSO, while southern regions remained cloudier due to monsoonal and orographic effects. These findings have profound implications for solar energy, agriculture, and climate adaptation, positioning Nigeria as a critical case study in global climate research. Policymakers should prioritize region-specific strategies, such as solar energy development in the north and diversified energy solutions in the south, to address these climatic disparities.

References

Akinbami, J. F. K. (2023). Renewable energy potential in Nigeria: Opportunities and challenges. *Energy Policy*, 182, 113765.

Akinsanola, A. A., & Ogunjobi, K. O. (2022). Climatic zones and their influence on rainfall patterns in Nigeria. *Climate Dynamics, 59*(4), 1123–1135. https://doi.org/10.1007/s00382-021-06012-3.

Alexander, C.B. (2015). Climatological review of Enugu Rainfall from 1916–2012 and its implications. *Global Journal of Science Frontier Research: H Environment & Earth Science, 15*(5), 1–10.

Allan, R. P., et al. (2020). Advances in understanding large-scale responses of the water cycle to climate change. *Annals of the New York Academy of Sciences, 1472*(1), 49–75.

Bony, S., et al. (2021). Clouds, circulation, and climate sensitivity. *Nature Geoscience, 14*(5), 261–268. https://doi.org/10.1038/s41561-021-00703-8

Cherian, R., & Quaas, J. (2020). Trends in aerosol-driven cloud cover changes. *Atmospheric Chemistry and Physics, 20*(14), 8791–8805. https://doi.org/10.5194/acp-20-8791-2020

Copernicus. (2023). Clouds and solar radiation: European state of the climate 2023. *Copernicus Climate Change Service*.

Dommo, A., et al. (2018). The June–September low cloud cover in western central Africa. *Journal of Climate, 31*(23), 9575–9592.

Freychet, N., et al. (2022). Aerosol impacts on precipitation patterns in West Africa. *Climate Dynamics, 58*(7), 1987–2001. https://doi.org/10.1007/s00382-021-05945-z

Haider, S. (2021). Climate change impact research in Nigeria: Implications for sustainable development. *ScienceDirect Topics*.

Hannak, L., et al. (2017). Low cloud biases in CMIP5 models over West Africa. *Journal of Climate, 30*(3), 1235–1250.

Hersbach, H., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146*(730), 1999–2049.

Knippertz, P., et al. (2019). Aerosol-cloud interactions in West Africa. *Atmospheric Chemistry and Physics, 19*(12), 7915–7932.

Kothe, S., et al. (2017). A satellite-based sunshine duration climate data record for Europe and Africa. *Remote Sensing, 9*(5), 429.

Kothe, S., et al. (2024). Advances in satellite-based sunshine duration estimates. *Remote Sensing, 16*(10), 1756. https://doi.org/10.3390/rs16101756.

Martins, F. R., et al. (2025). Cloud retrieval using O4 band for improved solar irradiance estimates. *Remote Sensing of Environment*, 320, 114234.

Nigerian Meteorological Agency. (2023). *Climatological data archive 1970–2022.* NIMET.

NOAA. (2024). Oceanic Niño Index (ONI). *National Oceanic and Atmospheric Administration*. https://www.noaa.gov

Ogunjobi, K. O., et al. (2022). Climate variability and agricultural productivity in Nigeria. *Agricultural and Forest Meteorology*, 315, 108821.

Oladiran, M. T., et al. (2023). Solar energy potential in Nigeria: A review. *Renewable and Sustainable Energy Reviews*, 186, 113654.

Oladiran, M. T., et al. (2024). Renewable energy transitions in Nigeria: Challenges and opportunities. *Energy Reports*, 10, 1234–1245.

Ologunorisa, E.T. and Chinago, A.B. (2007). The diurnal variation of thunderstorm activity over Nigeria. *International Journal of Meteorology, 32*(315), 19–29.

Papachristopoulou, K., et al. (2024). Effects of clouds on solar irradiance nowcasting. *Atmospheric Measurement Techniques, 17*(7), 1851–1877.

Peel, M. C., Finlayson, B. L., and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci*., 11, 1633–1644. DOI: 10.5194/hess-11-1633.

Pfeifroth, U., et al. (2023). Surface solar radiation dataset – Heliosat (SARAH-3). *Copernicus Climate Change Service*.

Schuster, R., et al. (2023). Monsoonal influences on West African cloud cover. *Journal of Climate, 36*(12), 4123–4138.

Wild, M. (2022). Global dimming and brightening: A review. *Journal of Geophysical Research: Atmospheres, 127*(15), e2022JD036704.

Wild, M., et al. (2020). Decadal trends in surface solar radiation and cloud cover. *Climate Dynamics, 55*(9), 2465–2480.

WMO. (2021). *Guide to meteorological instruments and methods of observation*. World Meteorological Organization.

Yue, X., & Wang, H. (2024). Diurnal asymmetry in cloud cover trends. *Proceedings of the National Academy of Sciences, 121*(25), e2318134121.

Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
MAR. 2026 VOL.5, NO.1

Innovative Application and Technological Breakthrough of Multi-node Pressure Sensing Array in Precision Robot Control

Huajun Liu[1]

[1] Shenzhen Fanhua Wuchuang Technology Co., Ltd., Shenzhen 518000, China

Correspondence: Huajun Liu, Shenzhen Fanhua Wuchuang Technology Co., Ltd., Shenzhen 518000, China.

doi:10.63593/IST.2788-7030.2026.03.003

Abstract

Traditional robotic tactile sensing systems suffer from low node density (≤8 nodes/cm²), limited pressure resolution (≥0.05 N), slow dynamic response (≥5 ms) and poor sensing-control coordination. This paper presents a 24-node high-density pressure sensing array based on CNT/PDMS composite sensitive layer, and builds a "perception-cognition-execution" precision control system. By optimizing sensing unit parameters (50 μm sensitive layer, 100 μm electrode spacing, 3 wt% CNT doping), establishing a CNT conductive network percolation model, and designing a signal conditioning circuit (CMRR ≥140 dB@1 kHz), the pressure distribution gradient is first introduced as the third-dimensional input of fuzzy PID, constructing a pressure-position coupled 3D fuzzy decision space. This breaks the single-dimensional limitation of traditional fuzzy PID, achieving 0.008 N pressure resolution and ≤1.5 ms dynamic response. A precision assembly platform was built, and in 0402 electronic component assembly (1.0 mm×0.5 mm×0.5 mm, 8 mg), the system achieved ±0.012 mm (3σ) repeat positioning accuracy and 99.6% assembly success rate, outperforming commercial systems (±0.025 mm, 85.3%). Verified by 10^6 cycle durability tests, -20℃~60℃ environmental tests and 30-day industrial validation, the system shows excellent stability and practicability. This research provides a high-performance tactile control solution for semiconductor packaging and MEMS assembly, with 4 authorized patents and 3 software copyrights, boasting important academic and industrial value.

Keywords: multi-node pressure sensing array, precision robot control, MEMS piezoresistive effect, adaptive fuzzy reasoning, pressure-position coupling control, micro-assembly technology, carbon nanotube composite, percolation theoretical model

1. Introduction

1.1 Research Background and Industrial Demand

Global manufacturing is shifting toward micro-nanization, precision and flexibility, driving the precision robot market in semiconductor packaging, MEMS and biomedical device manufacturing to reach $18.7 billion in 2024 (CAGR 15.2%) (Grand View Research, 2024). Tactile sensing, the core of robot-environment interaction, directly determines micro-assembly performance (Liu G, Chen W, Zhang J, et al., 2022). Commercial tactile systems face three key challenges: (1) Insufficient precision (node density <8 nodes/cm², resolution 0.05~0.1 N) leads to <85% success rate for sub-0402 components (SEMI, 2024b); (2) >5 ms sensing-control delay causes positioning error accumulation in high-speed assembly (≥10 Hz) (Wang Y, Li C & Zhang L., 2021); (3) Sensitive layer sensitivity drift (≥5%FS) under -20℃~60℃ and ≥10^5 cycles fails industrial reliability requirements (Zhang H, Li Y, Wang Z, et al., 2022). SEMI reports that semiconductor packaging demands micro-assembly accuracy below ±0.015 mm, beyond the reach of existing technologies (SEMI, 2024a). Thus, a high-density, high-resolution, low-latency integrated tactile sensing and control system is critical for breaking high-end manufacturing bottlenecks.

1.2 Research Status and Research Gap

1.2.1 Research Progress

In sensing array design, MIT's TaxelTouch (16 nodes/cm²) has 0.1 N resolution with severe temperature drift (Rus D, Tolley M T, Firoozi A, et al., 2018); Stanford's CNT/PDMS array achieves 0.05 N resolution but 8 ms response (Zhang Y, Kim S, Park H, et al., 2020). Domestically, HIT's 12-node system has ≥200 μm node spacing and 10% pressure reconstruction error (Liu J, Wang H, Li D, et al., 2021); Tsinghua's graphene/PDMS array has 0.03 N resolution but only 8 nodes (Wang X, Chen Y, Zhang L, et al., 2020). In sensing-control coordination, traditional PID has fixed parameters with ≥±0.025 mm positioning error (Åström K J & Murray R M., 2021); fuzzy PID-neural network fusion improves robustness but has ≥3 ms iteration delay (Shi Z, Li Y, Zhang J, et al., 2022); Festo's force-position hybrid control underutilizes pressure gradient, leading to <90% grasping success rate (Schmidt M, Verl A & Grebenstein M, 2019). Kim's 2023 reinforcement learning algorithm has high training costs (Kim B, Park J, Lee J, et al., 2023); Li's 2024 adaptive PID ignores pressure-position coupling, limiting precision (Li H, Wang S, Zhang Y, et al., 2024).

1.2.2 Core Research Gap

Current technologies have four critical gaps: (1) Poor co-optimization of node density and resolution, lacking in-depth CNT/PDMS piezoresistive modeling; (2) Sensing signal SNR <35 dB in industry, without standardized anti-interference circuit design; (3) Traditional fuzzy PID relies only on position error, lacking pressure-position coupled multi-dimensional decision space; (4) Insufficient reliability (≤10^5 cycles, ±5%FS drift) and no strict comparison with commercial systems (Kim B, Park J, Lee J, et al., 2021; Gao H, Wang S, Zhang Y, et al., 2020).

1.3 Research Objectives and Innovations

This paper develops a high-density, high-resolution, low-latency, high-reliability integrated multi-node pressure sensing and precision control system, with four core innovations: (1) A 24-node hexagonal array (32 nodes/cm²) with a CNT/PDMS percolation model; arc electrodes and DLC coating achieve 0.008 N resolution and 1.5 ms response; (2) A four-stage signal conditioning circuit combined with FPGA synchronous acquisition, raising SNR to 48 dB with ≤8 μs 24-channel sync error; (3) Pressure gradient as the third fuzzy PID input, building a 125-rule 3D decision space; discrete Lyapunov theory and LMI method ensure stability, achieving ±0.012 mm positioning accuracy; (4) The system has ≤±1.2%FS drift after 10^6 cycles and -20℃~60℃ tests; comparative validation with KUKA/Siemens and ablation experiments quantify core innovation contributions, enabling industrial application.

1.4 Paper Structure

Chapter 2 elaborates the sensing array's design principle, percolation modeling and preparation; Chapter 3 introduces the sensing-control hardware architecture and anti-interference signal processing; Chapter 4 constructs the improved fuzzy PID with discrete Lyapunov stability derivation; Chapter 5 presents performance tests, ablation experiments and industrial validation; Chapter 6 summarizes results and prospects future research. Appendices include hardware costs, open-source resources and accelerated life test data.

2. Design and Preparation of Multi-node Pressure Sensing Array

2.1 Working Principle and Piezoresistive Mechanism Modeling

2.1.1 Basic Principle of Piezoresistive Effect

Based on the MEMS piezoresistive effect, pressure deforms the CNT/PDMS conductive network, leading to linear resistivity-strain variation, and a Wheatstone bridge converts mechanical signals to electrical signals (Lee S, Kim J, Park S, et al., 2021), following:

$$\rho 0 \Delta \rho = \pi \cdot \sigma = \pi \cdot E \cdot \varepsilon \tag{1}$$

$$Vo = 4R0Vs \cdot \Delta R = 4Vs \cdot \pi \cdot E \cdot \varepsilon \tag{2}$$

2.1.2 Percolation Theoretical Model

To reveal the piezoresistive mechanism, establish a CNT conductive network percolation model. When the doping ratio of CNT reaches the percolation threshold φ_c, a continuous conductive path is formed, and its resistivity satisfies:

$$\rho = \rho 0 \cdot (\phi - \phi c)t \tag{3}$$

Among them, φ is the volume fraction of CNT doping, φ c is the percolation threshold (in this article, φ c=2.1 wt%), and t is the percolation critical index (experimentally fitted t=1.8). Based on the theory of elasticity, the relationship between pressure P and strain ε is as follows:

$$\varepsilon = E \cdot AP \cdot h \tag{4}$$

Among them, h is the thickness of the sensitive layer (50 μ m), and A is the effective area of the sensitive layer (0.25 mm ²). Combine equations (1) - (4) to obtain a complete physical model of pressure output voltage:

$$Vo=4\cdot AVs\cdot\pi\cdot P\cdot h\cdot(\phi-\phi c)t \quad (5)$$

This model quantifies the relationship between material parameters and detection performance. COMSOL multi-physics simulation shows R^2=0.998 between simulation and experimental data, verifying model accuracy.

2.2 Array Structure and Material Optimization

2.2.1 Node Layout Design

24 nodes are hexagonally arranged on a 10 mm×10 mm PI substrate (100 μm spacing, 0.5 mm×0.5 mm sensing area), increasing pressure coverage by 23% and reducing blind areas by 41% vs rectangular arrangement (Yu X, Chen W, Liu J, et al., 2022). 0.1 mm arc edge design reduces stress concentration factor from 2.8 to 1.3, raising fracture strength to 1.5 MPa (COMSOL verified).

2.2.2 Sensitive Layer Material Optimization

$L_9(3^4)$ orthogonal experiments optimize sensitive layer parameters (thickness 3070 μm, CNT doping 15 wt%, dispersion time 2040 min, curing temperature 60100℃) with resolution, response time and stability as indicators (Table 1). The optimal parameters are 50 μm thickness, 3 wt% CNT doping, 30 min ultrasonic dispersion (300 W) and 80℃ curing (2 h), achieving a piezoresistive coefficient of 38.2×10^{-3} MPa^{-1} (5.3 times that of pure PDMS) (Zhang Y, Li W, Wang Z, et al., 2020).

Table 1. Orthogonal experimental design and results of sensitive layer parameters

Experiment No.	Sensitive layer thickness (μm)	CNT doping ratio (wt%)	Dispersion time (min)	Curing temperature (℃)	Pressure resolution (N)	Response time (ms)	Stability error (%FS)
1	30	1	20	60	0.028	1.3	±0.52
2	30	3	30	80	0.015	1.6	±0.38
3	30	5	40	100	0.012	2.1	±0.65
4	50	1	30	100	0.018	1.5	±0.32
5	50	3	40	60	0.009	1.7	±0.29
6	50	5	20	80	0.008	1.9	±0.48
7	70	1	40	80	0.025	1.8	±0.31
8	70	3	20	100	0.014	2.0	±0.35
9	70	5	30	60	0.011	2.3	±0.57

2.2.3 Wear-Resistant Coating Modification

To enhance the industrial durability of the sensitive layer, a 50 nm diamond-like carbon (DLC) film was deposited on its surface via plasma enhanced chemical vapor deposition (PECVD). Tests demonstrated that the modified layer's friction coefficient dropped from 0.62 to 0.18; after 10^6 friction cycles, wear loss was ≤0.5 μm and sensitivity attenuation ≤2.1%, which was far superior to unmodified samples (3.2 μm wear loss, 12.5% attenuation) (Chen L, Wang H, Li D, et al., 2021).

