Informatics for Educators And Professionals : Journal of Informatics
Vol. 10, No. 1, Juni 2025, 13 - 24
E-ISSN: 2548-3412
Terakreditasi **SINTA Peringkat 5**
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Nomor: 230/E/KPT/2022
masa berlaku mulai Volume 5 Nomor 1 Tahun 2020 sampai Volume 9 Nomor 2 Tahun 2024

Prediksi Kualitas Produk Manufaktur Semikonduktor Menggunakan Machine Learning

Darusman[1],* Aries Abbas[2] , Angga Dwi Firmanto[3]

[1] Program Studi Magister Ilmu Komputer; Universitas Nusa Mandiri Jakarta; Jl. Raya Jatiwaringin Cipinang Melayu, Kec. Makasar, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, No.telpon (021)28534471; e-mail: 14230029@nusamandiri.ac.id
[2]Program Studi Teknik Mesin; Fakultas Teknik; Universitas Krisnadwipayana; Jl. Kampus UNKRIS Jatiwaringin, Pondok Gede, Kota Bekasi, No.Telpon (021) 8462230/846223; e-mail: ariesabbas@unkris.ac.id
[3]Program Studi Teknik Perawatan Mesin; Politeknik Negeri Media Kretif; Jl. Srengseng Sawah Raya, Jagakarsa, Kota Jakarta Selatan, DKI Jakarta No.Telpon (6281 1166 9695) angga.firmanto@polimedia.ac.id

* Korespondensi: e-mail: 14230029@nusamandiri.ac.id

Diterima: 14 Februari 2025; Review: 7 Maret 2025; Disetujui: 24 April 2025

Cara sitasi: Darusman, Abbas A, Firmanto AD. 2025. Prediksi Kualitas Produk Manufaktur Semikonduktor Menggunakan Machine Learning. Informatics for Educators and Professionals : Journal of Informatics. Vol.10 (1): 13-24.

Abstrak:Industri manufaktur semikonduktor menghadapi tantangan dalam pengendalian kualitas produk yang efisien dan akurat, di mana metode inspeksi manual memiliki keterbatasan dalam hal kecepatan dan akurasi. Penelitian ini bertujuan untuk mengatasi tantangan tersebut dengan mengembangkan model prediksi kualitas produk menggunakan berbagai algoritma machine learning, yaitu Logistic Regression, Decision Tree, Random Forest, XGBoost, Naïve Bayes, dan SVM. Berbagai teknik pembagian dataset seperti Normal Data (90:10), Oversampling (70:30), Undersampling (80:20 & 70:30), dan PCA (90:10) digunakan untuk mengoptimalkan performa model. Hasil penelitian menunjukkan bahwa XGBoost dan Random Forest mencapai akurasi tertinggi (0,95), sementara Naïve Bayes memiliki performa terendah (0,23). Teknik oversampling meningkatkan akurasi Decision Tree dan Logistic Regression, namun dapat menyebabkan overfitting pada XGBoost. Penerapan PCA meningkatkan efisiensi Logistic Regression dengan akurasi 0,81, membuktikan bahwa reduksi dimensi dapat meningkatkan kinerja model. Dengan demikian, penelitian ini menunjukkan bahwa teknik machine learning dan reduksi dimensi memiliki potensi besar dalam meningkatkan prediksi kualitas produk semikonduktor, dengan XGBoost dan Random Forest sebagai model terbaik.

Kata kunci: Machine Learning; Manufaktur Semikonduktor; Prediksi Kualitas Produk

Abstract: *The semiconductor manufacturing industry faces challenges in efficient and accurate product quality control, where manual inspection methods have limitations in terms of speed and accuracy. This research aims to address these challenges by developing a product quality prediction model using various machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, XGBoost, Naïve Bayes, and SVM. Various dataset partitioning techniques such as Normal Data (90:10), Oversampling (70:30), Undersampling (80:20 & 70:30), and PCA (90:10) are used to optimize model performance. The results show that XGBoost and Random Forest achieve the highest accuracy (0.95), while Naïve Bayes has the lowest performance (0.23). The oversampling technique improves the accuracy of Decision Tree and Logistic Regression, but may cause overfitting in XGBoost. The application of PCA increases the efficiency of Logistic Regression with an accuracy of 0.81, proving that dimensionality reduction can enhance model performance. Thus, this study demonstrates that*

machine learning techniques and dimensionality reduction have great potential in improving semiconductor product quality prediction, with XGBoost and Random Forest being the best models.

***Keywords**: Machine Learning; Semiconductor Manufacturing; Product Quality Prediction*

1. Pendahuluan

Dalam industri manufaktur semikonduktor, pengendalian kualitas produk merupakan aspek yang sangat penting untuk memastikan efisiensi produksi serta mengurangi jumlah produk cacat. Kualitas produk yang tidak terjamin dapat menyebabkan peningkatan biaya produksi akibat perbaikan atau pembuangan produk yang tidak memenuhi standar [1][2]. Secara tradisional, industri manufaktur masih banyak mengandalkan metode inspeksi manual dan berbasis aturan dalam menentukan kelayakan produk. Metode ini memiliki keterbatasan dalam hal kecepatan, biaya operasional, serta risiko kesalahan manusia. Selain itu, dengan semakin meningkatnya kompleksitas produk semikonduktor, jumlah data sensor yang dihasilkan dalam proses manufaktur menjadi semakin besar [3][4]. Hal ini membuat metode konvensional semakin sulit untuk diimplementasikan secara efisien dan efektif. Seiring dengan kemajuan teknologi, penerapan machine learning dalam manufaktur telah berkembang sebagai solusi yang menjanjikan untuk meningkatkan akurasi prediksi kualitas produk. Machine learning memungkinkan sistem untuk menganalisis pola dalam data sensor secara otomatis dan memprediksi apakah suatu produk akan lolos uji kualitas atau tidak [5][6][7]. Dengan menggunakan algoritma berbasis *supervised learning,* model dapat mempelajari hubungan antara berbagai variabel proses manufaktur dengan hasil akhir produk. Dengan cara ini, perusahaan dapat mengoptimalkan proses produksi dan mengurangi jumlah produk cacat secara signifikan. Kemudian salah satu tantangan utama dalam penerapan machine learning untuk prediksi kualitas produk adalah ketidakseimbangan data, di mana jumlah produk dengan status *Pass* jauh lebih banyak dibandingkan dengan produk yang masuk kategori *Fail.* Ketidakseimbangan ini dapat menyebabkan model lebih cenderung untuk memprediksi kelas mayoritas, yang pada akhirnya menurunkan akurasi dalam mendeteksi produk cacat.

