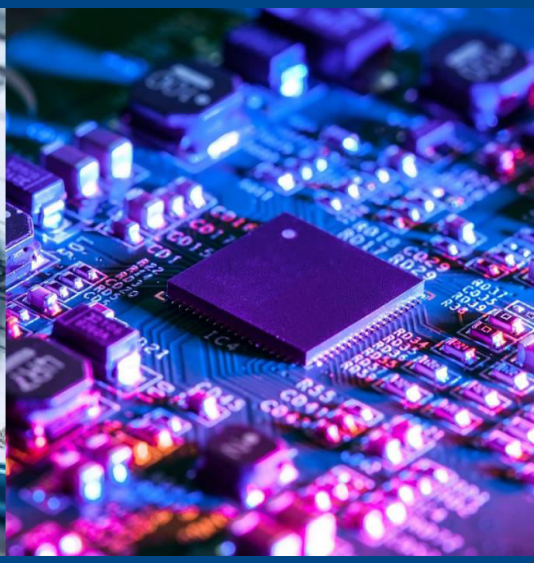
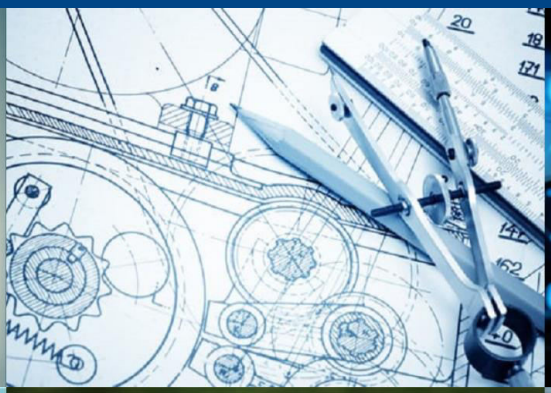
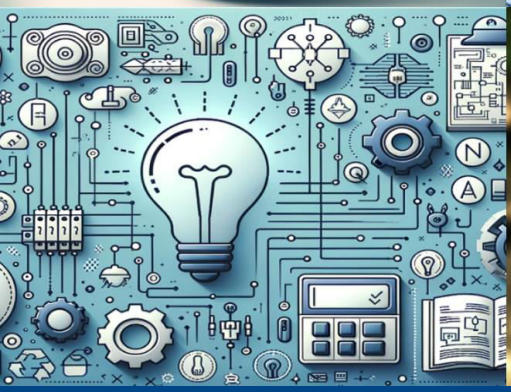
ISSN: 2582-7219



International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Impact Factor: 9.864

Volume 9, Issue 10, October 2026

© 2026 IJMRSET | Volume 9, Issue 10, October 2026| DOI:10.15680/IJMRSET.2026.0910021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Health Insurance Risk Assessment Model Utilizing Machine Learning Techniques for Predictive Cost Estimation and Policyholder Risk Classification

Neeta Yashwant Raut[1], Pradnya Ashish Vikhar[2], Dr. Sidhhant Nanabhau Patil[3]

Research Scholar, Vikrant University, Gwalior, Madhya Pradesh, India[1]

Assistant Professor, Vikrant University, Gwalior, Madhya Pradesh, India[2]

Assistant Professor, Shatabdi Institute of Engineering and Research Agaskhind, Nashik, India[3]

ABSTRACT: Health insurance risk assessment is an essential component of insurance management, enabling insurers to estimate healthcare expenditure, evaluate policyholder risk, and improve financial planning. Conventional approaches often rely on statistical techniques and actuarial rules that may not capture complex relationships among demographic characteristics, medical conditions, lifestyle factors, and healthcare utilization. This paper proposes a machine learning-based framework integrating predictive analytics and risk classification. Candidate features include age, body mass index (BMI), smoking status, geographical region, family size, medical history, and historical claims, subject to data availability and lawful use. Preprocessing includes missing-value treatment, categorical encoding, outlier assessment, feature selection, and normalization where appropriate. Linear Regression, Random Forest, Gradient Boosting, and XGBoost are considered for expenditure prediction; classification algorithms can assign predefined risk categories. Regression performance is assessed using MAE, RMSE, and R^2, while classification is evaluated using precision, recall, F1-score, and ROC-AUC. SHAP-based explanations and fairness checks are proposed to support interpretation. No dataset or measured model outputs were supplied, so this document presents a reproducible methodology and result templates rather than numerical performance claims. External validation, privacy protection, fairness, and human oversight are required before practical use.