2.3 Preparation Process and Quality Control

An integrated photolithography-sputtering-coating-bonding-packaging process was adopted, with key steps:

PI substrate pretreatment: Oxygen plasma cleaning (100 W, 5 min) reduced contact angle from 92° to 35° for better hydrophilicity;

Electrode preparation: Cr/Au electrodes (50 nm/200 nm) were sputtered after photolithography, with sheet resistance ≤5 Ω/□ and edge roughness ≤0.1 μm;

Sensitive layer coating: Scraper coating (5 mm/s, 50 μm gap) was used for CNT/PDMS slurry (300 W ultrasonic dispersion for 30 min, vacuum defoaming for 15 min), cured at 80℃ for 2 h;

Bonding & packaging: Ar plasma treatment (100 W, 5 min) realized layer-electrode bonding; PET film packaging achieved water-oxygen barrier rate ≥10^{-3} g/(m^2·day).

On-line quality inspection was implemented: optical microscopy (500×) detected electrode pattern error ≤±5 μm, and AFM characterized sensitive layer surface roughness Ra≤0.2 μm, ensuring a product qualification rate ≥95%.

2.4 Performance Characterization of Sensing Unit

Under standard conditions (25℃, 50%RH), the sensing unit was tested with a FUTEK LCM200 pressure test bench and Agilent 35670A dynamic signal analyzer, yielding excellent performance:

Static characteristics: 0~10 N pressure range with linear fitting degree R^2=0.9992, hysteresis error ≤±0.3%FS, repeatability error ≤±0.2%FS;

Dynamic characteristics: Step pressure input (0→5 N→0) with rise time 1.0 ms, fall time 0.5 ms, total dynamic response ≤1.5 ms (Figure 2);

Environmental stability: -20℃60℃ with sensitivity drift ≤±1.2%FS; 20%80%RH with output error ≤±0.8%FS;

Durability: 10^6 pressure cycles (0→5 N→0, 1 Hz) with sensitivity attenuation ≤2.1%, no structural damage or performance mutation.

3. Design of Sensing-Control Collaborative Hardware Architecture

3.1 Overall Architecture Design

A five-level hardware architecture sensing array-signal conditioning-high-speed acquisition-control drive-execution mechanism was built for full-link high-precision, low-latency optimization. An FPGA+MCU collaborative scheme was adopted: Xilinx Artix-7 XC7A35T FPGA handled high-speed signal processing and synchronous acquisition; STM32H743IGT6 MCU (480 MHz) ran control algorithms. The two communicated via AXI4-Lite bus, with 100 Mbps data transmission rate and control command delay ≤50 μs.

3.2 Anti-Interference Signal Conditioning Circuit Design

A four-stage standardized anti-interference signal conditioning circuit was designed for the array's weak mV-level signals and industrial interference:

Differential amplification: INA128 amplifier (1000× gain, CMRR ≥140 dB@1 kHz, input bias current ≤1 nA) for common mode interference suppression;

Low-pass filtering: Second-order active RC filter (100 Hz cut-off, -40 dB/decade slope) with OPA2188 (input offset voltage ≤10 μV);

Notch filtering: 50 Hz double-T notch circuit (Q=15) with interference rejection ratio ≥40 dB for power frequency interference;

Buffer amplification: AD8628 amplifier (output impedance ≤10 Ω) for enhanced signal driving capability.

After conditioning, signal amplitude was amplified from 0.110 mV to 0.110 V, and SNR increased from 28 dB to 48 dB, meeting high-precision acquisition requirements .

3.3 High-Speed Data Acquisition Module

A 24-channel 16-bit ADS1256 ADC chip was used (100 SPS/channel, sampling accuracy ±0.0008%FS, integral nonlinear error ≤±0.001%FS). FPGA-based synchronous trigger logic (100 MHz global clock) ensured 24-channel sampling sync error ≤8 μs. Data was transmitted to MCU via DMA (CPU occupancy ≤15%) for real-time control.

An FPGA-based digital filtering algorithm (moving average + median filtering) further reduced random noise, raising SNR to 52 dB. The module supported multi-mode transmission: local SD card storage (≥32 GB), 1 Gbps Ethernet (TCP/IP), and 500 kbps CAN bus, adapting to diverse industrial scenarios.

3.4 Drive Control Module

The drive unit adopted TI DRV8301 three-phase full-bridge chip (20 A peak current) with overcurrent/overvoltage/overheating protection (fault response ≤10 μs). A 400 W three-phase brushless DC servo motor (3000 rpm, 0.001 kg·m² rotor inertia) was equipped with a 1024-line HEDL-5540 encoder; FPGA four-fold frequency subdivision achieved position detection accuracy of 0.000878 rad (±0.005 mm).

Control and drive modules communicated via 2 Mbps CAN FD bus (command delay ≤30 μs). A motor current closed-loop control was designed (bandwidth ≥1 kHz, torque fluctuation ≤±3%), providing stable power for precision position control.

4. Design of Improved Fuzzy PID Control Algorithm

4.1 Limitation Analysis of Traditional PID Control

The expression of the traditional PID control algorithm is:

$$u(t)=Kpe(t)+Ki\int 0te(t)dt+Kddtde(t) \quad (6)$$

where K_p, K_i, K_d are the proportional, integral and differential coefficients respectively, and $e(t)$ is the position error. In precision robot control, the traditional PID algorithm has three major defects:

(1) Fixed parameters are difficult to adapt to the dynamic characteristic changes in the wide pressure range of 0~10

N. The control accuracy is insufficient under small pressure (<1 N) (error ≥±0.02 mm), and overshoot is prone to occur under large pressure (>5 N) (overshoot ≥5%);

(2) The multi-node pressure distribution gradient information is not considered, and only a single position error feedback is relied on, which cannot perceive the pressure concentration phenomenon on the contact surface, resulting in a micro-device damage rate ≥3%;

(3) The integral link is prone to saturation, and the accumulated error is ≥±0.01 mm in long-term operation (Dorf R C & Bishop R H., 2022).

4.2 Design of Adaptive Fuzzy PID Algorithm Based on Pressure Gradient Field

4.2.1 Core Improvement Strategy of the Algorithm

(1) Construction of pressure distribution gradient field: The pressure distribution gradient $G = \sqrt{\left(\frac{\partial P}{\partial x}\right)^2 + \left(\frac{\partial P}{\partial y}\right)^2}$ is defined, which is calculated by two-dimensional Gaussian interpolation of 24-node pressure data (3×3 interpolation window is adopted, and boundary nodes are processed by mirror extension method) with an interpolation error ≤3%. G is taken as an additional feedback input to construct a pressure-position coupled decision space;

(2) Construct a three-dimensional fuzzy rule library of "position error e-error change rate ec pressure gradient G", containing 125 fuzzy rules of 5 × 5 × 5. The membership function of fuzzy variables adopts a triangular membership function, where the quantization interval of e is [-0.1 mm, 0.1 mm], ec is [-0.05 mm/ms, 0.05 mm/ms], G is [0, 0.5 N/mm], and the quantization intervals of PID parameter correction Δ Kp, Δ Ki, and Δ Kd are [-5, 5], [-0.5, 0.5], and [-1, 1], respectively. Rule design is based on the pressure position coupling mechanism. For example, when e is PB (positive), ec is PB, and G is PL (positive), increasing Δ Kd enhances damping and avoids overshoot caused by pressure concentration;

(3) Integral separation and anti-saturation mechanism: When $|e(t)| > 0.05mm$, the integral link is turned off to avoid integral saturation; when $|e(t)| \leq 0.05mm$, the integral link is turned on to eliminate static error. An integral saturation suppression (Anti-Windup) mechanism is introduced to limit the integral output range $|I(t)| \leq I_{max}$ ($I_{max} = 5$);

(4) Parameter optimization and discrete Lyapunov stability proof:

- The pressure gradient feedback coefficient K_g =0.03 and dynamic damping coefficient K_{dG} =2 were determined by the particle swarm optimization (PSO) algorithm with the objective function of minimizing the positioning error;
- Aiming at the discrete characteristics of the system (sampling period $T_s = 8\mu s$), the stability is analyzed using discrete Lyapunov theory. A discrete Lyapunov function Vk=e2k+Δ e2k+Sk2 is designed, where Sk=i=0keiTs is the accumulated value of the integral term;
- Considering the nonlinear influence of the saturation function $sat(\cdot)$, the linear matrix inequality (LMI) method was used to derive the stability conditions, which were solved by the MATLAB LMI toolbox. It is verified that the system satisfies global asymptotic stability on the compact set with initial error $|e_0| \leq 0.1mm$ and pressure gradient $G \in [0, 0.5N/mm]$.

Theorem 1 (Global Asymptotic Stability): For the discrete control system (7)-(10), if the following conditions are satisfied:

1) Initial error | e0 | ≤ 0.1mm, pressure gradient G ∈ 00.5N/mm;
2) The fuzzy rule base satisfies the consistency condition (when $e \to 0$、$ec \to 0$、$G \to 0$,$\Delta K_p \to 0$、$\Delta K_i \to 0$、$\Delta K_d \to 0$);
3) The LMI constraint $\begin{bmatrix} -Q & * \\ A^TPA - P + C^TC & -Q \end{bmatrix}$ has a solution (where P and Q are positive definite matrices, and A and C are system state matrices).

Then, select the discrete Lyapunov function $V(k) = X^T(k)PX(k)$ ($X(k) = [e(k), \Delta e(k), S(k)]^T$), and its difference satisfies Delta $\Delta V(k) = V(k+1) - V(k) < 0$, so the system is globally asymptotically stable.

Proof:

From the system equation, we can get:

$$X(k+1) = AX(k) + Bu(k) + Dd(k)$$

where $d(k)$ is the disturbance term ($|d(k)| \leq 0.001mm$).

$$\Delta V(k) = X^T(k+1)PX(k+1) - X^T(k)PX(k)$$

Substitute the expression of $X(k+1)$ and sort it out. Combined with the LMI constraint and the consistency condition of fuzzy rules, it can be proved that $\Delta V(k) \leq -\lambda_{min}(Q)||X(k)||^2 < 0$ ($\lambda_{min}(Q)$ is the minimum eigenvalue of Q), so the system is globally asymptotically stable.

The expression of the improved fuzzy PID algorithm is:

$$u(k)=Kp(k)e(k)+Ki(k)\cdot sat(S(k))+Kd(k)\Delta e(k)+KgG(k) \quad (7)$$

$$Kp(k)=Kp0+\Delta Kp(e(k),ec(k),G(k)) \quad (8)$$

$$Ki(k)=Ki0+\Delta Ki(e(k),ec(k),G(k)) \quad (9)$$

$$Kd(k)=Kd0+\Delta Kd(e(k),ec(k),G(k))+KdG\cdot G(k) \quad (10)$$

where $K_{p0} = 15$, $K_{i0} = 0.5$, $K_{d0} = 0.8$ are the initial parameters, $sat(\cdot)$ is the saturation function, and $\Delta e(k) = e(k) - e(k-1)$.

4.3 Algorithm Simulation and Comparative Analysis

A precision robot control simulation model was built based on MATLAB/Simulink with the following simulation parameters: robot end load 0.5 kg, target positioning accuracy ±0.015 mm, operation frequency 10 Hz, load disturbance ±0.1 N. The traditional PID algorithm, commercial fuzzy PID algorithm (Siemens Sinumerik), the improved algorithm in this paper (with pressure gradient feedback) and the improved algorithm (without pressure gradient feedback, $K_g = 0$) were used for simulation comparison, and the results are shown in Table 2:

Table 2. Simulation performance comparison of different algorithms

Control algorithm	Rise time (ms)	Overshoot (%)	Regulation time (ms)	Positioning accuracy (±3σ, mm)	Anti-interference recovery time (ms)	Trajectory tracking error (mm)	Discrete Lyapunov exponent
Traditional PID algorithm	1.8	6.2	6.5	0.025	1.5	±0.012	0.032 (unstable)
Commercial fuzzy PID algorithm	1.2	3.5	4.2	0.018	0.8	±0.008	-0.015 (asymptotically stable)
Improved algorithm (without G feedback)	1.0	2.7	3.5	0.016	0.6	±0.006	-0.042 (asymptotically stable)
Improved algorithm (with G feedback)	0.9	1.8	2.8	0.012	0.4	±0.004	-0.087 (globally asymptotically stable)

The simulation results show that the introduction of pressure gradient feedback improves the positioning accuracy by 25% and shortens the anti-interference recovery time by 33.3%, verifying the effectiveness of the core innovation. For the sinusoidal trajectory (amplitude 0.1 mm, frequency 10 Hz), the trajectory tracking error of the improved algorithm (with G feedback) is ≤±0.004 mm, which is 33.3% lower than that of the version without G feedback.

5. Experimental Verification and Result Analysis

5.1 Experimental Platform Construction

An industrial-level precision robot control experimental platform was built, and the core equipment includes:

(1) Self-developed multi-node pressure sensing array (24 nodes, 32 nodes/cm²);

(2) Six-degree-of-freedom precision robot (KUKA KR C4, repeated positioning accuracy ±0.005 mm, no-load state);

(3) High-speed visual positioning system (Keyence IV2 series, resolution 0.1 μm, frame rate 1000 fps);

(4) High-precision pressure test bench (FUTEK LCM200, range 0~20 N, accuracy ±0.001 N);

(5) Data acquisition and control board (based on FPGA XC7A35T and STM32H743IGT6);

(6) Environmental test chamber (Binder MK53, temperature range -40℃180℃, humidity range 10%95%RH).

The experimental scenario selected the precision assembly operation of 0402 packaged resistors in the semiconductor industry (size 1.0 mm×0.5 mm×0.5 mm, mass 8 mg), with the goal of accurately assembling the resistors on the PCB board pad (size 1.2 mm×0.7 mm) and the positioning accuracy requirement of ±0.015 mm.

5.2 Experimental Design and Scheme

5.2.1 Experimental Variables and Evaluation Indicators

- Experimental variables: Control algorithms (traditional PID, commercial fuzzy PID, improved fuzzy PID (with G feedback), improved fuzzy PID (without G feedback));
- Evaluation indicators: (1) Assembly success rate (≥99% is qualified); (2) Repeated positioning accuracy (3σ criterion); (3) Dynamic response time; (4) Pressure distribution uniformity (coefficient of variation CV); (5) Environmental stability (-20℃~60℃); (6) Durability (10^6 cycles).

5.2.2 Experimental Design of Comparative Test Under the Same Conditions

To ensure the scientificity of the comparison, all tests were completed on the same experimental platform and under the same environmental conditions (25℃, 50%RH):

(1) Commercial systems include KUKA KR C4 (equipped with Siemens Sinumerik fuzzy PID controller) and ABB IRB 1200 (standard PID controller);

(2) Unified test process: Device grasping → visual positioning → pressure feedback adjustment → precision assembly → result judgment;

(3) Unified data statistical standard: 1000 assembly experiments were carried out for each algorithm/system, and abnormal values were eliminated by the 3σ criterion (abandonment ratio ≤0.3%). One-way analysis of variance (ANOVA) and post hoc multiple comparison (LSD test) were carried out by SPSS 26.0 with a significance level of $P<0.05$.

5.3 Experimental Results and Analysis

5.3.1 Benchmark Performance Test and Analysis of Variance

The experimental results show (Table 3) that when the improved fuzzy PID algorithm (with G feedback) is adopted, the assembly success rate reaches 99.6%, which is 16.4% higher than that of the traditional PID algorithm (83.2%), 4.5% higher than that of the commercial fuzzy PID algorithm (95.1%), and 2.3% higher than that of the version without G feedback (97.3%); the repeated positioning accuracy is ±0.012 mm (3σ), which meets the industrial requirement of ±0.015 mm.

Table 3. Comparative results of experiments with different algorithms/systems under the same conditions

Control algorithm/System	Assembly success rate (%)	Repeated positioning accuracy (±3σ, mm)	Dynamic response time (ms)	Pressure distribution CV (%)	Device damage rate (%)
Traditional PID algorithm (ABB)	83.2	0.025	6.8	16.8	3.5
Commercial fuzzy PID algorithm (Siemens)	95.1	0.018	4.1	10.5	1.2
Improved algorithm (without G feedback)	97.3	0.016	3.2	9.1	0.8
Improved algorithm (with G feedback)	99.6	0.012	2.8	7.3	0.4

Note: The ±0.005 mm of KUKA KR C4 is the no-load accuracy. In this experiment, in the scenario of 0402 micro-device (8 mg) grasping + pressure feedback, the positioning accuracy of KUKA is reduced to ±0.025 mm, mainly due to the flexible deformation of the gripper end and the contact force disturbance of the micro-device.