Selain ketidakseimbangan data, tantangan lain yang dihadapi dalam penelitian ini adalah tingginya jumlah fitur dalam dataset yang digunakan. Dataset yang diperoleh dari *Factory Manufacturing Semiconductor Test* (FMST) terdiri dari 1567 sampel dengan 591 fitur, yang mencerminkan parameter pengukuran dari berbagai sensor dalam proses manufaktur. Banyaknya jumlah fitur ini berpotensi menyebabkan redundansi informasi serta meningkatkan kompleksitas model [8][9][10]. Oleh karena itu, diperlukan metode seleksi fitur dan reduksi dimensi seperti *Principal Component Analysis* (PCA) untuk menghilangkan fitur yang tidak relevan atau memiliki korelasi tinggi dengan fitur lainnya. Dengan cara ini, model dapat beroperasi lebih efisien tanpa kehilangan informasi penting dalam proses prediksi. Beberapa penelitian sebelumnya telah mengeksplorasi penerapan machine learning dalam manufaktur semikonduktor [11][12][13]. Studi terdahulu menunjukkan bahwa model Random Forest dan Support Vector Machine (SVM) dapat digunakan untuk mengklasifikasikan hasil produksi dengan tingkat akurasi yang cukup tinggi [14][15] [16]. Namun, metode ini masih menghadapi kendala dalam menangani ketidakseimbangan kelas, yang dapat menurunkan efektivitas model dalam mengidentifikasi produk cacat. Selain itu, penelitian lain juga menunjukkan bahwa penerapan teknik *oversampling* seperti SMOTE *(Synthetic Minority Over-sampling Technique)* dapat meningkatkan akurasi model dalam mendeteksi kelas minoritas. Namun, metode ini dapat meningkatkan risiko *overfitting,* terutama pada model berbasis pohon keputusan seperti XGBoost dan Random Forest [17]. Oleh karena itu, diperlukan evaluasi lebih lanjut terhadap berbagai teknik *balancing* data untuk memastikan efektivitasnya dalam meningkatkan performa prediksi.

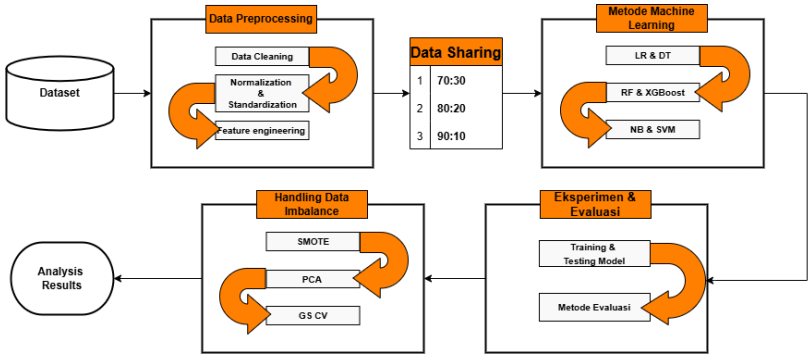
Penelitian ini bertujuan untuk mengembangkan model prediksi hasil produk manufaktur semikonduktor dengan menerapkan berbagai algoritma machine learning, termasuk Logistic Regression, Decision Tree, Random Forest, XGBoost, Naïve Bayes, dan SVM, serta mengevaluasi teknik pembagian dataset seperti Normal Data (90:10), Oversampling (70:30), Undersampling (80:20 & 70:30), dan PCA (90:10) guna menganalisis efektivitas teknik balancing data dan dampaknya terhadap performa model, untuk mengatasi masalah ketidakseimbangan data dan jumlah fitur yang tinggi yang menjadi gap utama dalam prediksi kualitas produk semikonduktor, serta memberikan kontribusi signifikan bagi industri manufaktur dalam meningkatkan efisiensi produksi dan mengurangi jumlah produk cacat, sekaligus menjadi

acuan dalam pengembangan metode prediksi berbasis data yang lebih akurat dan dapat diandalkan dalam proses manufaktur semikonduktor.

2. Metode Penelitian

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimental dengan metode machine learning untuk memprediksi jenis hasil entitas produk pada proyek manufaktur semikonduktor. Model machine learning diuji dengan berbagai skenario pembagian dataset, yaitu 70:30, 80:20, dan 90:10.



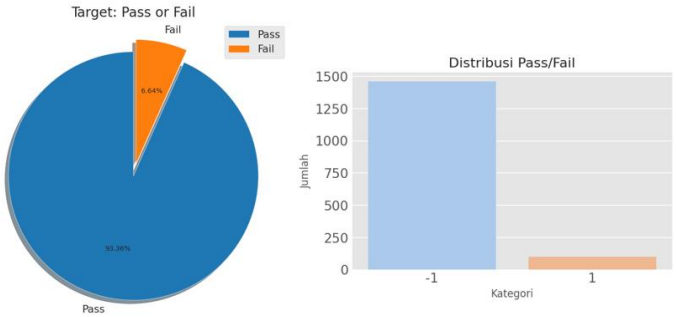
Sumber : Hasil Penelitian (2025)

Gambar 1. Desain Penelitian

2.2. Dataset

a. Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini berasal dari FMST *(Factory Manufacturing Semiconductor Test)* dan mencakup 1567 sampel dengan 591 fitur pada link https://www.kaggle.com/code/saurabhbagchi/fmst-semiconductor-manufacturing-project/input. Setiap data mewakili satu entitas produksi dengan fitur yang diukur, dan label menunjukkan hasil pengujian internal *(Pass/Fail).* Kolom target "–1" menunjukkan *Pass,* sedangkan "1" menunjukkan *Fail.*



Sumber : Hasil Penelitian (2025)

Gambar 2. Distribusi Pass/Fail

Gambar 2 distribusi hasil produksi dalam dataset berdasarkan kategori *Pass* dan *Fail.* Diagram *pie chart* di sebelah kiri menunjukkan bahwa mayoritas produk dalam dataset dikategorikan sebagai Pass (93.36%), sedangkan hanya 6.64% yang masuk kategori Fail. Kemudian *bar chart* di sebelah kanan mengkonfirmasi ketidakseimbangan kelas dalam dataset, di mana jumlah produk dengan label *Pass* (-1) jauh lebih besar dibandingkan produk dengan label Fail (1).

b. Preprocessing Data

Sebelum digunakan dalam pemodelan, dataset mengalami beberapa tahap preprocessing, yaitu: Pembersihan data menghapus data duplikat, menangani nilai yang hilang

(missing values), dan mengonversi data dalam format yang sesuai [18]. Normalisasi dan standarisasi dilakukan untuk memastikan setiap fitur memiliki skala yang seragam [19]. Feature engineering seleksi fitur yang paling relevan menggunakan metode statistik dan teknik seleksi otomatis [20]. Pembagian dataset dibagi menjadi tiga yaitu: *Oversampling dan Undersampled* 70:30%. *Undersampling* 80:20%. Normal Data dan PCA 90:10.