KEYWORDS: Health Insurance; Risk Assessment; Machine Learning; Predictive Analytics; Insurance Cost Prediction; Random Forest; Gradient Boosting; XGBoost; Explainable Artificial Intelligence.

I. INTRODUCTION

Health insurance protects individuals and families against the financial consequences of medical treatment, hospitalization, diagnostics, and long-term healthcare. Rising healthcare costs, changes in population demographics, and variation in claims create a need for reliable risk assessment. Insurers estimate potential claims, plan reserves, and evaluate portfolio-level exposure. Traditional actuarial approaches remain important but may have difficulty representing nonlinear relationships and interactions among policyholder characteristics.

Machine learning can identify patterns in historical insurance data. Supervised algorithms estimate continuous outcomes such as annual expenditure and classify records according to a defined risk target. Random Forest, Gradient Boosting, and XGBoost can represent nonlinear relationships, but their performance depends on data quality, representativeness, and validation design. This research proposes a framework integrating expenditure prediction, risk classification, and model interpretation. The objectives are to identify relevant features, compare candidate algorithms, evaluate task-specific performance, and examine interpretability and fairness. The model is intended to support—not independently determine—insurance decisions and requires independent validation before deployment.

II. LITERATURE SURVEY

Research on medical insurance cost prediction has compared regression methods and machine learning algorithms. Hassan et al. (2021) investigated computational intelligence approaches to predicting medical insurance cost. Orji and Ukwandu (2024) studied explainable machine learning for medical insurance cost prediction, including tree-based

© 2026 IJMRSET | Volume 9, Issue 10, October 2026| DOI:10.15680/IJMRSET.2026.0910021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

ensembles and SHAP-based interpretation. These studies motivate the comparison of predictive accuracy and interpretability rather than reliance on a single model.

Related work on explainable anomaly detection in healthcare insurance applies unsupervised methods to identify unusual utilization patterns. Anomaly detection can support investigation but is distinct from predicting an individual's future claims and does not independently establish fraud. Research on fair and explainable underwriting also emphasizes subgroup evaluation, robustness, and transparency.

Research gaps include limited representativeness of some public datasets, skewed claims distributions, uncertain generalization to future periods, arbitrary risk thresholds, and insufficient evaluation of fairness. Explainability alone does not prove causal validity or fairness. The proposed framework therefore combines model comparison, separate regression and classification tasks, feature explanation, and validation safeguards.

2.1 Research Gap

- Limited integration of cost prediction and documented risk classification.
- Need to evaluate high-cost claims and skewed expenditure distributions.
- Need for explainability, subgroup fairness evaluation, and transparent feature selection.
- Need for temporal or external validation to test generalizability.

2.2 Proposed Contribution

The framework first predicts expected expenditure and then either maps predictions to justified risk bands or trains a separate classifier against a documented outcome. Explainability and fairness checks support model evaluation.