The results of one-way analysis of variance are shown in Table 4. The differences between groups are significant ($F=24.3$, $P<0.001$). The post hoc LSD test shows that the differences between the improved algorithm (with G feedback) and the other three groups are statistically significant ($P<0.01$).

Table 4. Analysis of variance (ANOVA) results (taking positioning accuracy as the dependent variable)

Source of variance	Sum of squares (SS)	Degrees of freedom (DF)	Mean square (MS)	F value	P value
Between groups	0.0012	3	0.0004	24.3	<0.001
Within groups	0.0165	3996	4.13×10^{-6}	-	-
Total	0.0177	3999	-	-	-

Failure case analysis shows that the failure of the traditional algorithm is mainly due to material slipping during small pressure grasping (68%) and overshoot during high-speed positioning (32%); the failure of the commercial algorithm is concentrated in device offset caused by uneven pressure distribution (75%); the failure of the improved algorithm without G feedback is mostly due to fine-tuning lag caused by sudden pressure changes on the contact surface (60%); while the failure rate of only 0.4% of the improved algorithm with G feedback is due to the dimensional deviation of the device itself.

5.3.2 Pressure Gradient Feedback Ablation Experiment

To quantify the technical contribution of pressure gradient feedback, an ablation experiment was designed to compare the performance differences of the improved algorithm under the conditions of "with/without G feedback", and the results are shown in Table 5. The contribution of pressure gradient feedback to positioning accuracy is $\frac{0.016-0.012}{0.016-0.008}\times100\%=50\%$, and the contribution to assembly success rate is $\frac{99.6\%-97.3\%}{99.6\%-83.2\%}\times100\%=14\%$, verifying the necessity and effectiveness of the core innovation.

Table 5. Results of pressure gradient feedback ablation experiment

Experimental condition	Positioning accuracy (±3σ, mm)	Assembly success rate (%)	Dynamic response time (ms)	Pressure distribution CV (%)	Technical contribution
Without G feedback	0.016	97.3	3.2	9.1	-
With G feedback	0.012	99.6	2.8	7.3	50% for positioning accuracy; 14% for success rate

5.3.3 Environmental Stability Test

The results of the wide temperature range experiment show that the improved algorithm (with G feedback) maintains an assembly success rate of more than 98.5% in the range of -20℃ to 60℃, the repeated positioning accuracy fluctuation is ≤±0.002 mm, and the sensitivity drift is ≤±1.2%FS. Among them, the dynamic response time increases slightly (3.5 ms) in the low temperature environment of -20℃, but it still meets the industrial operation requirements; the pressure resolution is maintained within 0.01 N in the high temperature environment of 60℃ without obvious performance attenuation.

Humidity influence experiments show that in the range of 20%~80%RH, the system assembly success rate is ≥99.0%, and the output error is ≤±0.8%FS, which proves that the system has good environmental adaptability.

5.3.4 Durability and Accelerated Life Test

The results of 10^6 cycles of durability tests show that the performance attenuation trend of the system is gentle: the pressure resolution changes from the initial 0.008 N to 0.009 N with an attenuation of 12.5%; the dynamic response time increases from 1.5 ms to 1.8 ms with an increase of 20%; the assembly success rate decreases from 99.6% to 98.2%, which still meets the industrial requirements. After the test, the surface of the sensitive layer was observed by SEM, and no obvious cracks and wear were found, and the integrity of the DLC coating was maintained above 95%.

To evaluate the long-term reliability, an accelerated life test (temperature 85℃, humidity 85%RH, pressure cycle frequency 5 Hz) was supplemented with a test duration of 1000 h (equivalent to 10^7 cycles at room temperature). The failure data was fitted by Weibull distribution with a shape parameter $\beta=2.3$, a characteristic life $\eta=1.2\times10^7$ cycles, and a mean time between failures (MTBF) $\geq1.0\times10^7$ cycles, meeting the long-term operation requirements of semiconductor equipment. The failure mode analysis is shown in Table 6:

Table 6. Failure mode analysis of accelerated life test

Failure mode	Number of failures	Proportion (%)	Main cause	Improvement scheme
Electrode peeling	3	23.1	Insufficient bonding force between DLC coating and electrode interface	Increase plasma treatment time to 8 min
Sensitive layer crack	1	7.7	Fatigue aging of PDMS material	Add 1 wt% silane coupling agent to improve toughness
Signal drift	9	69.2	Oxidation of CNT network and humidity penetration	Adopt vacuum packaging to improve water and oxygen barrier rate

5.3.5 Empirical Verification in Industrial Scenarios

A 30-day empirical test was carried out on the production line of a semiconductor packaging enterprise, completing a total of 100,000 assembly operations of 0402 packaged resistors. The average assembly efficiency reached 3600 pieces/hour, an increase of 20% compared with the original equipment; the product defective rate was reduced from 1.8% to 0.3%, saving the enterprise about 1.2 million RMB in production costs annually. Field tests show that the system can still operate stably in an industrial environment with dust concentration of 0.1~0.5 mg/m³ and electromagnetic interference intensity ≤40 V/m without fault shutdown.

5.4 Comparison with Similar Research Results

The system in this paper is compared with the latest research results at home and abroad and commercial products in terms of performance (Table 7), and the differences in experimental conditions are clearly marked to ensure the objectivity of the comparison:

Table 7. Performance comparison of similar research results and commercial products (with experimental conditions marked)

Research team/Manufacturer	Experimental conditions (Temperature/Humidity)	Test task	Sensing node density (nodes/cm²)	Pressure resolution (N)	Dynamic response time (ms)	Positioning accuracy (±3σ, mm)	Assembly success rate (%)	Cycle durability (times)
MIT (Rus D, Tolley M T, Firoozi A, et al., 2018)	25℃/50%RH	Object grasping (≥1g)	16	0.1	8	0.035	90.2	1×10^5
Stanford University (Zhang Y, Kim S, Park H, et al., 2020)	25℃/50%RH	Flexible material grasping	20	0.05	5	0.025	95.1	5×10^5
Harbin Institute of Technology (Liu J, Wang H, Li D, et al., 2021)	25℃/50%RH	Mechanical part assembly (≥5g)	12	0.08	6	0.030	92.3	3×10^5
Siemens (Schmidt M, Verl A & Grebenstein M, 2019)	25℃/50%RH	0402 device assembly	10	0.06	4	0.022	94.8	1×10^6
ABB (ABB Robotics, 2023)	25℃/50%RH	0402 device assembly	8	0.1	7	0.028	93.5	2×10^6
This research	25℃/50%RH	0402 device assembly	32	0.008	1.5	0.012	99.6	1×10^7 (accelerated equivalent)

Explanation of experimental condition differences:

1) All benchmarks in this paper are completed on the same KUKA platform to eliminate performance deviations caused by mechanical structure differences;
2) The data of MIT/Stanford are from literatures, and the original test tasks are mostly grasping large-mass objects (≥1g), which cannot be directly compared with the difficulty of 0402 micro-device (8 mg) assembly in this paper;
3) This paper additionally verifies the wide temperature range (-20~60℃) and 10^7 cycles of durability (accelerated equivalent), while other studies are mostly room temperature single tests with a more clear reliability boundary;
4) The performance advantage is mainly reflected in the "micro-assembly working condition", which comes from the sensing-control optimization coupled with pressure gradient, and does not represent the overall superiority in general scenarios.

6. Conclusions and Prospects

6.1 Research Conclusions

To address the key issues of low tactile sensing accuracy, slow dynamic response and poor sensing-control coordination in precision robot control, this paper conducted systematic research on multi-node pressure sensing arrays and control algorithms, yielding four core outcomes: (1) A CNT/PDMS composite sensitive layer percolation model was established, and a 24-node high-density MEMS piezoresistive array (32 nodes/cm²) was fabricated, achieving 0.008 N pressure resolution and ≤1.5 ms dynamic response; DLC coating and vacuum packaging enabled the array to withstand 10^7 accelerated cycle tests and maintain ≤±1.2%FS sensitivity drift at -20℃~60℃. (2) A standardized anti-interference hardware architecture was built, with the four-stage signal conditioning circuit and FPGA high-speed acquisition realizing 48 dB sensing SNR and ≤50 μs control command delay, providing a standardized solution for industrial high-precision data acquisition. (3) Pressure distribution gradient was first used as the third-dimensional input for fuzzy PID to construct a 3D fuzzy decision space, whose stability was proven via discrete Lyapunov theory and LMI method; the system achieved ±0.012 mm repeat positioning accuracy and 99.6% assembly success rate in micro-assembly, with pressure gradient feedback contributing 50% to positioning accuracy and outperforming commercial systems significantly. (4) Ablation tests, identical-condition comparative experiments, accelerated life tests and a 30-day industrial validation fully verified the system's scientificity, effectiveness and practicability. The related technologies have obtained 4 authorized patents and 3 software copyrights, offering a high-performance tactile control solution for high-end manufacturing.

6.2 Research Limitations and Future Directions

This research has three limitations: (1) The sensitive layer's stability in strong corrosive environments (e.g., chlorine/sulfur-containing gases) awaits further verification; (2) The control algorithm does not account for interactive effects in multi-robot collaborative operations; (3) The high single-set hardware cost (≈$8000) restricts large-scale popularization. Corresponding future research directions are as follows: (1) Develop low-cost sensitive layer materials (e.g., graphene/cellulose composites) to reduce hardware costs to below $3000; (2) Expand the algorithm's multi-robot collaborative control capability and integrate machine vision-tactile sensing data fusion for complex micro-assembly tasks; (3) Optimize the array's packaging structure with ceramic-based packaging to enhance adaptability in extreme industrial environments such as strong corrosion and high humidity; (4) Promote technical standardization by participating in formulating industry specifications for robotic tactile sensing systems and expand the industrial influence of the research results.

References

ABB Robotics. (2023). Product Specification: IRB 1200 Industrial Robot. Zurich: ABB Group.

Åström K J, Murray R M. (2021). *Feedback systems: An introduction for scientists and engineers*. Princeton University Press.

Chen L, Wang H, Li D, et al. (2021). Diamond-like carbon coating for improving the durability of flexible tactile sensors. *Surface and Coatings Technology, 415*, 127189.

Dorf R C, Bishop R H. (2022). *Modern control systems*. 15th ed. Boston: Pearson Education Inc, 2022.

Gao H, Wang S, Zhang Y, et al. (2020). Design and fabrication of high-performance flexible tactile sensors: A review. *IEEE Sensors Journal, 20*(24), 14781-14798.

Grand View Research. (2024). Industrial Robotics Market Size Report, 2030.

Kim B, Park J, Lee J, et al. (2021). Recent advances in flexible tactile sensors: Materials, structures, and applications. *Small, 17*(3), 2007029.

Kim B, Park J, Lee J, et al. (2023). Reinforcement learning-based tactile control for robotic assembly. *IEEE*

Transactions on Robotics, 39(2), 1123-1138.

Lee S, Kim J, Park S, et al. (2021). Piezoresistive tactile sensors based on polymer composites for robotic applications. *Composites Science and Technology, 207*, 108789.

Li H, Wang S, Zhang Y, et al. (2024). Adaptive PID control for precision robotic manipulation with tactile feedback. *Mechatronics, 92*, 103189.

Liu G, Chen W, Zhang J, et al. (2022). Tactile sensing for robotic manipulation: A review. *Advanced Materials Technologies, 7*(4), 2101184.

Liu J, Wang H, Li D, et al. (2021). Design and application of a flexible piezoresistive tactile sensor array for robotic grasping. *IEEE Transactions on Industrial Electronics, 68*(11), 11183-11192.

Rus D, Tolley M T, Firoozi A, et al. (2018). Soft robots with proprioception via embedded soft sensors. *Science Robotics, 3*(21), eaar7807.

Schmidt M, Verl A, Grebenstein M. (2019). Force-torque sensing in robotic manipulation: A review. *Robotics and Autonomous Systems, 117*, 103322.

SEMI. (2024a). Advanced Packaging Market Trends 2024.

SEMI. (2024b). Semiconductor Equipment Market Statistics 2024.

Shi Z, Li Y, Zhang J, et al. (2022). Neural network-based adaptive PID control for robotic manipulators. *Neural Computing and Applications, 34*(12), 9879-9892.

Wang X, Chen Y, Zhang L, et al. (2020). Graphene-based flexible tactile sensor with high sensitivity and wide linear range. *ACS Applied Materials & Interfaces, 2020,* (23), 26201-26209.

Wang Y, Li C, Zhang L. (2021). A review of piezoresistive tactile sensors for robotic applications. *Sensors and Actuators A: Physical, 321*, 112458.

Yu X, Chen W, Liu J, et al. (2022). Node layout optimization of tactile sensor array for robotic grasping. *IEEE Transactions on Robotics, 38*(2), 1034-1046.

Zhang H, Li Y, Wang Z, et al. (2022). Environmental stability of flexible tactile sensors: A review. *Journal of Materials Chemistry C, 10*(15), 5823-5842.

Zhang Y, Kim S, Park H, et al. (2020). High-resolution flexible tactile sensor array based on carbon nanotube/polydimethylsiloxane composite. *Journal of Microelectromechanical Systems, 29*(3), 456-464.

Zhang Y, Li W, Wang Z, et al. (2020). Carbon nanotube/polydimethylsiloxane composite for flexible tactile sensors: A review. *Journal of Materials Science & Technology, 36*(12), 1-15.

Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
MAR. 2026 VOL.5, NO.1

Machine Learning: A Brief Review for the Beginners

Haradhan Kumar Mohajan[1]

[1] Chairman and Associate Professor, Department of Mathematics, Premier University, Chittagong, Bangladesh

Correspondence: Haradhan Kumar Mohajan, Chairman and Associate Professor, Department of Mathematics, Premier University, Chittagong, Bangladesh.

doi:10.63593/IST.2788-7030.2026.03.004

Abstract

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on developing models, studies statistical algorithm, teaches the systems to think and understand like humans by learning from the data, and performs tasks without explicit instructions. It is one of the most relevant technologies of the 21st century that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It opens an entirely new realm of what humans can do with computers and other machines. It describes the capacity of systems to learn from problem-specific training data to automate the process of analytical model building and solve associated tasks. It can enable an organization to autonomously learn and improve using neural networks and deep learning (DL), without being explicitly programmed, by feeding it large amounts of data. This paper tries to discuss elementary ideas of machine learning for the benefit of the new researchers in this field.

Keywords: machine learning, deep learning, artificial intelligence, cyber-attacks, cyber-security

1. Introduction

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on enabling computers and machines to imitate the way that humans learn to perform the tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data (Alpaydin, 2020). It is for designing algorithms that allow a computer to learn. It aims at enabling machines to perform their jobs skillfully by using intelligent software. It is using in nearly every industry and business activity that helps the logistics industry optimize shipping and delivery routes, the retail industry personalize shopping experiences and manage inventory, manufacturers automate factories, and helps secure organizations everywhere (Fujii & Managi, 2018).

In the 1930s, American brilliant electrochemical expert Thomas Neil Ross (1909-2010) made the first attempt to develop a machine that simulated the behavior of a living creature in performance (Ross, 1938). The term machine learning was first coined in 1959 by American pioneer in the field of computer gaming and artificial intelligence Arthur Lee Samuel (1901-1990) (Samuel, 1959). In recent years ML has grown rapidly in the context of data analysis and computing that typically allows the applications to function in an intelligent manner. Actually, the ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically programmed and is generally referred to as the most popular latest technologies in the fourth industrial revolution (Sarker, 2021). The ML is used in web search, drug design, spam filters, credit scoring, fraud detection, recommender systems, ad placement, stock trading, and many other applications (Domingos, 2012).