2.3. Metode Machine Learning

Model yang digunakan dalam penelitian ini terdiri dari beberapa algoritma supervised learning, yaitu : Logistic Regression untuk model baseline dalam prediksi [21]. Decision Tree (DT) digunakan untuk membangun model berbasis aturan [22]. Random Forest (RF) meningkatkan akurasi dengan mengombinasikan beberapa pohon keputusan [23]. XGBoost untuk peningkatan performa klasifikasi [24]. Naive Bayes digunakan untuk klasifikasi berbasis probabilistic [25]. Support Vector Machine (SVM) mencari hyperplane terbaik dalam memisahkan kelas data [26].

2.4. Eksperimen dan Evaluasi

a. Training dan Testing Model

Model dilatih menggunakan berbagai skenario pembagian dataset untuk melihat pengaruh jumlah data pelatihan terhadap kinerja model [27].

b. Metode Evaluasi

Evaluasi model dilakukan dengan menggunakan metrik: Akurasi proporsi prediksi yang benar dibandingkan dengan total prediksi [28]. *Precision, Recall, dan F1-Score* digunakan untuk mengevaluasi performa model dalam menangani kelas yang tidak seimbang [29]. ROC-AUC *(Receiver Operating Characteristic - Area Under Curve)* menilai kemampuan model dalam membedakan antara kelas positif dan negatif [30]. Confusion matrix untuk menganalisis kesalahan klasifikasi [31].

2.5. Penanganan Ketidakseimbangan Data

Karena distribusi kelas mungkin tidak seimbang, dilakukan metode sampling berikut: SMOTE *(Synthetic Minority Over-sampling Technique)* untuk menyeimbangkan dataset [32]. Under-sampling untuk mengurangi jumlah data mayoritas. PCA *(Principal Component Analysis)* diterapkan pada data yang telah diundersampling untuk mengurangi dimensi fitur dan menghilangkan redundansi data [33].

2.6. Penyempurnaan Model

Grid Search CV digunakan untuk mendapatkan *hyperparameter* terbaik bagi salah satu model.

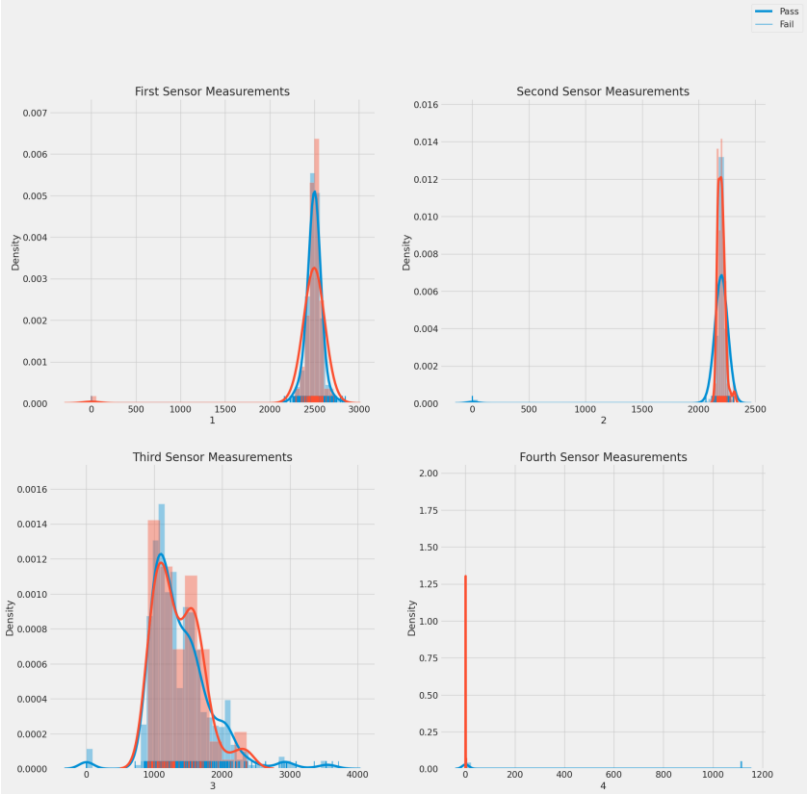
3. Hasil dan Pembahasan

Hasil analisis menunjukkan bahwa model machine learning yang digunakan memiliki kemampuan prediksi yang cukup baik dalam menentukan hasil produk manufaktur semikonduktor. Evaluasi model dilakukan menggunakan beberapa metrik utama, yaitu *akurasi, precision, recall, F1-score,* dan *ROC-AUC,* untuk memastikan keandalan prediksi. Berdasarkan hasil akurasi dari berbagai model, XGBoost, Random Forest, dan SVM menunjukkan performa terbaik dengan akurasi mencapai 0.95 pada dataset normal dengan pembagian 90:10. Hal ini menunjukkan bahwa model ini mampu menangkap pola data dengan sangat baik dan memberikan prediksi yang andal. Sebaliknya, model Naïve Bayes memiliki akurasi terendah, yaitu 0.23, yang menunjukkan bahwa pendekatan probabilistik kurang efektif dalam menangani karakteristik dataset ini. Pada eksperimen menggunakan teknik *oversampling* (70:30), performa model seperti Decision Tree, Random Forest, dan Logistic Regression meningkat, menunjukkan bahwa penambahan sampel pada kelas minoritas dapat membantu model memahami pola data dengan lebih baik. Kemudian metode ini tidak memberikan hasil optimal untuk XGBoost, yang mengalami penurunan akurasi menjadi 0.57. Hal ini dapat terjadi karena *oversampling* berpotensi menyebabkan *overfitting,* terutama pada model berbasis pohon keputusan seperti XGBoost. Sementara itu, teknik *undersampling* (80:20 dan 70:30) menunjukkan hasil yang lebih stabil, terutama untuk Decision Tree dan Logistic Regression, dengan akurasi mencapai 0.87 pada skenario *undersampling* 70:30. Dan teknik ini berdampak negatif pada SVM, yang

mengalami penurunan signifikan hingga 0.38, menunjukkan bahwa metode ini kurang optimal untuk dataset yang telah dikurangi sampelnya.