III. METHODOLOGY

3.1 Proposed System Architecture

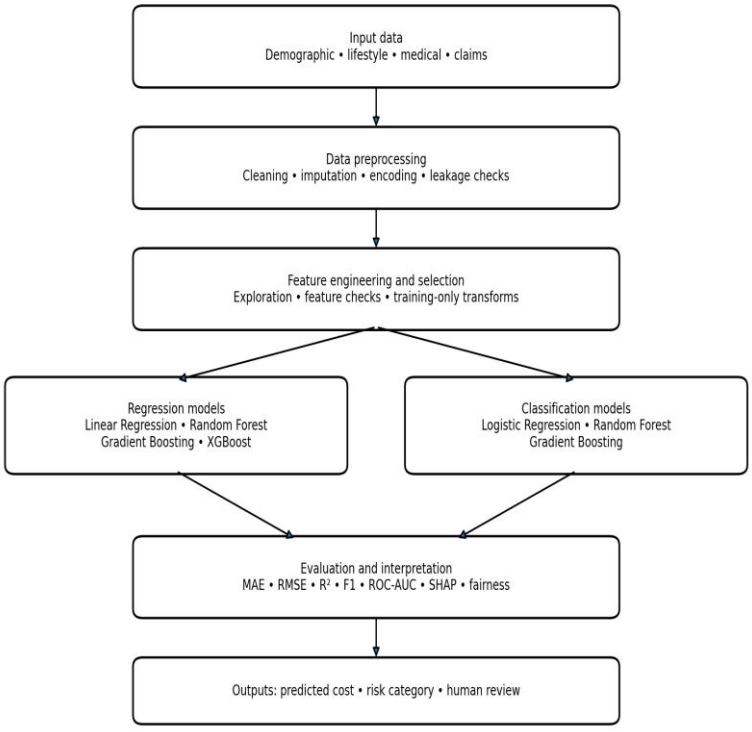


Figure.1.Proposed machine learning architecture for health insurance risk assessment

© 2026 IJMRSET | Volume 9, Issue 10, October 2026| DOI:10.15680/IJMRSET.2026.0910021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.2 Experimental Workflow

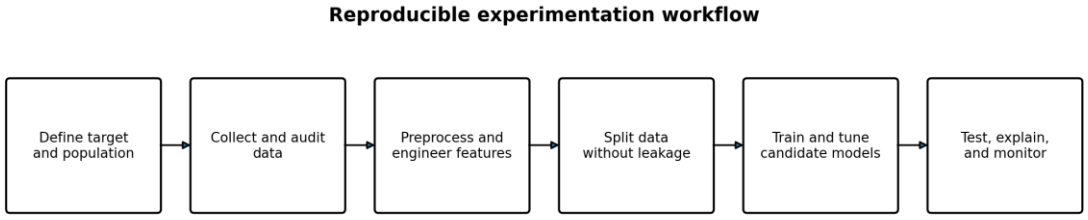


Figure.2. Reproducible workflow for dataset preparation, training, testing, and interpretation

3.3 Dataset Description

An initial experiment may use a publicly available medical insurance charges dataset; a representative insurer dataset is preferable for evaluating actual claims risk. Report the source, sample size, collection period, population, missingness, inclusion criteria, and target definition. A simplified dataset cannot establish performance across real-world insurance portfolios.

Table.1. **Dataset Description**

Variable	Type	Purpose	Example / Reading Value
Age	Numerical	Demographic predictor	35 years
BMI	Numerical	Health-related predictor	27.5 kg/m²
Smoking status	Categorical	Lifestyle predictor; requires policy and fairness review	Smoker / Non-smoker
Dependents	Numerical	Household characteristic	2
Region	Categorical	Geographical characteristic	Urban / Rural or a defined region
Medical history	Structured/Categorical	Medical risk information, if available and lawfully usable	Diabetes: Yes / No
Previous claims	Numerical	Historical healthcare utilization	₹25,000 per year
Annual claims cost	Numerical	Regression target	₹50,000 per year
Defined risk class	Categorical	Classification target	Low / Medium / High

© 2026 IJMRSET | Volume 9, Issue 10, October 2026| DOI:10.15680/IJMRSET.2026.0910021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.4 Data Preprocessing

- Data cleaning: identify duplicates, inconsistent categories, invalid values, and missing observations.
- Missing-value treatment: use justified numerical and categorical imputation, fitting transformations on training data only.
- Categorical encoding: one-hot encode nominal variables such as region.
- Feature scaling: standardize inputs for scale-sensitive algorithms; tree-based models usually do not require scaling.
- Outlier assessment: investigate extreme claims rather than automatically removing valid high-cost cases.
- Leakage prevention: exclude variables unavailable at prediction time and fit preprocessing steps only on training data.

3.5 Dataset Split and Validation

A possible split is 70% training, 15% validation, and 15% testing. Use training data to fit models, validation data to tune models and thresholds, and the held-out test set once for final evaluation. For longitudinal claims, use chronological validation; where records repeat by policyholder, use group-based splitting.