2. Literature Review

A literature review is an overview of previously published works on a particular topic. It provides the researchers general information of an existing knowledge of a particular topic (Bolderston, 2008). It is a comprehensive survey of scholarly sources on a specific topic that provides an overview of current knowledge, which

synthesizes, analyzes, and critically evaluates existing research to identify key themes, debates, and gaps in the literature (Galvan, 2015). A good literature review has a proper research question, a proper theoretical framework, and a chosen research methodology (Creswell, 2013). Arthur Lee Samuel has been investigated two ML procedures in some detail using the game of checkers, and has observed that a computer can be programmed so than a programmer. It can be done in a very short period of time when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters (Samuel, 1959). Pedro Domingos has shown that ML is widely used in computer science and other fields. He has summarized some key lessons, such as pitfalls to avoid important issues to focus on, and answers to common questions (Domingos, 2012).

Tom Mitchell has described that ML as a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E (Mitchell, 1997). Niklas Kühl and his coworkers have wanted to clarify the relationship between ML and DL, and to specify the contribution of machine learning to artificial intelligence. They have presented a conceptual framework which clarifies the role of ML to build AI agents with the more terminological clarity and a starting point for interdisciplinary discussions and future research (Kühl et al., 2019). Iqbal H. Sarker has presented a comprehensive view on ML algorithms that can be applied to enhance the intelligence and the capabilities of an application through the explaining the principles of different ML techniques and their applicability in various real-world application domains, such as cyber security systems, smart cities, healthcare, e-commerce, agriculture, etc. (Sarker, 2021).

Christian Janiesch and his coauthors have summarized the fundamentals of ML and DL to generate a broader understanding of the methodical underpinning of current intelligent systems. They have provided a conceptual distinction between relevant terms and concepts, explain the process of automated analytical model building through ML and DL, and discuss the challenges that arise when implementing such intelligent systems in the field of electronic markets and networked business (Janiesch et al., 2021). Juan D Pineda-Jaramillo has provided a brief explanation of some ML algorithms commonly used for transportation research, such as artificial neural networks (ANN), decision trees (DT), support vector machines (SVM) and cluster analysis (CA). Later, the characterization of ML algorithms is discussed and random forest (RF), a variant of decision tree algorithms, is presented as the best methodology for modeling travel mode choice (Pineda-Jaramillo, 2019). Hamed Alqahtani and his coworkers have studied various popular ML classification algorithms, such as Bayesian network, naive Bayes classifier, decision tree, random decision forest, random tree, decision table, and artificial neural network to detect intrusions due to provide intelligent services in the domain of cyber-security (Alqahtani et al., 2020).

3. Research Methodology of the Study

Research is a searching for knowledge and truth. It is a creative and systematic work undertaken to increase the stock of knowledge that involves the collection, organization, and analysis of evidence to increase understanding of a topic (Grover, 2015). The primary purposes of research are documentation, discovery, interpretation, the research and development (R&D) of methods, and systems for the advancement of human knowledge (Song et al., 2010). Methodology is the study of research methods that refers to the philosophical discussion of associated background assumptions. It is divided into quantitative and qualitative research areas. Quantitative research is the main methodology of the natural sciences that uses precise numerical measurements (Adams et al., 2007). On the other hand, qualitative research is more characteristic of the social sciences, such as surveys, interviews, focus groups, and the nominal group technique that aim more at an in-depth understanding of the meaning of the studied phenomena and less at universal and predictive laws (Berg, 2009). Also, many social scientists use mixed-methods research that combines quantitative and qualitative methodologies (Creswell, 2013, Mohajan, 2018b, 2020). A research methodology is a way of explaining how a researcher intends to carry out the research (Kara, 2012). It describes the techniques and procedures used to identify and analyze information regarding a specific research topic (Eyler, 2020). It provides a detailed plan that helps to keep researchers on track, making the process smooth, effective, and manageable (Mohajan, 2017; Groh, 2018).

4. Objective of the Study

Machine learning (ML) is a subfield of artificial intelligence (AI) that replaces the need for developing computer programs manually and lets computers to create programs themselves from the data. It also relates broadly with many fields, such as statistics, mathematics, physics, theoretical computer science, etc. (Koza et al., 1996). It enables machines to make predictions, perform clustering, extract association rules, and make decisions from a given dataset. It studies computer algorithms for learning to do stuff, complete a task, make accurate predictions, and behave intelligently that are being done always based on some sort of data, such as examples, direct experience, and instruction. The goal of ML is to devise learning algorithms that does the learning automatically without human intervention. Hence, without codifying knowledge into computers, ML seeks to automatically learn meaningful relationships and patterns from examples and observations (Janiesch et al., 2021). Main objective of this article is to provide introductory ideas of ML. Other minor objectives of the study are as follows

(Mohajan, 2018a):

1) to highlight on overview and types of ML,
2) to focus on common algorithms of ML, and
3) to discuss application and importance of ML.

5. An Overview of ML

Machine learning and artificial intelligence, as well as the terms data mining, deep learning and statistical learning are related, often present in the same context and sometimes used interchangeably (Bousquet et al., 2011). The ML is defined as an application of AI where available information is used through algorithms for processing the statistical data. It involves concepts of automation, a high level of generalization to get a system, and requires human guidance (Mohri et al., 2012). In brief, ML is the process of turning data into programs. From a statistical point of view, ML can be regarded as an implementation of statistical learning. But, in computer science, it has the focus of designing efficient algorithms to solve problems with computational resources (Brink, 2017). The ML algorithms are organized into taxonomy, based on the desired outcome of the algorithm. Common algorithm types are supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, etc. (Anzai, 2012). The ML is used to define a group of methods or algorithms that allow computers to mechanize data driven model programming and build models by means of a methodical detection of patterns in statistically significant data (Bhavsar et al., 2017).

6. Types of ML

Machine learning is about designing algorithms that allow a computer to learn. There are three kinds of models used in ML: i) supervised learning, ii) unsupervised learning, and iii) reinforcement learning. I have also highlighted on the other two types, such as semi-supervised learning and transduction inference. Among these five types, supervised learning is considered as the most developed branch of ML (Alpaydin, 2020).

6.1 Supervised Learning

Supervised learning is a subcategory of machine learning model (MLM) in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answers. It uses labeled training data and a collection of training examples to infer a function. Therefore, we are given a labeled training dataset from which a machine learning algorithm can learn a model that can predict labels of unlabeled data points. It is carried out when certain goals are identified to be accomplished from a certain set of inputs. For example, given a corpus of spam and non-spam email, a supervised learning task would be to learn a model that predicts to which class, spam or non-spam, and new emails belong (Sarker, 2021).

It uses structured data to map a specific feature to a label, where the output is known accurately, and the model is trained on data of the known output. For example, names, dates, addresses, credit card numbers, stock information, geo-location, etc. are structured data. Therefore, it covers three main portions; labeled data, direct feedback, and predict outcome (Hirt et al., 2017). It indicates the presence of a supervisor or a teacher. It always aims to build a model by applying an algorithm on a set of known data points to gain insight on an unknown set of data. Therefore, it has a well-defined structure, conforms to a data model following a standard order that is highly organized and easily accessed, and used by an entity or a computer program. The supervised tasks are "classification" that separates the data, and "regression" that fits the data (Sarker et al., 2020).

Mathematical Notations

We define a function that use to approximate some unknown function,

$$y = f(x),$$

where x is a vector of input features associated with a training example and y is the outcome we want to predict. In classification, we define the hypothesis function as,

$$h: X \to Y$$

where $X = R^m$ and $Y = \{1,2,...,k\}$ with class labels k. In regression, the task is to learn a function,

$$h: R^m \to R.$$

Given a training set

$$D = \left\{ \left\langle \mathbf{x}^i, \mathbf{y}^i \right\rangle, i = 1,2,...,m \right\}$$

where $\mathbf{x} \in X$ and $\mathbf{y} \in Y$, and the training pairs $\langle \mathbf{x}^i, \mathbf{y}^i \rangle$ are drawn from a joint distribution $P(X,Y) = P(X)P(Y/X) = P(Y)P(X/Y)$. Here $\mathbf{x} \in X$ is a training example with *m* features, such as height, weight, age, etc. of a person, represented as a column vector,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_m \end{bmatrix}.$$

Similarly, $\mathbf{y} \in Y$ can be represented as a column vector,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}.$$

Supervised learning can be classified into two categories of algorithms; i) classification: a classification problem is used when the output variable is a category, where the outputs are discrete labels, as in spam filtering, such as red or blue, disease or no disease, and ii) regression: a regression problem is used when the output variable is a real-valued, such as dollars or weight (Mohammed et al., 2016).

6.2 Unsupervised Learning

Unsupervised learning is a MLM that uses unstructured data to learn patterns without the need for human interference, where the correctness of the output is not known ahead of time. It mainly deals with the unlabeled data and tries to analyze and discover patterns within (Han et al., 2011). It is widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes. The algorithm learns from the data without human input and categorizes it into groups based on attributes. It allows one to perform more complex processing tasks compared to supervised learning. Therefore, it covers three main portions; no labels, no feedback, and find hidden structure in data (Hirt et al., 2017).

Unsupervised learning is good at descriptive modeling and pattern matching. The most common unsupervised learning algorithms used present are fuzzy means, k-means clustering, hierarchical clustering, and partial least squares. It is classified into two categories of algorithms: clustering and dimension reduction (Sarker et al., 2020). It is used by clustering algorithms to find patterns in data so that it can be grouped, and is also used for dimensionality reduction, such as compressing data onto a lower-dimensional subspace or manifold. By spotting distinctions between data points that humans have missed, computers can aid data scientists. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, finding association rules, anomaly detection, etc. (Mohammed et al., 2016).

6.3 Reinforcement Learning

Reinforcement learning is a MLM that can be described as “learn by doing” through a series of trial and error experiments. It aims at using observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk (Mohammed et al., 2016). An agent, such as a robot or controller learns to perform a defined task through a feedback loop until its performance is within a desirable range. Instead of having correct or false label for each step, the learner must discover or learn a behavior that maximizes the reward for a series of actions. The agent receives positive reinforcement when it performs the task well and negative reinforcement when it performs poorly (Raschka & Mirjalili, 2017). In that sense, it is not a supervised setting and somewhat related to unsupervised learning; however, reinforcement learning really is its own category of machine learning. Therefore, it covers three main portions; decision process, reward system, and learn series of actions (Hirt et al., 2017). Typical applications of reinforcement learning involve playing games, such as chess, Atari video games; and some form of robots, such as drones, warehouse robots, and self-

driving cars (Raschka & Mirjalili, 2017).

6.4 Semi-Supervised Learning

The semi-supervised learning can be described as a mix between supervised and unsupervised learning, as it operates on both labeled and unlabeled data, where the algorithm must figure out how to organize and structure the data to achieve a known result (Han et al., 2011). In semi-supervised learning tasks, some training examples contain outputs, but some do not. We then use the labeled training subset to label the unlabeled portion of the training set, which we then also utilize for model training. For instance, in the ML model it is told that the result is a pear, but only some training data is labeled as a pear. It is applied in machine translation, fraud detection, labeling data, and text classification. Actually, the labeled data could be rare in several contexts, and unlabeled data are numerous, and semi-supervised learning is useful to provide a better outcome for prediction than that produced using the labeled data alone from the model (Mohammed et al., 2016).

6.5 Transduction Inference

Transduction inference is reasoning from observed, specific training cases to specific cases. It is similar to supervised learning, but does not explicitly construct a function. Instead, it tries to predict new outputs based on training inputs, training outputs, and new inputs. It was introduced in a computer science context by Russian statistician, researcher, and academician Vladimir Vapnik in the 1990s (Vapnik & Kotz, 2006). Transduction algorithms can be broadly divided into two categories: those that seek to assign discrete labels to unlabeled points, and those that seek to regress continuous labels for unlabeled points (de Finetti, 1970).

7. Common Algorithms of ML

There are numerous ML algorithms in use, and the most common learning algorithms used at present are linear classifiers (logical regression, naïve Bayes classifier, perceptron, and support vector machine), quadratic classifiers, k-means clustering, boosting, decision tree, random forest, neural networks, and Bayesian networks. Now we will describe neural networks, k-means clustering, decision trees, and random forests in briefly (Cao, 2017).

7.1 Neural Networks

Neural network has a massive number of connected processing nodes that is inspired by the structure and functions of biological neural networks, and mimics how the human brain functions (Bishop, 2006). It consists of connected units or nodes called artificial neurons, which loosely model the neurons in the brain. It is actually performing a number of regression and classification tasks at once, although commonly each network performs only one (Bishop, 1995). *American neurophysiologist and cybernetician* Warren S. Mcculloch (1898-1969) and *American logician* Walter Pitts (1923-1969) introduced the concept of artificial neural network (ANN), and it was designed to simulate the functions and structure of the nervous systems in living beings (McCulloch & Pitts, 1943). The ANNs are composed of a large number of neurons that are interconnected in parallel and work in unison to solve diverse problems (Bishop, 1995).

In most of the cases, the network will have a single output variable, although in the case of many-state classification problems, this may correspond to a number of output units (Gurney, 1997). The neurons are aggregated into different layers and may perform different transformations on their inputs. Signals travel from the input layer to the output layer, possibly passing through multiple intermediate hidden layers (Dawson, 1998). Neural networks are used to extract complex patterns from the data, and perceive trends that are too complex to be observed by humans or other computer methods with their outstanding ability to derive meaning from data that is complex or inaccurate (Bishop, 1995). These are effective at identifying patterns and are crucial in applications, such as speech recognition, image creation, natural language translation, and image recognition (Bhadeshia, 1999). These are very powerful tools that have been used for numerous applications, such as medicine, transportation, optimization, and even quantum physics (Cantarella & de Luca, 2003).

7.2 K-Means Clustering

Clustering problems are seen in many different applications, such as data compression and vector quantization, data mining and knowledge discovery, and pattern recognition and pattern classification (Fayyad et al., 1996). The k-means clustering is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem through a certain number of clusters (assume k-clusters) fixed a priori. It does the three steps to convergence, determine the center coordinate, determine the distance of each object to the center, and group the object based on minimum distance (Bishop, 1995). The main idea is to define k centroids, one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed (Alsabti et al., 1998). Then re-calculate k new centroids, and if a loop has been generated, centroids move no more. The k-means is a simple algorithm that has been adapted to many problem domains. It aims at minimizing an objective function (Kanungo et al., 2004).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centers.

7.3 Decision Trees

A decision tree (DT) is a flowchart-like structure and a non-parametric method that is oriented graphs formed by a finite number of nodes departing from the root nodes. It is a decision support recursive partitioning structure that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility (von Winterfeldt & Ward, 1986). It simulates a real tree, which begins at a wide trunk, and as it rises is divided into narrower branches and the paths from root to leaf represent classification rules. These are powerful algorithms for classifying data, where a tree structure is used for modeling the different relationships between the features and potential output data. When a final decision is made, the DT ends with leaf nodes (Bhavsar et al., 2017).

Decision trees are simple to understand and interpret; and can determine worst, best, and expected values for different scenarios (Xu et al., 2023). These are useful for both categorizing data and regression that are the prediction of numerical values, which are simple to validate and audit. Functioning of DT is easy to understand and interpret, and needs little data preparation from the user to build an optimal DT (de Oña, 2016). Decision trees are commonly used in operations research, specifically in decision analysis, statistics, data mining, and machine learning to identify a strategy most likely to reach a goal. The measurements used to evaluate decision trees are accuracy, sensitivity, specificity, precision, miss rate, false discovery rate, and false omission rate (Kamiński et al., 2017).

7.4 Random Forests

Random forests (RFs) are tree-based algorithms that are associated with bagging. These are machine learning algorithms that *use many decision trees to make better predictions* (Dessi et al., 2013). These are ensemble learning method for classification, regression, and other tasks that work by creating a multitude of decision trees during training (Ye et al., 2008). In a random forest, the machine learning algorithm combines the output from various decision trees to predict a value or category (Denisko & Hoffman, 2018). The RFs combine both the different set of data called bootstrap aggregation and also numerous features selection to predict the outcome (Alqahtani et al., 2020). The first algorithm for random decision forests was created in 1995 by Chinese computer scientist Tin Kam Ho using the random subspace method (Ho, 1995). An extension of this algorithm was developed by Leo Breiman and Adele Cutler in 2001 (Breiman, 2001).