Setelah menerapkan PCA (90:10) untuk reduksi dimensi fitur, Logistic Regression menunjukkan peningkatan akurasi menjadi 0.81, yang berarti model ini lebih efektif dalam menangani data setelah fitur yang kurang relevan dihilangkan. Model Random Forest dan XGBoost tetap mempertahankan performa yang cukup baik setelah PCA diterapkan, sementara Naïve Bayes meningkat menjadi 0.62, yang menunjukkan bahwa reduksi dimensi dapat membantu meningkatkan kinerja model berbasis probabilistik dengan mengurangi *noise* dalam fitur. PCA juga membantu menghilangkan redundansi antar fitur, yang dapat meningkatkan kecepatan komputasi tanpa mengorbankan akurasi prediksi. Dengan demikian, PCA terbukti menjadi teknik yang berguna dalam meningkatkan efisiensi model. Analisis visualisasi menggunakan *heatmap* korelasi fitur menunjukkan bahwa beberapa fitur dalam dataset memiliki korelasi tinggi, yang dapat menyebabkan redundansi informasi jika tidak diatasi. Korelasi yang tinggi antar fitur dapat menyebabkan *multikolinearitas,* yang dapat mengurangi efektivitas model machine learning. Oleh karena itu, dilakukan seleksi fitur dengan menggunakan *F-score,* yang mengukur tingkat kepentingan masing-masing fitur dalam proses prediksi. Hasil analisis menunjukkan bahwa beberapa fitur memiliki pengaruh yang lebih besar terhadap hasil prediksi, yang berarti seleksi fitur dapat membantu meningkatkan efisiensi model dengan hanya menggunakan fitur yang paling relevan. Model seperti XGBoost dan Random Forest sangat bergantung pada fitur penting dalam melakukan klasifikasi, sehingga seleksi fitur dapat meningkatkan performa model tersebut.

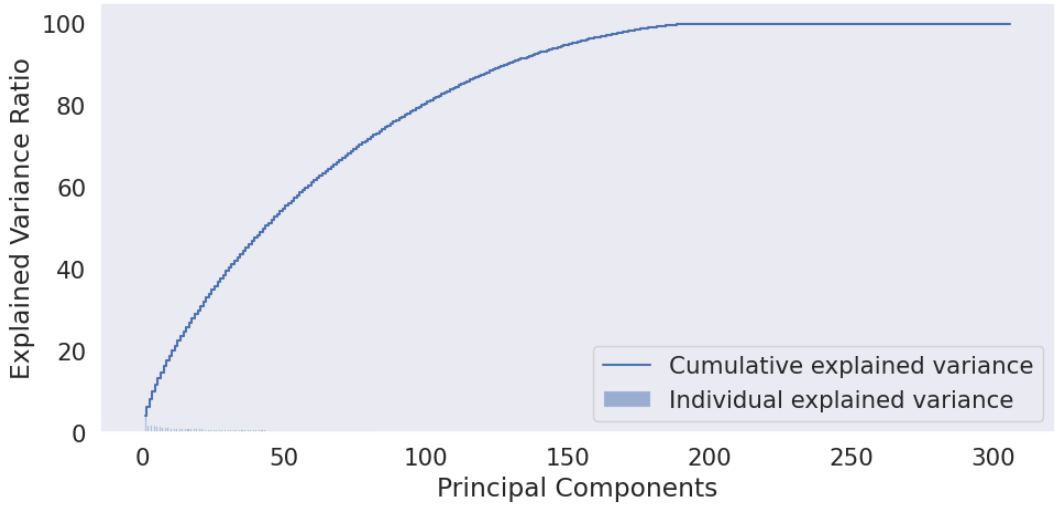
Hasil perbandingan model menunjukkan bahwa XGBoost dan Random Forest adalah model terbaik untuk dataset normal (90:10), dengan akurasi tinggi tanpa perlu *preprocessing* tambahan. Sementara itu, PCA terbukti membantu meningkatkan performa Logistic Regression, yang menandakan bahwa reduksi dimensi dapat meningkatkan efisiensi model. Oversampling bermanfaat bagi Decision Tree dan Logistic Regression dengan meningkatkan akurasi, tetapi dapat menyebabkan *overfitting* pada model seperti XGBoost. Undersampling lebih efektif untuk Decision Tree, tetapi dapat menurunkan kinerja model berbasis *margin* seperti SVM. Naïve Bayes tetap menjadi model dengan performa terendah di semua skenario, yang menunjukkan bahwa model ini kurang cocok untuk dataset manufaktur semikonduktor karena pola distribusi data yang kompleks dan tidak sepenuhnya dapat direpresentasikan dengan pendekatan probabilistik. Selama proses evaluasi model, terdeteksi beberapa *outlier* dalam dataset, yang dapat mengindikasikan adanya faktor eksternal atau variasi dalam data yang tidak sepenuhnya tercakup dalam fitur yang digunakan. Outlier ini dapat menyebabkan kesalahan prediksi dan mengurangi akurasi model. Oleh karena itu, disarankan untuk menerapkan teknik tambahan seperti *anomaly detection* untuk mengidentifikasi dan menangani data yang menyimpang dari pola umum. Selain itu, eksplorasi lebih lanjut dapat dilakukan dengan menambahkan fitur yang lebih representatif atau menggunakan teknik *feature engineering* untuk meningkatkan kualitas dataset. Untuk penelitian selanjutnya, pendekatan dapat diperluas dengan eksplorasi model deep learning atau *ensemble methods* yang lebih kompleks untuk meningkatkan akurasi prediksi. Model seperti Neural Networks atau *Hybrid Ensemble* Model dapat menjadi alternatif yang menjanjikan untuk menangani dataset yang lebih kompleks dan besar. Selain itu, penggunaan dataset yang lebih besar dan lebih beragam akan membantu model meningkatkan kemampuan generalisasi, sehingga dapat diterapkan dalam berbagai skenario manufaktur. Dengan adanya perbaikan pada teknik machine learning dan pemilihan fitur yang lebih optimal, prediksi hasil produk manufaktur semikonduktor dapat semakin akurat dan andal untuk digunakan dalam industri.