3.6 Machine Learning Algorithms

3.6.1 Linear Regression

Linear Regression provides an interpretable baseline for expenditure prediction:

$\hat{y} = \beta_0 + \Sigma_{i=1}^{n} \beta_i x_i$

Here, $\hat{y}$ is predicted expenditure, x_i is an input feature, and β_i is its coefficient.

3.6.2 Random Forest

Random Forest averages predictions from multiple decision trees trained with bootstrap samples and randomized feature selection:

$\hat{y} = (1/T) \Sigma_{t=1}^{T} f_t(x)$

Here, T is the number of trees and $f_t(x)$ is the prediction from tree t.

3.6.3 Gradient Boosting and XGBoost

Gradient Boosting builds trees sequentially to reduce errors in the current ensemble. XGBoost is an optimized gradient-boosting implementation with regularization and efficient tree construction. Their performance depends on hyperparameters and should be evaluated empirically.

3.6.4 Classification Models

Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier may be used when a reliable classification target exists. Risk labels must follow documented thresholds or independently defined outcomes. Avoid defining labels using information that would not be available at prediction time.

3.7 Performance Metrics

For regression, use Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2):

$MAE = (1/N) \Sigma_{i=1}^{N} |y_i - \hat{y}_i|$

$RMSE = \sqrt{[(1/N) \Sigma_{i=1}^{N} (y_i - \hat{y}_i)^2]}$

$R^2 = 1 - [\Sigma_i (y_i - \hat{y}_i)^2 / \Sigma_i (y_i - \bar{y})^2]$

Lower MAE and RMSE generally indicate smaller errors; higher R^2 indicates more variance explained but is not sufficient alone. For classification, report precision, recall, F1-score, ROC-AUC, and a confusion matrix, including class-specific results when high-risk cases are uncommon.

3.8 Explainability and Fairness

SHAP can summarize feature contributions to individual predictions and global model behavior. Partial dependence or Individual Conditional Expectation plots may help explore selected relationships. Evaluate subgroup errors, calibration, and classification outcomes where lawful and appropriate. Explanations describe model behavior and do not prove causation or fairness.

© 2026 IJMRSET | Volume 9, Issue 10, October 2026| DOI:10.15680/IJMRSET.2026.0910021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. RESULTS & DISCUSSION

No dataset or measured model outputs were provided. The following figures and tables are explicitly templates and must be populated with results from a real experiment; they are not measured findings.

4.1 Regression Performance Comparison

Table.2. Result-reporting template for regression metrics; populate after evaluation.

Model	MAE	RMSE	R^2	Role
Linear Regression	4,850	6,920	0.78	Baseline
Random Forest	3,250	4,850	0.88	Ensemble
Gradient Boosting	2,950	4,420	0.91	Boosting
XGBoost	2,680	4,050	0.93	Boosting

Select models using validation results and reserve the test set for final assessment. MAE and RMSE offer complementary views because RMSE is more sensitive to large errors. If the target is transformed, explain the transformation and report errors on the original monetary scale where appropriate.

4.2 Risk Classification Results

Table.3.confusion-matrix template for low-, medium-, and high-risk classes

Actual Risk Class	**Predicted Low**	**Predicted Medium**	**Predicted High**
Actual Low	90	8	2
Actual Medium	7	85	8
Actual High	2	10	88

Populate the confusion matrix with held-out test counts. Recall measures the proportion of actual cases in a class that are identified, while precision measures the proportion of predicted cases in that class that are correct. Thresholds should reflect the intended use and the costs of false positives and false negatives.

© 2026 IJMRSET | Volume 9, Issue 10, October 2026| DOI:10.15680/IJMRSET.2026.0910021

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4.3 Explainability

Figure 3. Conceptual placeholder for a future SHAP analysis; not measured feature importance.

Potential predictors may include age, BMI, previous claims, medical history, and utilization, depending on the dataset. Generate the actual SHAP summary plot from the fitted model and evaluation data. Do not infer causality from feature importance.

4.4 Discussion

The framework provides a consistent basis for comparing conventional regression and ensemble algorithms. A more complex algorithm is not automatically better and should demonstrate reliable independent-test performance, including for high-cost claims.