8. Applications of ML

The ML can be applied in almost every aspect of our life, such as web search (Bing, Google), email spam detection, game playing, weather prediction, sports predictions, stock predictions, identify credit card fraud, medical diagnoses, drug design, product recommendations, face detection and matching, ATMs, language translation, fraud detection, sentiment analysis, customer segmentation, self-driven vehicles (e.g., cars, drones), etc. (Warner & Hirschberg, 2012). It can be applied to filter spam emails. Various ML models and algorithms can be used to solve complex data science and analytics problems using Bayes' theorem (Balducci et al., 2018). In medical field, it is widely used to predict mortality in injured patients of Trauma & Injury Severity Score (TRISS) using logistic regression. It can contribute in areas as disparate as helping in the treatment of chronic diseases, fighting climate change, and anticipating cyber security threats (Fatima & Pasha, 2017). Using ML in practice requires that we make use of our own prior knowledge and experimentation to solve problems (Mohri et al., 2012).

9. Importance of ML

The ML helps to identify fraud, security threats, personalization and recommendations, automated customer service through the chatbots, transcription and translation, data analysis, etc. It opens an entirely new realm of what humans can do with computers and other machines (Russell & Norvig, 2015). It can control autonomous vehicles, drones, and airplanes, augmented and virtual reality, and robotics efficiently. Different ML techniques are used to meet the challenges of growing travel demands, safety concerns, energy consumption, emissions, and environmental degradation (Abduljabbar et al., 2019).

10. Conclusion

Machine learning is a branch of AI that aims at enabling machines to perform their jobs skillfully by using intelligent software. It is a key technology driver that encompasses the intelligent power to harness the

knowledge from the available data. It enables computers to imitate and adapt human-like behavior. In this study, I have briefly discussed various types of machine learning methods that can be used for making solutions to various real-world issues. In this article, a comprehensive review of ML process and algorithms are presented. I believe that my study on ML will be helpful to the new researchers of academia and industry professionals, and decision-makers.

References

Abduljabbar, R., et al. (2019). Applications of Artificial Intelligence in Transport: An Overview. *Sustainability*, *11*(1), 189-190.

Adams, J., et al. (2007). *Research Methods for Graduate Business and Social Science Students*. New Delhi: SAGE Publications.

Alpaydin, E. (2020). *Introduction to Machine Learning*, (4th Ed.). MIT Press, Cambridge.

Alqahtani, H. (2020). Cyber Intrusion Detection Using Machine Learning Classification Techniques. In N. Chaubey, S. Parikh, & K. Amin (Eds.), *Computing Science, Communication and Security: First International Conference,* pp. 121-131. COMS2 2020 Gujarat, India, March 26-27, 2020.

Alsabti, K., et al. (1998). An Efficient k-Means Clustering Algorithm. In Proceedings of the First Workshop on High Performance Data Mining, Orlando, FL, March.

Anzai, Y. (2012). *Pattern Recognition and Machine Learning*. Elsevier.

Balducci, F., et al. (2018). Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. *Machines*, *6*(3), 38.

Berg, B. L. (2009). *Qualitative Research Methods for the Social Sciences* (7th Ed.). Boston MA: Pearson Education Inc.

Bhadeshia, H. K. D. H. (1999). Neural Networks in Materials Science. *ISIJ International*, *39*(10), 966-979.

Bhavsar, P., et al. (2017). Machine Learning in Transportation Data Analytics. In: Chowdhury, M., Apon, A. and Dey, K., (Eds.). *Data Analytics for Intelligent Transportation System*, pp. 283-307, Elsevier.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford, England: Oxford University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bolderston, A. (2008). Writing an Effective Literature Review. *Journal of Medical Imaging and Radiation Sciences*, *39*(2), 86-92.

Bousquet, O., et al. (2011). Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures, vol. 3176. Springer.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Brink, J. A. (2017). Big Data Management, Access, and Protection. *Journal of the American College of Radiology*, *14*(5), 579-580.

Cantarella, G., & de Luca, S. (2003). Modeling Transportation Mode Choice through Artificial Neural Networks. Fourth International Symposium on Uncertainty Modeling and Analysis (ISUMA), College Park, USA.

Cao, L. (2017). Data Science: A Comprehensive Overview. *ACM Computing Surveys*, *50*(3), 43.

Creswell, J. W. (2013). *Research Design: Qualitative, Quantitative, and Mixed Method Approaches* (4th Ed.). Thousand Oaks, California: SAGE Publications.

Dawson, C. W. (1998). An Artificial Neural Network Approach to Rainfall-Runoff Modelling. *Hydrological Sciences Journal*, *43*(1), 47-66.

de Finetti, B. (1970). *Theory of Probability: A Critical Introductory Treatment*. New York: John Wiley.

Denisko, D., & Hoffman, M. M. (2018). Classification and Interaction in Random Forests. *Proceedings of the National Academy of Sciences*, *115*(8), 1690-1692.

de Oña, J., et al. (2016). Transit Service Quality Analysis Using Cluster Analysis and Decision Trees: A Step Forward to Personalized Marketing in Public Transportation. *Transportation*, *43*(5), 725-747.

Dessi, N., et al. (2013). *Enhancing Random Forests Performance in Microarray Data Classification*. Conference on Artificial Intelligence in Medicine in Europe.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, *55*(10), 78-87.

Eyler, A. A. (2020). *Research Methods for Public Health*. New York: Springer Publishing Company.

Fatima, M., & Pasha, M. (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(1), 1-16.

Fayyad, U. M., et al. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Fujii, H., & Managi, S. (2018). Trends and Priority Shifts in Artificial Intelligence Technology Invention: A Global Patent Analysis. *Economic Analysis and Policy*, *58*, 60-69.

Galvan, J. L. (2015). *Writing Literature Reviews: A Guide for Students of the Social and Behavioral Sciences* (6th Ed.). Pyrczak Publishing.

Groh, A. (2018). *Research Methods in Indigenous Contexts*. New York: Springer.

Grover, V. (2015). Research Approach: An Overview. *Golden Research Thoughts*, *4*(5), 1-8.

Gurney, K. (1997). *An Introduction to Neural Networks*. UCL Press.

Han, J., et al. (2011). *Data Mining: Concepts and Techniques*. Amsterdam: Elsevier.

Hirt, R. et al. (2017). An End-to-End Process Model for Supervised Machine Learning Classification: From Problem to Deployment in Information Systems. In Maedche, A., vom Brocke, J., Hevner, A. (Eds.) in *Proceedings of the Design Science Research in Information Systems and Technology (DESRIST) 2017 Research-in-Progress*, pp. 55-63, Karlsruhe, Germany.

Ho, T. K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278-282.

Janiesch, C., et al. (2021). Machine Learning and Deep Learning. *Electronic markets*, *31*(3), 685-695.

Kamiński, B., et al. (2017). A Framework for Sensitivity Analysis of Decision Trees. *Central European Journal of Operations Research*, *26*(1), 135-159.

Kanungo, T., et al. (2004). A Local Search Approximation Algorithm for k-Means Clustering. *Computational Geometry*, *28*(2-3), 89-112.

Kara, H. (2012). *Research and Evaluation for Busy Practitioners: A Time-Saving Guide*. Bristol: The Policy Press.

Koza, J. R., et al. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In Gero, J. S., and Sudweeks, F. (Eds.), *Artificial Intelligence in Design '96*. Springer, Dordrecht.

Kühl, N., et al. (2019). *Machine Learning in Artificial Intelligence: Towards a Common Understanding*. Proceedings of the 52nd Hawaii International Conference on System Sciences, pp. 5236-5245, Grand Wailea, Maui, Hawaii.

McCulloch, W., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115-133.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Higher Education, New York.

Mohajan, H. K. (2017). Two Criteria for Good Measurements in Research: Validity and Reliability. *Annals of Spiru Haret University Economic Series*, *17*(3), 58-82.

Mohajan, H. K. (2018a). Aspects of Mathematical Economics, Social Choice and Game Theory. PhD Dissertation. University of Chittagong, Chittagong, Bangladesh.

Mohajan, H. K. (2018b). Qualitative Research Methodology in Social Sciences and Related Subjects. *Journal of Economic Development, Environment and People*, *2*(1), 19-46.

Mohajan, H. K. (2020). Quantitative Research: A Successful Investigation in Natural and Social Sciences. *Journal of Economic Development, Environment and People*, *9*(4), 50-79.

Mohammed, M., et al. (2016). *Machine Learning: Algorithms and Applications*. CRC Press, Taylor & Francis Group, New York.

Mohri, M., et al. (2012). *Foundations of Machine Learning*. MIT press.

Pineda-Jaramillo, J. D. (2019). A Review of Machine Learning (ML) Algorithms Used for Modeling Travel Mode Choice. *DYNA*, *86*(211), 32-41.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow*. Packt Publishing, Birmingham, Mumbai.

Ross, T. (1938). The Synthesis of Intelligence: Its Implications. *Psychological Review*, *45*(2), 185-189.

Russell, S. J., & Norvig, P. (2015). *Artificial Intelligence: A Modern Approach*, (3rd Ed.), Pearson Publisher.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 535-554.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2, 160.

Sarker, I. H., et al. (2020). Cybersecurity Data Science: An Overview from Machine Learning Perspective. *Journal of Big Data*, *7*(1), 1-29.

Song, F., et al. (2010). Dissemination and Publication of Research Findings: An Updated Review of Related Biases. *Health Technology Assessment*, *14*(8), 1-193.

Vapnik, V., & Kotz, S. (2006). Estimation of Dependences Based on Empirical Data. *Journal of the Royal Statistical Society Series D*, *41*(3), Publisher: Springer.

von Winterfeldt, D., & Ward, E. (1986). *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge.

Warner, W., & Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. *Proceedings of the Second Workshop on Language in Social Media*, pp. 19-26, Montréal, Canada.

Xu, N., et al. (2023). Predicting and Assessing Wildfire Evacuation Decision-Making Using Machine Learning: Findings from the 2019 Kincade Fire. *Fire Technology*, *59*(2), 793-825.

Ye, Y., et al. (2008). Feature Weighting Random Forest for Detection of Hidden Web Search Interfaces. *Journal of Computational Linguistics and Chinese Language Processing*, *13*, 387-404.

Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
MAR. 2026 VOL.5, NO.1

Cascading Resilience Through Predictive Multi-Dimensional Safeguards: System Stability Architecture for Billion-Scale Concurrent Platforms

Yuheng Liu[1]

[1] Xiaohongshu Inc., Xi'an, Shaanxi 710032, China

Correspondence: Yuheng Liu, Xiaohongshu Inc., Xi'an, Shaanxi 710032, China.

doi:10.63593/IST.2788-7030.2026.03.005

Abstract

Modern billion-scale concurrent internet platforms suffer from explosive traffic bursts, multiplicative failure propagation, and resource contention spirals, while traditional static defense mechanisms and reactive stabilization strategies lag in prediction, lack integrated state awareness, and fail to prevent cascading failures. This paper proposes CoReliance, a cascading resilience architecture empowered by predictive multi-dimensional safeguards for system stability in ultra-large-scale concurrent platforms. The framework integrates ensemble demand forecasting (TCN, seasonal decomposition, and causal feature fusion), state-coupled dynamic rate-limiting, reinforcement-learned hierarchical degradation, multi-modal fault detection, causal root-cause localization, and closed-loop autonomous recovery. It abandons isolated component defense and realizes proactive capacity pre-positioning, real-time adaptive regulation, progressive service degradation, and closed-loop verifiable recovery. Validated in 12-month production across two tier-1 platforms with over 1.2 billion users, CoReliance lifts system availability from 97.08% to 99.87%, reduces mean time to recovery (MTTR) by 86.5% to 103 seconds, cuts unplanned outages by 94.3%, prevents 31 major incidents, and achieves 490% annual return on investment with a 1.8-month payback period. This architecture provides end-to-end stability assurance for high-concurrency social commerce, ride-hailing, and similar large-scale internet systems.

Keywords: cascading resilience, predictive safeguard, system stability, billion-scale concurrency, adaptive rate-limiting, hierarchical degradation, autonomous recovery, ensemble forecasting

1. Introduction

1.1 The Emerging Challenge of Extreme-Scale Concurrency

Modern internet platforms face an unprecedented dichotomy: sustained high throughput interspersed with explosive demand bursts. A leading social commerce platform with 1.2 billion registered users experiences baseline traffic of 15,000 requests per second, yet encounters peaks of 38,000 RPS during promotional campaigns—a 2.5-fold amplification within minutes. Similar patterns manifest across ride-hailing networks (up to 127,000 RPS during peak hours) and financial transaction systems.

This non-stationary behavior defies traditional capacity planning models, which assume either steady-state or predictable load curves. The underlying mechanisms intensify the challenge:

1) **Multiplicative failure propagation**: Microservice architectures mean that a single component bottleneck cascades upstream, affecting services that initiated calls minutes prior.
2) **Resource contention spirals**: When database connection pools deplete, queuing delays compound, triggering application-level timeouts that paradoxically increase connection consumption.
3) **Asymmetric user impact**: Rate-limiting mechanisms, while protective for system integrity,

disproportionately affect paying users when indiscriminately applied.

Financial impact quantification reveals the severity: our analysis of 18-month operational logs across two tier-1 platforms indicates that each hour of degraded service (response time > 500ms affecting > 10% of users) correlates with 3–8 million RMB revenue loss through abandoned transactions, churn, and reputation damage. Notably, 71% of peak-period incidents were preceded by identifiable warning signals 15–45 minutes prior, yet went unaddressed due to lack of predictive-actuated defense mechanisms.

1.2 Why Existing Approaches Fall Short

Contemporary system stabilization employs layered defenses—rate limiting, circuit breakers, graceful degradation—but treats them as independent components. This fragmentation creates critical gaps:

Gap 1: Static-to-Dynamic Mismatch

Traditional rate-limiting applies fixed thresholds (e.g., 30,000 RPS global limit) computed offline during capacity planning. However, system capacity varies dynamically: during a database slow-query incident, safe throughput may drop to 8,000 RPS, yet the fixed limiter allows 30,000, causing backlog accumulation and eventual cascade. Conversely, when a redundant cache layer activates, capacity may increase to 45,000 RPS, but fixed limiters reject legitimate requests.

Gap 2: Reaction Lag

Monitoring systems detect failures through continuous metric polling (typical 10–30s intervals). Upon detection, humans must interpret root cause, decide remediation, and execute—consuming 3–8 minutes. During this window, user experience degrades, queued requests accumulate, and the system may enter irrecoverable states.

Gap 3: Absence of Integrated Assessment

Existing health checks evaluate metrics independently: “P99 latency is 800ms” and “error rate is 1.5%” are reported separately, without synthesizing implications. A system with high latency but low error rate requires different action than one with low latency but high errors due to systematic faults.

Gap 4: Insufficient Empirical Evidence

While Google, Facebook, and Alibaba have published system design papers, few provide comprehensive **comparative evaluations** across multiple strategies on identical production workloads, nor do they quantify the interaction effects between different defense layers.

1.3 Novel Contributions

This work introduces **CoReliance** (Cascading Resilience), a tightly-integrated framework departing from component-isolation models. Core contributions:

Contribution 1: Predictive Safeguard Architecture

Rather than react to failures, CoReliance **anticipates** demand 15–60 minutes ahead using ensemble learning combining temporal convolution networks (TCN), seasonal decomposition, and causal features (promotional calendars, viral indicators, regional patterns). This enables **pre-positioning** capacity, pre-tuning parameters, and pre-activating degradation policies before surges materialize.

Contribution 2: State-Coupled Dynamic Regulation

A novel rate-limiting mechanism that couples limiting decisions to a **real-time system state vector** encompassing latency percentiles, error rates, resource saturation across layers, and downstream service health. Unlike prior work that adjusts limits based on single metrics, this approach recognizes that a system with 30% CPU but 90% connection pool saturation requires different actions than 90% CPU with 30% connection pool.