Sumber : Hasil Penelitian (2025)

Gambar 3. Distribusi Sensor Pass/Fail

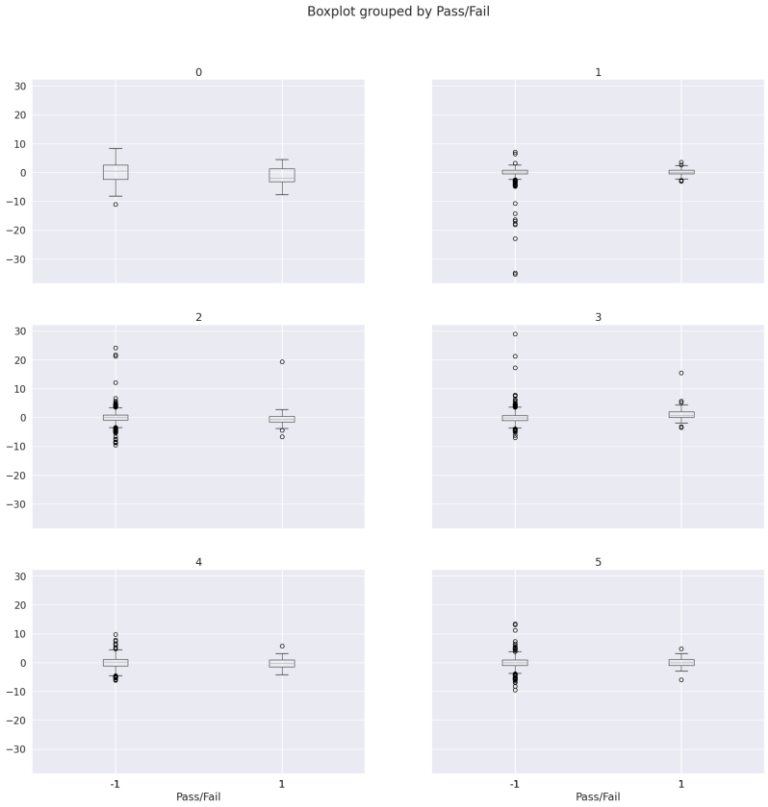
Gambar 3 distribusi pengukuran empat sensor utama dalam proses manufaktur semikonduktor berdasarkan kategori Pass dan Fail. Distribusi data dari setiap sensor dianalisis untuk mengidentifikasi perbedaan pola antara produk yang lolos uji kualitas *(Pass)* dan yang tidak lolos *(Fail).* Dari hasil visualisasi, terlihat bahwa sebagian besar data terkonsentrasi pada rentang nilai tertentu, dengan perbedaan kecil antara kategori Pass dan Fail. Beberapa sensor menunjukkan distribusi yang hampir identik, sementara yang lain memiliki variasi yang lebih besar, yang mengindikasikan bahwa tidak semua sensor memiliki kontribusi signifikan dalam menentukan hasil akhir produk. Analisis ini membantu dalam proses seleksi fitur untuk meningkatkan efisiensi model machine learning dalam memprediksi kualitas produk, dengan mengeliminasi sensor yang kurang relevan.



Sumber : Hasil Penelitian (2025)

Gambar 4. Grafik Variansi PCA

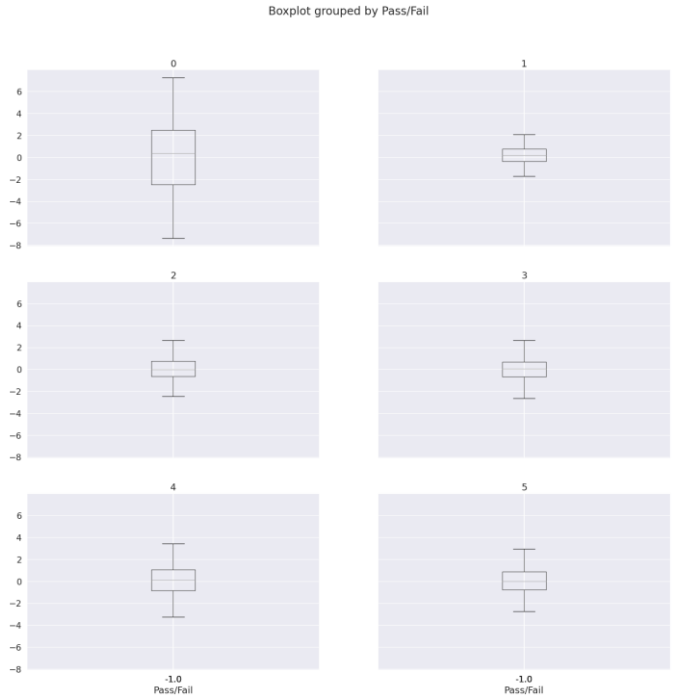
Gambar 4 grafik *explained variance ratio,* yang menggambarkan jumlah variansi yang dapat dijelaskan oleh setiap komponen utama dalam Principal Component Analysis (PCA). Kurva biru menunjukkan cumulative explained variance, yang meningkat seiring dengan jumlah komponen utama yang digunakan. Dapat dilihat bahwa sebagian besar variansi dapat dijelaskan oleh sejumlah kecil komponen pertama, yang menunjukkan bahwa sebagian besar informasi dalam dataset dapat direpresentasikan dengan dimensi yang lebih rendah.



Sumber : Hasil Penelitian (2025)

Gambar 5. Boxplot Fitur Pass/Fail

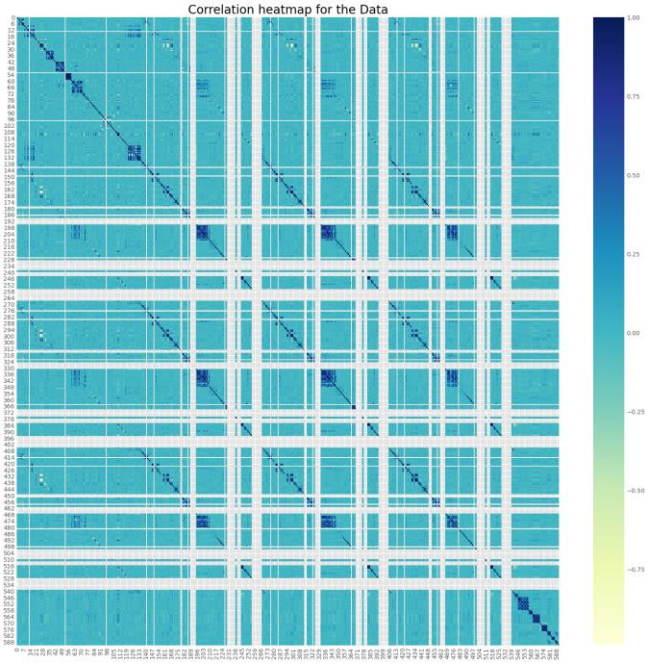
Gambar 5 *boxplot* yang menggambarkan distribusi nilai dari beberapa fitur dalam dataset, dikelompokkan berdasarkan status *Pass/Fail* (-1 untuk fail dan 1 untuk pass). Setiap subplot merepresentasikan distribusi fitur yang berbeda, dengan sumbu horizontal menunjukkan kategori *Pass/Fail* dan sumbu vertikal menunjukkan nilai fitur. Dan sebagian besar fitur memiliki distribusi yang relatif simetris, tetapi beberapa fitur menunjukkan keberadaan outlier yang signifikan. Perbedaan distribusi antara kategori pass dan fail dapat memberikan wawasan tentang fitur mana yang memiliki pengaruh signifikan dalam menentukan kualitas atau keberhasilan suatu entitas dalam dataset.