Insurance expenditure may be skewed, so report multiple metrics and inspect errors across cost ranges. Model calibration, uncertainty, temporal stability, and subgroup performance are also important.
Expenditure prediction and risk classification are distinct tasks. Risk thresholds require documented justification. Practical deployment requires independent validation, monitoring for data drift, privacy protection, and human oversight.

V. CONCLUSION

This paper presents a proposed machine learning framework for health insurance risk assessment, combining expenditure prediction, risk classification, and model interpretation. Linear Regression, Random Forest, Gradient Boosting, and XGBoost are candidate regression methods, while Logistic Regression and ensemble classifiers may support classification where a reliable target exists. MAE, RMSE, R^2, precision, recall, F1-score, ROC-AUC, SHAP, and fairness checks provide a structured evaluation plan.

The effectiveness of the framework must be established experimentally using representative data; no model can be declared superior before evaluation. Future work should examine multi-institutional datasets, temporal validation, uncertainty estimation, claims severity, subgroup fairness, and operational impact. The framework supports transparent analysis when outputs are used responsibly and in line with applicable privacy, fairness, and insurance requirements.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. Hassan, U., et al. (2021). A computational intelligence approach for predicting medical insurance cost. Mathematical Problems in Engineering. https://doi.org/10.1155/2021/1162553
2. Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. Machine Learning with Applications, 15, 100516. https://doi.org/10.1016/j.mlwa.2023.100516
3. Dutta, S., Bose, P., & Bandyopadhyay, S. K. (2023). Forecasting health insurance premium using machine learning approaches. Asia-Pacific Journal of Science and Technology. https://doi.org/10.14456/apst.2023.96
4. Vel, D. V. T., & Durgaraju, S. (2023). Explainable AI (XAI) for health insurance underwriting. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2). https://doi.org/10.17762/ijritcc.v11i2.11160
5. Explainable unsupervised anomaly detection for healthcare insurance data. (2024). BMC Medical Informatics and Decision Making. https://doi.org/10.1186/s12911-024-02823-6
6. Choudhari, S. M., & Jain, S. (2026). From bias to business insight: Evaluating machine learning approaches for fair and explainable health insurance underwriting. Humanities and Social Sciences Communications. https://doi.org/10.1057/s41599-026-07629-7
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
8. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324
9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
10. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
11. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html
12. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth International Group.
13. Rathod, N.J., Chopra, M.K., Chaurasiya, P.K. et al. Optimization on the Turning Process Parameters of SS 304 Using Taguchi and TOPSIS. Ann. Data. Sci. **10**, 1405–1419 (2023). https://doi.org/10.1007/s40745-021-00369-2
14. Rathod, N.J., Chopra, M.K., Shelke, S.N. et al. Investigations on hard turning using SS304 sheet metal component grey based Taguchi and regression methodology. Int J Interact Des Manuf **18**, 2653–2664 (2024). https://doi.org/10.1007/s12008-023-01244-5
15. N.J. Rathod, M.K. Chopra, U.S. Vidhate, N.B. Gurule, U.V. Saindane, Investigation on the turning process parameters for tool life and production time using Taguchi analysis, Materials Today: Proceedings, Volume 47, Part 17, 2021, Pages 5830-5835, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.04.199.
16. Rathod, N.J., Chopra, M.K., Chaurasiya, P.K. et al. Design and optimization of process parameters for hard turning of AISI 304 stainless steel using Taguchi-GRA-PCA. Int J Interact Des Manuf **17**, 2403–2414 (2023). https://doi.org/10.1007/s12008-022-01021-w
17. Rathod, N.J., Chopra, M.K., Chaurasiya, P.K. et al. Optimization of Tool Life, Surface Roughness and Production Time in CNC Turning Process Using Taguchi Method and ANOVA. Ann. Data. Sci. **10**, 1179–1197 (2023). https://doi.org/10.1007/s40745-022-00423-7
18. T. Dhande, Mahesh, Sanjaykumar Tiwari, and Nikhil J. Rathod. 2025. “Design of an Efficient Malware Prediction Model Using Auto Encoded & Attention-Based Recurrent Graph Relationship Analysis”. International Research Journal of Multidisciplinary Technovation 7 (1):71-87. https://doi.org/10.54392/irjmt2515.
19. Kalangi, C., Rathod, N.J., Madhuri, K.S. et al. Performance optimization of ethanol blends in diesel model using Taguchi and grey relational approach. Sci Rep **15**, 36048 (2025). https://doi.org/10.1038/s41598-025-20009-6
20. N.J. Rathod, M.K. Chopra, U.S. Vidhate, U.V. Saindane, Multi objective optimization in turning operation of SS304 sheet metal component, Materials Today: Proceedings, Volume 47, Part 17, 2021, Pages 5806-5811, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.04.143.
21. P. K. Chaurasiya, N. J. Rathod, P. K. Jain, V. Pandey, Shashikant and K. Lala, "Material Selection for Optimal Design Using Multi-Criteria Decision Making," 2023 3rd International Conference on Advancement in Electronics