Contribution 3: Reinforcement-Learned Degradation Sequencing

Rather than pre-defined degradation policies, an offline-trained policy network learns **optimal degradation sequences** specific to failure types and current system state. A cache failure triggers different degradation steps than a database failure, and the sequencing adapts to ongoing conditions.

Contribution 4: Closed-Loop Recovery Verification

Upon partial recovery, automated verification ensures that restored capacity is genuine (not temporary) before expanding traffic. This prevents **pseudo-recovery** scenarios where traffic briefly succeeds before re-failure.

Quantified Outcomes (12-month validation, 1.2B+ users):

- Availability: 97.1% → 99.87% (+2.77 percentage points)
- Mean failure duration: 7.3 min → 68 sec (−91.2%)
- Unplanned outages: 14/year → 0.8/year (−94.3%)

- Forecast-prevented failures: 31 incidents detected and preempted
- Cost efficiency: 23.4% reduction through optimized resource scheduling

2. System Model and Formal Framework

2.1 Multi-Layer Dependency Model

We model the backend as a directed service graph where nodes represent logical service clusters and edges represent inter-service calls:

$$G=(V,E,\mathrm{R},\mathrm{M})$$

where:

- $V=v_1,\ldots,v_n$: Service clusters (API gateways, business logic, data layers)
- $E\subseteq V\times V$: Dependency edges; $(u,v)\in E$ means cluster u calls cluster v
- $\mathrm{R}=r_1,\ldots,r_m$: Resource pools (database connections, thread pools, memory)
- $\mathrm{M}(t)$: Metrics at time t for each node (latency, error rate, throughput)

Traffic demand is modeled as a time-varying stochastic process:

$$Q(t) = Q\mathrm{trend}(t) + Q\mathrm{seasonal}(t) + Q\mathrm{anomaly}(t) + \epsilon(t)$$

where $Q\mathrm{trend}(t)$ captures long-term growth, $Q\mathrm{seasonal}(t)$ captures daily/weekly cycles, $Q\mathrm{anomaly}(t)$ captures event-driven spikes, and $\epsilon(t)$ is irreducible noise.

2.2 Stability Objective and Constraints

The primary objective maximizes sustained availability under resource constraints:

$$\max f(\mathrm{G},\mathrm{P}(t)) = \int 0\ T\ 1[\text{SLA met at time } t]\ dt$$

subject to:

- Latency constraint: $P_{99}(RT)\leq\tau_{\max}$ (typically 500ms for user-facing APIs)

Reliability constraint: P (ErrorRate > εmax) ≤ δtolerance

- (e.g., ≤ 1% probability of error rate exceeding 0.5%)
- Resource constraint: $\forall r_i\in\mathrm{R}$, utilization$(r_i)\leq C_i$ (80% as safety threshold)
- Graceful degradation: \text{AvailableCapacity}(t) \geq 0.5 \times Q(t) (maintain 50% capacity for core services even under stress)

2.3 Failure Propagation Dynamics

When service v_i experiences capacity constraint (e.g., database connection pool at 100%), outbound latency increases. Upstream services v_j that call v_i accumulate responses in buffers, and if buffer saturation exceeds threshold, v_j itself becomes resource-constrained. This creates a **failure wavefront** that propagates upstream.

We model this propagation as:

$$\mathrm{Latency}j\ (t + \Delta t) = \mathrm{Latency}j\ (t) + 1[\mathrm{CallDependent}(j,i)] \times \mathrm{CascadeDelay}(i,t)$$

where CascadeDelay(i,t) depends on how saturated service i is. This formulation clarifies why isolated metrics are insufficient: measuring only v_i's latency misses the upstream impact.

3. Predictive Demand Anticipation

3.1 Ensemble Forecasting Architecture

We employ a three-model ensemble with learned weighting:

Model A: Temporal Convolution Network (TCN)

TCN captures non-linear temporal dependencies better than RNNs for traffic patterns:

$$Q\text{^}\mathrm{TCN}(t) = \mathrm{TCN}\theta\ (\{Q(t-k),\ldots,\ Q(t-1)\})$$

trained on 24-month historical data. TCN's receptive field enables capturing hour-scale patterns (e.g., lunch rush, evening peak).

Model B: Seasonal-Trend Decomposition

MSTL (Multiple Seasonal-Trend decomposition using Loess) isolates:

- Daily seasonality (lunch hours, evening)
- Weekly seasonality (weekday vs. weekend)

- Yearly seasonality (holidays, promotional calendars)

$$Q(t) = T(t) + S\text{daily}(t) + S\text{weekly}(t) + S\text{yearly}(t) + \varepsilon(t)$$

where $Q\text{^MSTL}(t) = T\text{^}(t) + S\text{^ daily}(t) + S\text{^weekly}(t) + S\text{^ yearly}(t)$.

Model C: Causal Feature Integration

External features significantly improve predictions:

$$Q\text{^ causal}(t) = \text{XGBoost}(\{Q(t - k : t - 1)\} \cup \text{Xext})$$

Viral score is derived from social media trending: we track mention volume of platform-related keywords on external platforms, with 2–4 hour lag indicating incoming traffic surge.

Ensemble Integration:

$$Q\text{^ final}(t) = w\text{TCN}(t) \times Q\text{^TCN}(t) + w\text{MSTL}(t) \times Q\text{^MSTL}(t) + w\text{causal}(t) \times Q\text{^ causal}(t)$$

Weights $w_i(t)$ are learned via gradient boosting on validation set errors, with adaptive adjustment based on each model's recent (last 7 days) performance:

$$wi(t + 1) = \sum j \exp(-\text{MAPE}j(t)/\alpha) \exp(-\text{MAPE}i(t)/\alpha)$$

where a= 10% is the temperature parameter.

3.2 Empirical Forecast Performance

Table 1. Baseline Comparison

Forecast Horizon	Single-Model Baselines	Ensemble	Improvement
5 min (LSTM only)	MAPE 4.8%	3.1%	−35.4%
15 min (Prophet)	MAPE 7.2%	5.4%	−25.0%
30 min (XGBoost)	MAPE 9.1%	6.8%	−25.3%
60 min (Ensemble baseline)	MAPE 12.5%	8.9%	−28.8%

The ensemble significantly outperforms single-model baselines, with gains increasing at longer horizons where temporal patterns become ambiguous.

Case Study: Lunar New Year Campaign

Historical median peak: 22,000 RPS. Ensemble forecast issued 36 hours in advance: 37,800 ± 1,200 RPS. Actual peak: 38,200 RPS. Error: +1.0%, well within tolerance.

4. State-Coupled Adaptive Rate-Limiting

4.1 System State Vectorization

Rather than limit decisions based on a single metric, we construct a **state vector** synthesizing system health:

$\mathbf{s}(t)$ = [P99_RT(t), P95_RT(t), ErrorRate(t), CPUmax(t), Memorymax(t), ConnPoolusage (t), QueueDepth(t), DependencyHealth(t)] $T \in$ R 8

Each component is normalized to [0,1]:

$$si(t) = m - m\ i \max\ i \min$$

where mi min, mi max are respectively the 5th and 95th percentiles from 90-day historical windows.

4.2 Learned Rate-Limit Policy

Instead of hand-crafted rules, we train a policy network via offline reinforcement learning on historical incident data:

$$\pi(\text{limit}t|\mathbf{s}(t); \theta) = \mu(t) \times \exp(\sigma \cdot \pi\theta\ (\mathbf{s}(t)))$$

where $\mu(t)$ is the base limit derived from traffic forecast, $\exp\left(\cdot^{\sigma}_{\pi_\theta(s(t))}\right)$ is a learned multiplier (where σ=0.3 prevents extreme adjustments), and π_θ is a shallow neural network (2 hidden layers, 64 units each, trained on 18 months of operational data).

The reward signal during training was:

$$R(t) = 1.0 - - 2.\ 0.\ 0\ 5 + 0.3$$

if SLA met and no cascades

if SLA valid

if proactive limiting prevented future failure

Offline RL prevents live exploration risks.

4.3 User-Tier Differentiation

Beyond state-coupled limits, we apply fractional allocation favoring high-value users:

$$\text{Allocation}u = \text{BaseLimit} \times \text{Priority}(u) \times (1 + \text{Fairness\_Adjustment}(t))$$

where:

if $u \in$ VIP (annual spend > 5k RMB)

increments allocations for users recently throttled, ensuring temporal fairness.

Table 2. Baseline Comparison, Lunar New Year campaign

Method	Effective RPS	P99 Latency	Error Rate	User Satisfaction
No limiting	38,200	3,200 ms	15.3%	2.1/5.0
Static limit (30k)	29,800	950 ms	2.8%	3.4/5.0
Threshold-based adaptation (Verma, A., et al., 2015)	32,100	680 ms	1.2%	3.9/5.0
State-coupled + policy	35,400	310 ms	0.28%	4.7/5.0

The proposed method sustains 92.7% of peak traffic with near-baseline latencies, compared to 78.2% for static limits.

5. Hierarchical Intelligent Degradation

5.1 Service Tier Classification

We classify services across two dimensions:

Dimension 1: User Impact (critical ↔ optional)

- **Tier 1**: Authentication, payment confirmation, order persistence
- **Tier 2**: Primary functionality (product search, recommendations)
- **Tier 3**: Secondary features (user profiles, social signals)
- **Tier 4**: Non-critical (analytics, A/B test logging)

Dimension 2: Recoverability

- **Async-friendly**: Recommendations, notifications (can queue without user-visible impact)
- **Cache-compatible**: Product metadata, user profiles (stale data acceptable)
- **Real-time-required**: Payments, inventory (must be current)

5.2 Graduated Degradation Strategy

Upon health degradation, the system activates a sequence of strategies, not just one:

Stage 1: Asynchronization (Latency increasing, health score < 0.75)

Non-critical path operations defer execution:

- Trigger → Enqueue to Kafka → Async processing via Flink
- Response time reduction: 40–60%
- User experience: Imperceptible (results delivered via async notification)

Stage 2: Cache Amplification (Latency high, health score < 0.50)

Extend TTLs for cached data and reduce freshness requirements:

$$\text{NewTTL}(t) = \text{BaseTTL} \times (1 + \text{HealthGap}(t) \times \lambda)$$

where $\text{HealthGap}(t)=\max(0, 0.5-H(t))$ and $\lambda=5$ (i.e., if health drops 20 points below threshold, TTL extends 5×).

Cache hit rate improvement: typically 15–25% additional hits.

Stage 3: Feature Suppression (System critical, health score < 0.30)

Disable Tier 3 and Tier 4 services, redirecting requests:

Response degraded = {core_field_1, core_field_2, …, null_placeholder}

Tier 1 and Tier 2 services continue, but with reduced dependencies.

Stage 4: Fallback Mode (System critical, health score < 0.10)

Route to read-only replica or local cache; transactions temporarily blocked with user notification.

5.3 Empirical Impact

Table 3. Scenario: Database latency surge to 2.0 seconds

Degradation Stage	System Throughput	P99 Latency	User Timeout Rate
No degradation	8,200 RPS (21%)	5,800 ms	28.3%
Stage 1 only	14,500 RPS (38%)	2,100 ms	8.2%
Stages 1+2	24,100 RPS (63%)	420 ms	0.8%
Stages 1+2+3	31,800 RPS (83%)	240 ms	0.1%

Progressive activation allows fine-tuned response rather than binary on/off decisions.

6. Autonomous Failure Recovery Engine

6.1 Multi-Modal Fault Detection

Failures manifest differently. We employ multi-modal detection combining four orthogonal signals:

Signal 1: Statistical Anomalies (robust to noise)

$$\text{Anomaly1}\ (t) = |M(t) - \mu_{t-48h:t-2h}| > 2.5 \times \text{MAD}_{t-48h:t-2h}$$

where MAD is median absolute deviation (robust to outliers). Detects sustained shifts (e.g., error rate drifting from 0.2% to 1.5%).

Signal 2: Rate-of-Change Spikes (immediate response)

$$\text{Anomaly2}\ (t) = |\nabla M(t)| > \text{threshold}_{\text{slope}}$$

captures sudden degradation within seconds (e.g., CPU spiking from 40% to 95% in 10 seconds).

Signal 3: Correlation Breaks (captures cascades)

Historical correlation between service A and B latencies is $\rho_{A,B}$=0.45. If correlation suddenly drops to 0.05, indicates decoupling (e.g., B failing, A's requests to B queuing without correlation).

Signal 4: Dependency Graph Anomalies (structural)

If service A's error rate is 0.1% but all callers of A see 5% errors, indicates A is healthy but callers have issues.

Fusion: Detection fires if ≥ 2 of 4 signals trigger, reducing false positives.

6.2 Causal Root-Cause Localization

Upon detection, rather than blanket mitigation, we identify the primary fault location via layered analysis:

```
Step 1: Identify degraded service(s)
        e.g., {API-gate, DB-conn-pool} show anomalies

Step 2: Perform correlation analysis
        - High error rate but low latency → client-side issue
        - High latency but low error rate → resource saturation
        - Both high → compound failure

Step 3: Check resource constraints
        - DB connections at 100% → DB bottleneck
        - CPU at 30% but queue depth high → threading issue
        - Disk I/O high → storage bottleneck

Step 4: Identify primary root
        Choose the component with most downstream impact

Step 5: Select recovery action
        Different roots warrant different actions
```

6.3 Canary-Based Recovery Protocol

Rather than flip a switch from broken → working, we employ graduated restoration:

$$Q\text{restored}(t) = Q\text{baseline} \times \alpha(t),$$

$\alpha(t) \in \{0.05, 0.25, 0.50, 1.0\}$

Canary 1 (5% traffic, 2 min): Verify that the fix actually addresses the root cause. If error rate remains elevated, escalate to human operator.

Canary 2 (25% traffic, 3 min): Expand to detect second-order issues (e.g., the fix works for 5% but degrades when scaled).

Canary 3 (50% traffic, 5 min): Near-full load test.

Full Rollout: Only if all canaries show healthy metrics.

Metric for Success: Error rate < 0.5%, P99 latency < 600ms, dependency health > 0.7.

6.4 Measured Recovery Improvement

Table 4. Comparative MTTR (Mean Time to Recovery)

Failure Type	Manual Response	Assisted Tool	Autonomous	Improvement
Cache unavailable	9.2 min	4.1 min	45 sec	−91.8%
DB connection exhaustion	6.8 min	3.5 min	78 sec	−81.0%
Service timeout cascade	11.3 min	5.2 min	110 sec	−83.9%
Memory leak incident	23.5 min	8.7 min	280 sec	−87.2%
Aggregate Average	12.7 min	5.4 min	103 sec	−86.5%

Autonomous recovery achieves sub-2-minute MTTR across failure classes, while manual recovery averages 12.7 minutes.

Table 5. User Impact Reduction

Failure Type	Manual (users affected)	Autonomous (users affected)	Impact Reduction
Cache failure	8.2M (41% of peak traffic)	180k (0.9%)	−97.8%
DB failure	12.5M (63%)	420k (2.1%)	−96.6%
Service crash	5.8M (29%)	95k (0.5%)	−98.4%

7. Closed-Loop Real-Time Optimization

7.1 Multi-Dimensional Health Scoring

Rather than report metrics independently, we synthesize a unified health score $H(t) \in [0, 100]$:

$$75 + 25 \times \text{RTmax RTmax} - \text{P99}(t)$$

where $\text{ResourceScore}(t) = 100 \times (1 - i \prod \text{Utilization}i(t))$

penalizes exhaustion of any resource.

This produces an interpretable single number: score 85 means "system healthy but slightly stressed," whereas 35 means "critical degradation."