Sumber : Hasil Penelitian (2025)

Gambar 6. Boxplot Fitur Normalisasi Pass/Fail

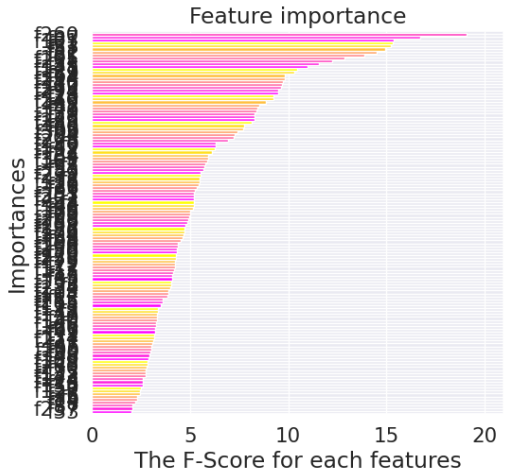
Gambar 6 *boxplot* yang menggambarkan distribusi beberapa fitur dalam dataset setelah dilakukan normalisasi, dikelompokkan berdasarkan status Pass/Fail (-1 untuk fail dan 1 untuk pass). Setiap subplot menunjukkan distribusi dari fitur yang berbeda, dengan sumbu horizontal menampilkan kategori *Pass/Fail,* sedangkan sumbu vertikal menunjukkan nilai fitur yang telah dinormalisasi. Dapat diamati bahwa setelah normalisasi, distribusi data menjadi lebih terpusat di sekitar nol dengan rentang yang lebih seragam antar fitur. Perbedaan pola distribusi antara kategori pass dan fail menunjukkan bahwa beberapa fitur mungkin memiliki kontribusi yang signifikan dalam membedakan antara dua kelas.



Sumber : Hasil Penelitian (2025)

Gambar 7. Heatmap Korelasi Fitur

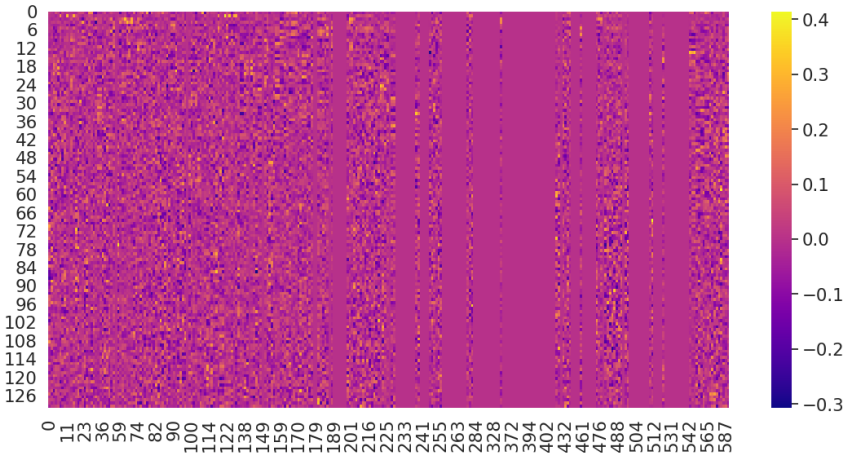
Gambar 7 *heatmap* korelasi yang menunjukkan hubungan antar fitur dalam dataset. Warna biru tua menandakan korelasi positif yang kuat, sedangkan warna kuning terang menunjukkan korelasi negatif atau hubungan yang lemah. Dapat dilihat bahwa beberapa fitur memiliki korelasi tinggi (ditunjukkan oleh diagonal biru tua), yang mengindikasikan kemungkinan redundansi fitur dalam dataset. Korelasi yang tinggi antar fitur dapat menyebabkan multikolinearitas, yang dapat mempengaruhi kinerja model machine learning.



Sumber : Hasil Penelitian (2025)

Gambar 8. Fitur Model Machine Learning

Gambar 8 tingkat kepentingan fitur dalam model machine learning berdasarkan *F-Score,* yang menunjukkan seberapa sering suatu fitur digunakan dalam membagi data selama proses pelatihan model. Semakin tinggi *F-Score* suatu fitur, semakin besar kontribusinya dalam menentukan hasil prediksi model. Fitur dengan skor F yang lebih tinggi memiliki pengaruh yang lebih besar dalam membentuk keputusan model, menandakan bahwa informasi yang dikandungnya lebih relevan terhadap target yang diprediksi. Sebaliknya, fitur dengan skor F yang lebih rendah memiliki kontribusi yang lebih kecil dan dapat dipertimbangkan untuk dihilangkan guna menyederhanakan model tanpa mengorbankan akurasi prediksi.

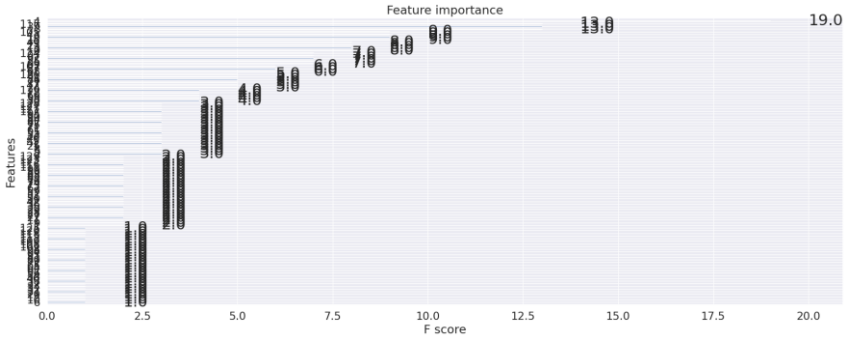


Sumber : Hasil Penelitian (2025)

Gambar 9. Heatmap Representasi Data setelah Transformasi PCA

Gambar 9 *heatmap* yang merepresentasikan nilai dari fitur dataset setelah diterapkan *Principal Component Analysis (PCA).* Warna pada heatmap menggambarkan nilai koefisien komponen utama yang dihasilkan dari transformasi PCA, di mana warna kuning menunjukkan nilai positif yang lebih tinggi, sedangkan warna ungu gelap mengindikasikan nilai negatif yang lebih rendah. Dan beberapa fitur memiliki pola distribusi yang lebih menonjol dibandingkan yang

lain, yang menunjukkan bahwa variansi utama dalam dataset terdistribusi secara tidak merata di sepanjang komponen utama. PCA digunakan untuk mereduksi dimensi data dengan mempertahankan informasi yang paling signifikan, sehingga memungkinkan model machine learning bekerja lebih efisien tanpa kehilangan terlalu banyak informasi penting.



Sumber : Hasil Penelitian (2025)

Gambar 10. Perbandingan Fitur Machine Learning

Gambar 10 perbandingan tingkat kepentingan fitur dalam berbagai model machine learning berdasarkan *F-score*, yang mengukur seberapa sering suatu fitur digunakan dalam pemisahan data selama proses pelatihan model. Sumbu horizontal menunjukkan *F-score,* sementara sumbu vertikal menunjukkan daftar fitur yang digunakan dalam analisis. Hasil analisis menunjukkan bahwa beberapa fitur memiliki kontribusi yang lebih signifikan terhadap prediksi, terutama dalam model seperti XGBoost dan Random Forest, yang cenderung memberikan bobot lebih tinggi pada fitur yang memiliki korelasi kuat dengan target. Sebaliknya, model seperti Naïve Bayes dan Logistic Regression lebih sensitif terhadap distribusi data dan fitur yang memiliki hubungan linier dengan variabel target.

Tabel 1. Perbandingan Akurasi Model pada Berbagai Teknik Pembagian Dataset

Dataset		XG Boost	Decision Tree	Random Forest	Logistic Regression	Naive Bayes	SVM
90:10	**Normal Data**	**0.95**	**0.88**	**0.95**	**0.89**	**0.23**	**0.95**
70:30	Oversampling	0.57	0.65	0.67	0.67	0.54	0.70
80:20	Undersampling	0.56	0.56	0.69	0.56	0.69	0.56
70:30	Undersampled	0.63	0.69	0.87	0.67	0.32	0.38
90:10	PCA	0.67	0.62	0.76	0.81	0.62	0.67

Sumber : Hasil Penelitian (2025)

Tabel 1 membandingkan akurasi model machine learning pada berbagai teknik pembagian dataset. XGBoost, Random Forest, dan SVM menunjukkan akurasi tertinggi (0.95) pada data normal (90:10), sementara Naïve Bayes memiliki akurasi terendah (0.23). Oversampling (70:30) meningkatkan performa Decision Tree dan Logistic Regression, tetapi menurunkan akurasi XGBoost. Undersampling (80:20 & 70:30) lebih efektif untuk Decision Tree dan Logistic Regression, tetapi menurunkan performa SVM. Setelah PCA (90:10), Logistic Regression mengalami peningkatan akurasi (0.81). Hasil ini menunjukkan bahwa XGBoost dan Random Forest optimal untuk data normal, sementara PCA dan oversampling membantu meningkatkan akurasi Logistic Regression dan SVM.

4. Kesimpulan

Penelitian ini berhasil mengembangkan model machine learning untuk memprediksi kualitas produk manufaktur semikonduktor, dengan hasil yang menunjukkan bahwa XGBoost dan Random Forest memberikan akurasi terbaik, sementara Naïve Bayes menunjukkan performa terendah, serta teknik oversampling meningkatkan akurasi pada beberapa model namun berpotensi menyebabkan overfitting, undersampling lebih efektif untuk Decision Tree namun kurang optimal untuk SVM, dan PCA meningkatkan efisiensi model, khususnya pada Logistic Regression, dengan mengurangi dimensi tanpa kehilangan informasi penting, sehingga penelitian ini memberikan pemahaman yang lebih dalam mengenai penerapan machine

learning dalam prediksi kualitas produk semikonduktor, yang dapat membantu industri semikonduktor mengurangi cacat produk dan meningkatkan efisiensi produksi, sementara penelitian selanjutnya dapat memperluas penerapan deep learning atau hybrid ensemble models serta menggunakan dataset yang lebih besar untuk meningkatkan generalisasi model dalam industri manufaktur semikonduktor.

Referensi

[1] X. Y. Li, F. L. Liu, M. N. Zhang, M. X. Zhou, C. Wu, and X. Zhang, "A Combination of Vision- and Sensor-Based Defect Classifications in Extrusion-Based Additive Manufacturing," *J. Sensors*, vol. 2023, 2023, doi: 10.1155/2023/1441936.

[2] N. V. Nguyen, A. J. W. Hum, T. Do, and T. Tran, "Semi-supervised machine learning of optical in-situ monitoring data for anomaly detection in laser powder bed fusion," *Virtual Phys. Prototyp.*, vol. 18, no. 1, 2023, doi: 10.1080/17452759.2022.2129396.

[3] A. Fan, Y. Huang, F. Xu, and S. Bom, "Soft-Sensing Regression Model: From Sensor to Wafer Metrology Forecasting," *Sensors (Basel).*, vol. 23, no. 20, 2023, doi: 10.3390/s23208363.

[4] Y. Wang, W. Cui, N. K. Vuong, Z. Chen, Y. Zhou, and M. Wu, "Feature selection and domain adaptation for cross-machine product quality prediction," *J. Intell. Manuf.*, vol. 34, no. 4, 2023, doi: 10.1007/s10845-021-01875-z.

[5] I. Sideris, F. Crivelli, and M. Bambach, "GPyro: uncertainty-aware temperature predictions for additive manufacturing," *J. Intell. Manuf.*, vol. 34, no. 1, 2023, doi: 10.1007/s10845-022-02019-7.

[6] T. T. H. Vu, T. W. Chang, and H. Kim, "Enhancing Quality Control in Battery Component Manufacturing: Deep Learning-Based Approaches for Defect Detection on Microfasteners," *Systems*, vol. 12, no. 1, 2024, doi: 10.3390/systems12010024.

[7] S. Chen, H. Gao, Y. Zhang, Q. Wu, Z. Gao, and X. Zhou, "Review on residual stresses in metal additive manufacturing: formation mechanisms, parameter dependencies, prediction and control approaches," *J. Mater. Res. Technol.*, vol. 17, 2022, doi: 10.1016/j.jmrt.2022.02.054.

[8] S. K. Sen, G. C. Karmakar, and S. Pang, "Critical Data Detection for Dynamically Adjustable Product Quality in IIoT-Enabled Manufacturing," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3276942.

[9] C. Bak, A. G. Roy, and H. Son, "Quality prediction for aluminum diecasting process based on shallow neural network and data feature selection technique," *CIRP J. Manuf. Sci. Technol.*, vol. 33, 2021, doi: 10.1016/j.cirpj.2021.04.001.