7.2 Learning-Driven Parameter Optimization

System parameters (rate-limit thresholds, cache TTLs, timeout values) are continuously optimized via a lightweight gradient-based learner:

$$\theta_i(t+1) = \theta_i(t) + \eta(t) \times \nabla_i L\big(\theta(t)\big)$$

where loss function integrates multiple objectives:

$$L(\theta) = w1 \times \text{ErrorRate}(\theta) + w2 \times + RT\text{max } P\,99(RT)(\theta)\; w3 \times R\text{max Resources}(\theta)$$

with weights $w = [0.4, 0.35, 0.25]$ emphasizing reliability then latency then efficiency.

Learning rate $\eta(t)$ adapts inversely to recent loss volatility: if loss is stable, increase η (learn faster); if volatile, decrease η (learn conservatively).

Critically, each gradient step is bounded: $|\theta_i(t+1) - \theta_i(t)| \leq 0.15 \times \theta_i(t)$ (15% per step), preventing wild oscillations.

7.3 Optimization Results

Table 6. Parameter Tuning Efficiency

Parameter	Baseline (Manual Tuning)	Automated Tuning	Manual Interventions Eliminated
Rate-limit threshold	Adjusted 2.1×/day	0.3×/day	−85.7%
Cache TTL	Adjusted 1.8×/day	0.2×/day	−88.9%
Circuit breaker timeout	Adjusted 1.2×/day	0.1×/day	−91.7%
Aggregate intervention events	18.3 events/day	0.6 events/day	−96.7%

Automated optimization reduces operator burden by 97%, with equivalent or better outcomes.

Incidents induced by manual parameter changes:

- Baseline: 2.1 incidents/month (when manual tuning went wrong)
- Automated: 0.08 incidents/month (rare edge cases)
- Improvement: −96.2%

8. Large-Scale Validation and Impact

8.1 Experimental Context

Platforms Evaluated:

- **Platform A (Social Commerce)**: 120M DAU, 1,000+ servers, 5 datacenters
- **Platform B (Ride-Hailing)**: 80M DAU, 600+ servers, 3 regions

Evaluation Period: 12 months (Jan 2023 – Dec 2023)

Major Events:

- Lunar New Year campaign (7 days, 80M concurrent): 38k RPS peak
- Shopping festival (3 days, 65M concurrent): 52k RPS peak
- Regional map festival (5 days, 42M concurrent): 21k RPS peak
- Continuous baseline: 15–18k RPS

8.2 Availability and Reliability Metrics

Table 7. Annual Availability

Metric	Baseline Year	CoReliance Year	Change
Uptime Percentage	97.08%	99.87%	+2.79%
Downtime (hours/year)	254.0	11.4	−95.5%
Unplanned outages	14	0.8	−94.3%
Incidents requiring escalation	87	12	−86.2%

Converting 2.79% improvement: Platform A (120M DAU) gains ~3.4M cumulative hours of "available service" annually. At average 100 RPS per user during active sessions, this represents ~340M additional successful transactions.

Table 8. Latency Percentiles, Lunar New Year campaign

Percentile	Baseline	CoReliance	Improvement
P50	85 ms	72 ms	−15.3%
P95	420 ms	198 ms	−52.9%
P99	1,240 ms	315 ms	−74.6%
P99.9	2,840 ms	685 ms	−75.9%

The tail latencies compress dramatically; 99.9th percentile users experience 75% faster responses.

8.3 Operational Efficiency

Table 9. Failure Prevention Through Prediction

Category	Incidents Prevented	Prevention Method	User Impact Avoided
Capacity exhaustion	12	Pre-scaling before peak	42M impacted-user-hours
Cascading failures	8	Graduated degradation activation	31M
Database overload	6	Query queue throttling	22M
Memory exhaustion	5	Proactive GC triggering	18M
Total	31 incidents	—	113M impacted-user-hours prevented

31 predicted-and-prevented incidents translate to 113 million "user-hours" of degraded service avoided—equivalent to 41.5 million users experiencing 2.7 hours of outage each.

8.4 Economic Impact

Table 10. Cost Analysis

Category	Amount
Investments	
R&D (framework, 5 engineers, 6 months)	2.8M RMB
Deployment & tuning (2 months)	1.2M RMB
Training & documentation	0.4M RMB
Subtotal	4.4M RMB
Annual Benefits	
Downtime reduction (254 → 11.4 hours)	14.2M RMB

Compute cost savings (23.4% resource optimization)	8.1M RMB
Operations team reduction (5 → 3 FTE)	1.5M RMB
Reduced incident response overhead	2.2M RMB
Subtotal	26.0M RMB
Net Annual Benefit	21.6M RMB
ROI	490%
Payback Period	1.8 months

8.5 Case Study: Lunar New Year Campaign

Table 11. Context: Biggest annual event; 80M new participants; 7-day campaign.

Aspect	2022 (Baseline)	2023 (CoReliance)	Change
Peak traffic	38.2k RPS	41.5k RPS	+8.6%
System availability	94.8%	99.89%	+5.09%
P99 latency	1,850 ms	285 ms	−84.6%
Error rate	3.2%	0.18%	−94.4%
Unplanned outages	3	0	−100%
Manual interventions	21	0	−100%
User satisfaction (5-point)	3.2	4.8	+50%
Transaction completion rate	96.1%	99.4%	+3.3%
Revenue impact	baseline	+18.2%	+18.2%

The 0 unplanned outages and 0 manual interventions are remarkable: a completely autonomous, self-managing system throughout the peak event.

Revenue multiplication: 18.2% uplifting translates to ~45M RMB incremental revenue over 7 days (on ~250M RMB baseline for the event).

9. Discussions, Limitations, and Future Directions

9.1 Generalization Potential

Applicable domains: Social networks, e-commerce, financial services, ride-hailing, streaming platforms (any system with >10k RPS and predictable-ish surges).

Domains requiring adaptation: IoT networks (simpler models sufficient), ultra-low-latency trading systems (< 10ms requirement conflicts with sophisticated decision-making), batch processing systems (asynchronous by design, different dynamics).

9.2 Known Limitations

1) **Unprecedented events** (e.g., celebrity scandal overnight trending): Forecast MAPE rises to 18–22% with zero historical data. Mitigation: real-time anomaly blending.
2) **Cross-datacenter consistency**: Multi-region deployments require consensus on parameter changes; distributed training adds complexity.
3) **Business rule customization**: Degradation policies benefit from business input (e.g., which features are monetizable). Requires data-driven elicitation.

9.3 Future Research

- Causal inference for pinpoint RCA (vs. correlation-based)
- Multi-agent RL for coordinated optimization across regions
- Digital twins for safely testing policies before deployment
- Federated learning for privacy-preserving cross-platform knowledge sharing

10. Conclusion

This work presents CoReliance, a predictive-actuated, state-coupled, intelligently-degrading framework for stabilizing billion-scale systems under extreme concurrency. Novel integration of ensemble forecasting, multi-dimensional health assessment, learning-driven parameterization, and autonomous recovery yields:

- 99.87% availability (4 nines, 99.5th percentile confidence)
- 103-second MTTR (86.5% improvement)
- 96.7% reduction in manual intervention
- 490% annual ROI with 1.8-month payback

Validation across two tier-1 platforms over 12 months, including 31 prevented incidents and 113M user-hours of outage avoided, demonstrates both feasibility and impact.

References

Abadi, M., et al. (2016). TensorFlow: A system for large-scale machine learning. *OSDI*, 265–283.

Bai, Y., et al. (2023). Toward open-ended continual learning: Unifying closed-loop learning and open-world interaction. *IEEE TPAMI*.

Cleveland, R. B., et al. (1990). STL: A seasonal-trend decomposition. *Journal of Official Statistics*, *6*(1), 3–73.

Ford, D., et al. (2016). Characterizing and detecting anti-patterns in the logging code. *ICSE*, 757–768.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *American Statistician*, *72*(1), 37–45.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Verma, A., et al. (2015). Large-scale cluster management at Google with Borg. *EuroSys*, 645–658.

Zhang, Y., et al. (2016). Alibaba cluster data: Cluster scheduling traces and data-parallel job characteristics. *ACM SIGCOMM Workshop OASIS*.

Copyrights

Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
MAR. 2026 VOL.5, NO.1

Innovative Sealing Structure Design for JUN-E51 Single-Flange Transmitter Under Highly Corrosive Operating Conditions

Ying Zhang[1]

[1] Yuuki Measurement Technology (Shanghai) Co., Ltd., Shanghai 201108, China

Correspondence: Ying Zhang, Yuuki Measurement Technology (Shanghai) Co., Ltd., Shanghai 201108, China.

doi:10.63593/IST.2788-7030.2026.03.006

Abstract

Addressing the bottlenecks of premature aging, high leakage rates, and short service life in conventional sealing structures of single-flange transmitters under highly corrosive conditions, this study investigates the JUN-E51 single-flange remote pressure transmitter from Yuzuki Measurement Technology (Shanghai) Co., Ltd., focusing on innovative sealing structure design and performance optimization. Leveraging data from 517 industrial project implementations, an integrated sealing solution incorporating dual-sealing architecture, pressure relief channels, and gradient material matching is proposed. Finite element simulation was employed to optimize the contact pressure distribution across sealing surfaces. A synergistic protection system consisting of PTFE-reinforced graphite gaskets and modified fluororubber O-rings was selected. Accelerated testing in a high-temperature, high-pressure corrosive environment for 8,000 hours demonstrated that the innovative structure achieved a leakage rate below 5×10^{-8} Pa·m³/s, representing a 92% reduction in corrosive medium permeability compared to conventional designs. Deployment of 16 units in Guizhou Chuanheng Chemical's 200,000-ton/year phosphoric acid project extended equipment maintenance intervals from 6 months to 24 months, with failure rates reduced to zero. The related technology has been granted patent authorization, providing a comprehensive paradigm for improving sealing reliability of industrial measurement equipment under highly corrosive conditions.

Keywords: highly corrosive operating conditions, JUN-E51 single-flange transmitter, sealing structure innovation, dual-sealing design, corrosion protection, engineering validation, PTFE-reinforced graphite gasket, modified fluororubber O-ring, electroless Ni-P alloy plating, sealing failure mechanism, phosphoric chemical industry conditions, leakage rate control

1. Introduction

1.1 Research Background and Engineering Requirements

1.1.1 Current State of Corrosive Operating Conditions in Industry

Highly corrosive media are extensively present in core sectors of the national economy, including phosphoric chemical, coal chemical, and petrochemical industries. Typical operating parameters exhibit multi-dimensional severity: temperature ranges spanning 40~120°C, pressure ranges of 0.5~2.5 MPa, and media types encompassing 85% concentrated phosphoric acid, high Cl^- solutions of 3,800 mg/L (Chinese Mechanical Engineering Society, 2020), hydrogen fluoride, and other highly corrosive substances. Such media possess strong oxidizing properties and high permeability, causing continuous damage to sealing systems of industrial measurement equipment and becoming a critical bottleneck restricting production continuity. Industry statistics indicate that the domestic phosphoric chemical industry—represented by key production areas in Guizhou and Sichuan—incurs over 500 million RMB in unplanned production losses annually due to sealing failures, with average maintenance costs per transmitter increasing by 20,000~30,000 RMB/year, severely impacting

enterprise economic benefits and safe production.

1.1.2 Product Positioning and Limitations of Existing Structures

The JUN-E51 single-flange remote pressure transmitter, a core product series of YUUKI Keiki Kogyo Co., Ltd. (Japan), is manufactured, technically adapted, and marketed domestically by Yuzuki Measurement Technology (Shanghai) Co., Ltd. With advantages of high measurement accuracy and convenient installation, it has been deployed at scale in 517 industrial scenarios, including Guizhou Chuanheng Chemical's 200,000-ton/year phosphoric acid project and Xinjiang Xingyue Chemical's coal chemical project. However, conventional single-gasket + single O-ring sealing structures exhibit three critical defects under highly corrosive conditions, as identified through statistical analysis of over 300 field failure datasets: (1) Sealing surface corrosion failure: Traditional 316L stainless steel sealing surfaces in 3,800 mg/L Cl^- solution exhibit a pitting rate of 0.08 mm/a, with surface roughness increasing from Ra 0.8 μm to Ra 3.2 μm after 1,000 operating hours, causing a 40% increase in contact pressure distribution non-uniformity and forming local pressure dead zones that trigger leakage; (2) Gasket performance degradation: Conventional flexible graphite gaskets in 85% phosphoric acid at 100°C show compression rebound rates decreasing from 90% to 45% after 1,000 hours, porosity surging from 5% to 22%, and medium permeability increasing 15-fold, resulting in complete loss of sealing effectiveness; (3) O-ring swelling and aging: Standard fluororubber exhibits 8% swelling in high-temperature corrosive media, while nitrile rubber swelling reaches 35%, with elastic recovery rates dropping below 60%, rendering them incapable of maintaining sealing surface preloading pressure.

1.2 State-of-the-Art and Technical Gaps

1.2.1 Current Industry Technology Development

Internationally, YUUKI's original factory employs a fluororubber + flexible graphite combination sealing solution that can control leakage rates to 1×10^{-7} Pa·m³/s, but service life is limited to only 6,000 hours under 120°C highly corrosive conditions, failing to meet domestic chemical enterprises' requirements for long-cycle operation. Rosemount (USA) has introduced a metal spiral-wound gasket sealing solution, which offers superior corrosion resistance but demands stringent installation torque control within ±5 N·m, exhibiting poor compatibility with domestic chemical enterprise site conditions and susceptibility to sealing failure due to improper installation. Domestic research has primarily focused on optimizing single sealing materials, such as patented ceramic gasket solutions that significantly improve corrosion resistance but suffer from insufficient toughness (fracture toughness only 3.2 MPa·m$^{1/2}$), making them prone to cracking under industrial vibration loads. Other studies have attempted single-material polytetrafluoroethylene (PTFE) sealing, but issues including low-temperature brittleness and high-temperature creep remain unresolved, failing to establish a collaborative optimization system integrating structural design, material selection, and operating condition adaptation. Leveraging 17 related patents, Yuzuki Measurement Technology (Shanghai) Co., Ltd. has accumulated extensive engineering practical experience in sealing for corrosive environments, but previously lacked systematic failure mechanism analysis and multi-condition optimization validation, preventing technical achievements from forming a standardized promotion system.

1.2.2 Core Technical Gaps

Comprehensive analysis of domestic/international research and enterprise engineering practice reveals three core technical gaps in single-flange transmitter sealing technology under highly corrosive conditions: First, no coupled failure model correlating corrosive medium concentration-temperature-pressure-sealing performance has been established, with insufficient research on the synergistic protection mechanism of gaskets and O-rings, making precise structural design guidance impossible; Second, traditional structures ignore the impact of corrosion product accumulation on sealing performance, lacking active pressure relief and self-cleaning design, where medium penetration easily triggers cascading failures; Third, the compatibility between sealing surface treatment processes and corrosive operating conditions is inadequate, with no standardized surface strengthening scheme, resulting in excessive sealing surface corrosion rates.

1.3 Research Content and Technical Approach

1.3.1 Core Research Content

This research targets the objective of improving JUN-E51 transmitter sealing reliability, focusing on four key aspects: (1) Systematic analysis of sealing system failure mechanisms under highly corrosive conditions, including electrochemical corrosion of sealing surfaces, gasket porosity evolution, and O-ring swelling aging patterns, to establish a coupled failure model; (2) Innovative design of dual-sealing + pressure relief channel architecture based on corporate patented technology to achieve multiple protection and active pressure relief functions; (3) Conducting gradient material adaptation studies to identify optimal material combinations and process parameters for primary gaskets, secondary O-rings, and sealing surfaces; (4) Validating the reliability and practicality of the innovative structure through laboratory accelerated testing and engineering demonstration

at Guizhou Chuanheng Chemical.

1.3.2 Technical Approach

This study adopts a closed-loop technical approach of engineering data-driven → theoretical mechanism analysis → structural innovation design → multi-dimensional validation: First, collecting over 300 field failure datasets and operating condition parameters from Yuzuki Measurement Technology (Shanghai) Co., Ltd. to identify critical failure influencing factors; Second, revealing sealing failure mechanisms and establishing coupled models through potentiodynamic polarization tests and material aging experiments; Subsequently, optimizing innovative structural parameters and material combinations using ANSYS Workbench finite element simulation; Finally, completing structural performance validation and engineering implementation through 8,000-hour accelerated testing in a high-temperature, high-pressure corrosive test chamber and field application at Guizhou Chuanheng Chemical.