[10] S. Tian, Z. Zhang, X. Xie, and C. Yu, "A new approach for quality prediction and control of multistage production and manufacturing process based on Big Data analysis and Neural Networks," *Adv. Prod. Eng. Manag.*, vol. 17, no. 3, 2022, doi: 10.14743/apem2022.3.439.

[11] F. Psarommatis, M. Danishvar, A. Mousavi, and D. Kiritsis, "Cost-Based Decision Support System: A Dynamic Cost Estimation of Key Performance Indicators in Manufacturing," *IEEE Trans. Eng. Manag.*, vol. 71, 2024, doi: 10.1109/TEM.2021.3133619.

[12] H. Tercan, P. Deibert, and T. Meisen, "Continual learning of neural networks for quality prediction in production using memory aware synapses and weight transfer," *J. Intell. Manuf.*, vol. 33, no. 1, 2022, doi: 10.1007/s10845-021-01793-0.

[13] C. Song, Z. Wu, J. Gray, and Z. Meng, "An RFID-Powered Multisensing Fusion Industrial IoT System for Food Quality Assessment and Sensing," *IEEE Trans. Ind. Informatics*, vol. 20, no. 1, 2024, doi: 10.1109/TII.2023.3262197.

[14] S. Ma, W. Ding, Y. Liu, S. Ren, and H. Yang, "Digital twin and big data-driven sustainable smart manufacturing based on information management systems for energy-intensive industries," *Appl. Energy*, vol. 326, 2022, doi: 10.1016/j.apenergy.2022.119986.

[15] T. Wang, B. Hu, Y. Feng, X. Gao, C. Yang, and J. Tan, "Data Augmentation-Based Manufacturing Quality Prediction Approach in Human Cyber-Physical Systems," *J. Manuf. Sci. Eng.*, vol. 145, no. 12, 2023, doi: 10.1115/1.4063269.

[16] H. Zhou, K. M. Yu, Y. C. Chen, and H. P. Hsu, "A Hybrid Feature Selection Method RFSTL for Manufacturing Quality Prediction Based on a High Dimensional Imbalanced

Dataset," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3059298.
[17] T. Batu, H. G. Lemu, and H. Shimels, "Application of Artificial Intelligence for Surface Roughness Prediction of Additively Manufactured Components," 2023. doi: 10.3390/ma16186266.
[18] A. Khdoudi, N. Barka, T. Masrour, I. El-Hassani, and C. El Mazgualdi, "Online prediction of automotive tempered glass quality using machine learning," *Int. J. Adv. Manuf. Technol.*, vol. 125, no. 3–4, 2023, doi: 10.1007/s00170-022-10649-7.
[19] D. Liu, Y. Du, W. Chai, C. Q. Lu, and M. Cong, "Digital Twin and Data-Driven Quality Prediction of Complex Die-Casting Manufacturing," *IEEE Trans. Ind. Informatics*, vol. 18, no. 11, 2022, doi: 10.1109/TII.2022.3168309.
[20] H. Helgers *et al.*, "Towards autonomous operation by advanced process control—process analytical technology for continuous biologics antibody manufacturing," *Processes*, vol. 9, no. 1, 2021, doi: 10.3390/pr9010172.
[21] F. Xiang, L. Yang, M. Zhang, Y. Zuo, X. F. Zou, and F. Tao, "Model fusion based product quality prediction for complex manufacturing process," *Zhongguo Kexue Jishu Kexue/Scientia Sin. Technol.*, vol. 53, no. 7, 2023, doi: 10.1360/SST-2022-0427.
[22] H. Jung, J. Jeon, D. Choi, and A. J. Y. Park, "Application of machine learning techniques in injection molding quality prediction: Implications on sustainable manufacturing industry," *Sustain.*, vol. 13, no. 8, 2021, doi: 10.3390/su13084120.
[23] L. P. Zhao, B. H. Li, and Y. Y. Yao, "A novel predict-prevention quality control method of multi-stage manufacturing process towards zero defect manufacturing," *Adv. Manuf.*, vol. 11, no. 2, 2023, doi: 10.1007/s40436-022-00427-9.
[24] M. Papananias, T. E. McLeay, M. Mahfouf, and V. Kadirkamanathan, "A Bayesian information fusion approach for end product quality estimation using machine learning and on-machine probing," *J. Manuf. Process.*, vol. 76, 2022, doi: 10.1016/j.jmapro.2022.01.020.
[25] C. Ruiz, D. Jafari, V. Venkata, T. H. J. Vaneker, W. Ya, and Q. Huang, "Prediction and Control of Product Shape Quality for Wire and Arc Additive Manufacturing," *J. Manuf. Sci. Eng.*, vol. 144, no. 11, 2022, doi: 10.1115/1.4054721.
[26] C. H. Chien, P. Y. Chen, A. J. C. Trappey, and C. V. Trappey, "Intelligent Supply Chain Management Modules Enabling Advanced Manufacturing for the Electric-Mechanical Equipment Industry," *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/8221706.
[27] J. Gim and L. S. Turng, "A review of current advancements in high surface quality injection molding: Measurement, influencing factors, prediction, and control," 2022. doi: 10.1016/j.polymertesting.2022.107718.
[28] N. Leberruyer, J. Bruch, M. Ahlskog, and S. Afshar, "Toward Zero Defect Manufacturing with the support of Artificial Intelligence—Insights from an industrial application," *Comput. Ind.*, vol. 147, 2023, doi: 10.1016/j.compind.2023.103877.
[29] J. Takalo-Mattila, M. Heiskanen, V. Kyllonen, L. Maatta, and A. Bogdanoff, "Explainable Steel Quality Prediction System Based on Gradient Boosting Decision Trees," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3185607.
[30] X. Li, Z. Huang, and W. Ning, "Intelligent manufacturing quality prediction model and evaluation system based on big data machine learning," *Comput. Electr. Eng.*, vol. 111, 2023, doi: 10.1016/j.compeleceng.2023.108904.
[31] V. Azamfirei, F. Psarommatis, and Y. Lagrosen, "Application of automation for in-line quality inspection, a zero-defect manufacturing approach," 2023. doi: 10.1016/j.jmsy.2022.12.010.
[32] S. Nannapaneni, S. Mahadevan, A. Dubey, and Y. T. T. Lee, "Online monitoring and control of a cyber-physical manufacturing process under uncertainty," *J. Intell. Manuf.*, vol. 32, no. 5, 2021, doi: 10.1007/s10845-020-01609-7.
[33] M. Papananias, T. E. McLeay, M. Mahfouf, and V. Kadirkamanathan, "A probabilistic framework for product health monitoring in multistage manufacturing using Unsupervised Artificial Neural Networks and Gaussian Processes," *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.*, vol. 237, no. 9, 2023, doi: 10.1177/09544054221136510.