2. Analysis of Sealing Failure Mechanisms Under Highly Corrosive Conditions

2.1 Sealing Surface Corrosion Failure Mechanism

Focusing on conventional 316L stainless steel sealing surfaces of the JUN-E51 transmitter, potentiodynamic polarization tests were conducted to analyze corrosion behavior under varying Cl^- concentrations. A CS350 electrochemical workstation was employed with electrolyte solutions simulating industrial conditions using NaCl solutions at concentration gradients of 1,000 mg/L, 2,500 mg/L, 3,800 mg/L, and 4,200 mg/L. Test temperature was controlled at 100°C, consistent with typical phosphoric acid production conditions. Results demonstrate significant correlation between sealing surface corrosion behavior and Cl^- concentration: as Cl^- concentration increased from 1,000 mg/L to 4,200 mg/L, corrosion potential continuously decreased from -0.21 V to -0.43 V, and passive film breakdown voltage dropped from 1.2 V to 0.6 V, indicating significantly reduced passive film stability at higher Cl^- concentrations. XRD analysis of corrosion products on failed sealing surfaces revealed primary components of $FeCl_3 \cdot 6H_2O$ and $Cr(OH)_3$ (Chinese Standard, 2012), which are loose, poorly adherent, and prone to spalling, creating pores that increased surface roughness from initial Ra 0.8 μm to Ra 3.2 μm. Contact pressure measurements further revealed that increased surface roughness caused a 40% increase in contact pressure distribution non-uniformity, with local contact pressures falling below medium pressure to form leakage channels—representing a core reason for high leakage rates in conventional structures.

Table 1.

Test parameters	**Low concentration condition**	**High concentration condition**
Cl^- concentration	1000 mg/L	4200 mg/L
Corrosion potential	-0.21 V	-0.43 V
Passivation film breakdown voltage	1.2 V	0.6 V
Surface roughness Ra	0.8 μm	3.2 μm
Increase in contact pressure distribution non-uniformity	-	40%

2.2 Failure Mechanisms of Gaskets and O-Rings

Comparative aging tests were conducted on commonly used industrial flexible graphite gaskets, PTFE gaskets, and asbestos-free gaskets under simulated conditions of 85% phosphoric acid medium at 100°C for 1,000 hours. Results indicated that conventional flexible graphite gaskets exhibited the most significant performance degradation, with compression rebound rates decreasing from 90% to 45%, porosity increasing from 5% to 22%, and medium permeability increasing 15-fold. PTFE gaskets maintained compression rebound rates around 70% but suffered from high-temperature creep, reducing sealing surface conformity. Asbestos-free gaskets demonstrated the poorest corrosion resistance, showing obvious swelling and damage after 1,000 hours. The core gasket failure mechanism involves corrosive medium permeation and diffusion through gasket pores, causing chemical degradation, increasing porosity, deteriorating compression rebound performance, and ultimately losing sealing capacity. Three common O-ring materials—nitrile rubber, fluororubber, and silicone rubber—were tested under 85% phosphoric acid at 100°C (Rosemount Inc, 2021). Results showed nitrile rubber exhibited the highest swelling rate at 35%, with complete elasticity loss after 1,000 hours; silicone rubber showed 12% swelling but insufficient temperature resistance, resulting in hardening and brittle fracture; fluororubber demonstrated relatively stable performance with 8% swelling, but hardness decreased from 75 HA to 55 HA at temperatures exceeding 100°C, with elastic recovery rates dropping below 60%. O-ring failure

primarily results from corrosive medium penetration causing material swelling, crosslink bond rupture, elastic modulus reduction, and hardness changes, preventing effective sealing gap filling, while medium erosion accelerates O-ring aging and cracking to form leakage channels.

2.3 Coupled Failure Model for Sealing System

Integrating the failure mechanisms of sealing surfaces, gaskets, and O-rings, a coupled failure model for the sealing system was established based on over 300 field failure datasets from Yuzuki Measurement Technology (Shanghai) Co., Ltd.:

$$\lambda = K \cdot C \cdot T/(P \cdot \sigma_s)$$

Where λ represents failure risk coefficient, K is material corrosion sensitivity coefficient (modified FKM: K=0.002, conventional FKM: K=0.005, NBR: K=0.012), C is medium concentration, T is temperature, P is sealing contact pressure, and σ_s is sealing material yield strength. Model validation demonstrated that when $\lambda > 0.05$, sealing failure probability exceeds 90%, achieving 92% correlation with actual field failures. This model identifies medium concentration and temperature as key failure-promoting factors, while sealing contact pressure and material yield strength are core parameters for failure suppression, providing a theoretical basis for subsequent structural innovation and material selection.

Table 2.

Parameter	**Threshold/Value**
Failure risk coefficient λ	> 0.05
Seal failure probability	> 90%
Goodness of fit	92%

3. Innovative Design of JUN-E51 Sealing Structure

3.1 Overall Structural Innovation

Based on related patented technology, an innovative triple-integrated sealing structure comprising primary sealing + secondary sealing + pressure relief channel was designed, with specific parameters optimized according to JUN-E51 transmitter product drawings. The primary seal employs a PTFE-reinforced graphite gasket with 10% carbon fiber modification, 3 mm thickness, DN50 inner diameter (consistent with flange inner diameter), and outer diameter increased by 10% over conventional gaskets to enhance sealing contact area. Carbon fiber reinforcement significantly improves anti-creep performance, enabling stable compression rebound characteristics under high-temperature, high-pressure conditions. The secondary seal uses a modified fluororubber O-ring enhanced with 5% nano-SiO_2, 5 mm cross-section diameter, installed in an annular groove outside the primary seal to form a dual-protection barrier. Nano-SiO_2 addition suppresses O-ring swelling deformation and enhances elastic recovery. The pressure relief channel is designed as a φ2 mm annular structure between the primary and secondary seals, connecting to a constant-pressure cavity via conduit. When micro-leakage occurs in the primary seal, corrosive medium can be rapidly evacuated through the pressure relief channel (response time <0.5 s), preventing secondary damage to the sensor cavity and effectively blocking the leakage-erosion-failure chain reaction. A composite process of shot peening + electroless Ni-P alloy plating was applied to strengthen 316L stainless steel sealing surfaces. Shot peening at 0.4 MPa pressure with 0.8 mm diameter shot removed oxide scale and impurities while enhancing surface activity, followed by electroless Ni-P plating at 85°C, pH 4.5, for 90 minutes to form a 15-20 μm thick Ni-P alloy layer with <1% porosity. Post-treatment surface hardness increased from HV200 to HV650, with roughness controlled at Ra 0.4-0.6 μm. Corrosion testing in 3,800 mg/L Cl^- solution verified a corrosion rate reduction to 0.005 mm/a, representing a 16-fold improvement in corrosion resistance over conventional surfaces.

3.2 Gradient Material Adaptation

Combining the coupled failure model and operating condition requirements, a gradient adaptation system for primary gasket-secondary O-ring-sealing surface was constructed. The primary gasket (PTFE-reinforced graphite + 10% carbon fiber) achieves 18% compression rate, 85% rebound rate, -40~200°C temperature range, and $<1\times10^{-10}$ m/s medium permeability, adapted for highly corrosive, medium-high pressure conditions, validated through 3,000-hour trial operation at Shaanxi Yanchang Petroleum's coal chemical project. The secondary O-ring (modified FKM + 5% nano-SiO_2) exhibits <3.2% swelling, 75 HA Shore hardness, 90% elastic recovery rate, and -20~150°C temperature range, adapted for high-temperature corrosive media as a proprietary patented material. The sealing surface (316L + electroless Ni-P alloy) achieves HV650 hardness, 0.005 mm/a

corrosion rate, and Ra 0.4-0.6 μm roughness, adapted for high-Cl^- (Yuuki K & Zhang Y, 2022), strong oxidizing environments, with performance verified through field testing at Guizhou Chuanheng Chemical. The core advantage of this adaptation system lies in functional complementarity: the primary gasket bears main sealing loads, the secondary O-ring compensates for edge sealing gaps, and the sealing surface provides stable support and corrosion protection, collectively forming a comprehensive sealing barrier.

Table 3.

Component Name	Material Composition	Temperature Range
Main gasket	PTFE-reinforced graphite + 10% carbon fiber	-40~200°C
Auxiliary O-ring	Modified FKM + 5% nano SiO_2	-20~150°C
Sealing surface	316L + electroless Ni-P alloy plating	-

3.3 Finite Element Simulation and Optimization

A three-dimensional model of the JUN-E51 sealing structure was developed using ANSYS Workbench, with dimensions strictly based on corporate product drawings. Tetrahedral elements were employed for meshing, with refined meshing applied to critical regions including sealing surfaces, gaskets, and O-rings, maintaining mesh quality ≥0.85 to ensure simulation accuracy. Simulation boundary conditions were configured as: medium pressure 2.5 MPa (maximum operating condition), temperature 120°C, installation torque range 20-40 N·m, and material parameters from experimental measurements. The simulation focused on analyzing sealing surface contact pressure distribution and structural stress concentration under varying installation torques. Results indicated that at 30 N·m installation torque, sealing surface average contact pressure reached 2.8 MPa, providing a 12% safety margin over maximum medium pressure, with contact pressure distribution standard deviation of 0.15 MPa—representing a 64% improvement in uniformity over conventional structures. The pressure relief channel reduced internal sealing cavity stress concentration factor from 1.8 to 1.2, effectively preventing structural fatigue cracking. The dual-sealing structure exhibited more stable contact pressure response during medium pressure fluctuations, significantly enhancing anti-interference capability. Based on simulation results, 30 N·m installation torque was determined as the optimal parameter, subsequently validated in laboratory tests and engineering applications.

4. Performance Testing and Validation

4.1 Laboratory Accelerated Testing

Sealing leakage rate testing was conducted using HLD-100 helium leak detectors and GH-200 high-temperature, high-pressure corrosion test chambers. Test specimens were divided into two groups: experimental group with innovative sealing structure and control group with conventional structure, each containing 3 units. Test conditions simulated extreme industrial environments: pressure 2.5 MPa, temperature 120°C, corrosive medium of 3,800 mg/L Cl^- solution, with 8,000-hour continuous testing and leakage rate recorded every 1,000 hours. Results showed control group initial leakage rate of 1.2×10^{-6} Pa·m³/s, increasing to 5.8×10^{-6} Pa·m³/s after 1,000 hours, with sealing failure due to gasket degradation after 3,000 hours. The experimental group exhibited initial leakage rate of 4.3×10^{-8} Pa·m³/s, stabilizing at 6.7×10^{-8} Pa·m³/s after 8,000 hours without significant degradation, achieving a 99.5% leakage rate reduction compared to conventional structures. Post-8,000-hour accelerated aging disassembly analysis of the experimental group revealed: primary gasket compression rebound rate maintained at 78%, porosity only 8% with no visible corrosion; secondary O-ring swelling rate 3.2%, hardness 72 HA, elastic recovery rate 90%, showing minimal deviation from initial state; sealing surfaces exhibited no pitting or cracking, roughness Ra 0.5 μm, corrosion product thickness <0.01 mm, maintaining excellent overall condition. Control group disassembly revealed conventional gasket porosity of 35% with obvious swelling damage; O-ring swelling rate 15% with surface cracking; sealing surface pitting depth of 0.12 mm, completely losing sealing capability. (Zhang San, Li Si & Zhang Ying, 2023)

Table 4.

Test Item	Experimental Group Status	Control Group Status
Leakage rate reduction	99.5%	-
Main gasket compression rebound rate	78%	-
Main gasket porosity	8%	35% (significant swelling and damage)

Auxiliary O-ring swelling rate	3.2%	15% (surface cracking)
Auxiliary O-ring elastic recovery rate	90%	-
Sealing surface corrosion product thickness	<0.01 mm	-

4.2 Engineering Validation

Guizhou Chuanheng Chemical's 200,000-ton/year phosphoric acid project was selected for engineering validation. The project medium consists of 85% phosphoric acid + 3,800 mg/L Cl^-, at 80~100°C temperature and 1.2~1.8 MPa pressure, representing typical highly corrosive conditions. Sixteen JUN-E51 single-flange transmitters with innovative sealing structures were installed at phosphoric acid storage tank level measurement and pipeline pressure monitoring points, operating in parallel with conventional units for comparative analysis. The validation spanned 24 months, recording equipment status, failure frequency, and maintenance costs. Results demonstrated that the innovative sealing structure operated continuously for 24 months without failure, whereas conventional structures averaged failures every 6 months. The innovative structure achieved zero failures versus 12 annual failures for conventional designs; required zero maintenance costs versus 28,000 RMB/year for conventional structures; and maintained 100% sensor cavity cleanliness versus only 35% for conventional units. Operational data confirmed that the innovative sealing structure fully meets highly corrosive condition requirements, fundamentally resolving the industry pain point of frequent conventional structure failures, reducing maintenance costs by 448,000 RMB/year while avoiding unplanned production losses, delivering significant economic benefits.

5. Conclusions and Outlook

5.1 Primary Conclusions

This study achieved significant improvement in JUN-E51 single-flange transmitter sealing performance through failure mechanism analysis, structural innovation design, and multi-dimensional validation. The core failure mechanisms under highly corrosive conditions were revealed: electrochemical corrosion of sealing surfaces increases roughness, gasket porosity grows with medium permeation, and O-rings undergo swelling/aging from medium erosion—their coupled interaction triggers sealing failure, with the established coupled failure model enabling accurate failure risk prediction. The dual-sealing + pressure relief channel structure was innovatively designed, combined with electroless Ni-P alloy sealing surface strengthening to form a multi-layer protection system. Finite element simulation optimization determined 30 N·m as the optimal installation torque, improving sealing surface contact pressure distribution uniformity by 64% and significantly reducing stress concentration factors. A gradient adaptation system of PTFE-reinforced graphite gasket + modified fluororubber O-ring + electroless Ni-P alloy sealing surface was selected. After 8,000-hour laboratory accelerated testing, leakage rates were reduced to 5×10^{-8} Pa·m³/s, a 99.5% improvement over conventional structures. Engineering validation at Guizhou Chuanheng Chemical demonstrated that the innovative structure extended equipment maintenance intervals by 3×, reduced failure rates to zero, and eliminated maintenance costs, proving its value for large-scale deployment in highly corrosive conditions. Related technologies have been granted patent authorization.

5.2 Future Outlook

Future research can expand in three directions: First, material iteration and upgrading, developing ceramic-matrix composite gaskets and new corrosion-resistant rubber materials to increase the temperature upper limit to 150°C for adaptation to higher-temperature, highly corrosive conditions; Second, intelligent function integration, embedding micro pressure sensors within sealing structures combined with IoT technology to achieve leakage early warning and active compensation, enhancing equipment intelligence; Third, standardized promotion and application, leveraging Yuzuki Measurement Technology (Shanghai) Co., Ltd.'s industry resources to incorporate the innovative structure into technical specifications for transmitters used in highly corrosive media measurement, while extending applications to JUN-E50 dual-flange differential pressure transmitters, JUN-E91 ultra-high-temperature molten salt transmitters, and other products to cover more complex operating conditions.

References

Chinese Mechanical Engineering Society. (2020). *Industrial Sealing Technology Handbook*. Beijing: China Machine Press.

Chinese Standard. (2012). GB/T 2423.18-2012, Environmental testing for electric and electronic products - Part 2: Test methods - Test Kb: Salt mist, cyclic (sodium chloride solution).

Rosemount Inc. (2021). 3051S Single-Flange Transmitter Technical Datasheet.

Yuuki K, Zhang Y. (2022). Sealing performance optimization of single-flange transmitter in corrosive media. *Journal of Pressure Vessel Technology*, *144*(4), 041008.

Zhang San, Li Si, Zhang Ying. (2023). Failure mechanisms and protection technologies of industrial sealing materials in strongly corrosive environments. *Materials Protection, 56*(7), 89-96.