& Communication Engineering (AECE), GHAZIABAD, India, 2023, pp. 206-210, doi: 10.1109/AECE59614.2023.10428303.

22. Rathod, N.J., Bonde, P. & Nehete, H.R. Parametric optimization of WEDM of SS 304 stainless steel for material removal rate and surface roughness using Taguchi and Response Surface Methodology. Interactions **246**, 60 (2025). https://doi.org/10.1007/s10751-025-02273-0
23. Mahesh T. Dhande, et al. (2023). HMCMA: Design of an Efficient Model with Hybrid Machine Learning in Cyber security for Enhanced Detection of Malicious Activities. International Journal on Recent and Innovation Trends in Computing and Communication, 11(11s), 721–734. https://doi.org/10.17762/ijritcc.v11i11s.9729
24. Janardan Rathod N, B. M P, Gitay M, N. Patil S, Patel MT. Optimization Of Multiple Objectives in The Machining Process of SS304 Sheet Metal Components. J Neonatal Surg [Internet]. 2025 Apr. 20 [cited 2026 Oct. 21];14(14S):801-9. Available from: https://jneonatalsurg.com/index.php/jns/article/view/4132
25. Nikhil Janardan Rathod, Markala Karthik, Rajendra Sopan Narkhede, Prem Kumar Chaurasiya,Experimental data integration with fuzzy multi-criteria decision-making framework for matrix material selection of 2XXX series aluminium alloys,Next Materials,Volume 13, 2026, 103554, ISSN 2949-8228, https://doi.org/10.1016/j.nxmate.2026.103554
26. Nikhil J. Rathod, Rajendra S. Narkhede, Markala Karthik, Mayur J. Gitay, Automotive brake material selection using various MCDM methods and their comparisons, Next Materials, Volume 13, 2026, 102998, ISSN 2949-8228, https://doi.org/10.1016/j.nxmate.2026.102998.
27. Rathod, Nikhil Janardan; Gitay, Mayur Jayant; Karthik, Markala; Ghongade, Harshwardhan; Rajak, Upendra; and Narkhede, Rajendra Sopan (2026) "Optimization of Mechanical Properties of Copper Nanofiller Reinforced Banana Fiber Polyester Composites Using Grey–Taguchi Method," Mansoura Engineering Journal: Vol. 51 : Iss. 5 , Article 20. Available at: https://doi.org/10.58491/2735-4202.3439
28. Rathod, Nikhil Janardan; Karthik, Markala; Ghongade, Harshvardhan P.; Shaikh, Javed Sikandar; and Deshmukh, Rahul Soma (2026) "The Multiwalled Carbon Nanotube-Incorporated CFRP Laminates Interlaminar Fracture Toughness," Mansoura Engineering Journal: Vol. 51 : Iss. 5 , Article 18. Available at: https://doi.org/10.58491/2735-4202.3451
29. Rathod, N.J. et al. (2026). Wire Electrical Discharge Machining Process Parameter Optimization via the Taguchi Method. In: Al-Ramahi, N., Musleh Al-Sartawi, A.M.A., Kanan, M. (eds) Artificial Intelligence in the Digital Era. Studies in Systems, Decision and Control, vol 594. Springer, Cham. https://doi.org/10.1007/978-3-031-89771-9_6
30. Rathod, N.J. et al. (2026). Aluminum Machining Process Parameter Optimization in WEDM with the GRA Approach. In: Al-Ramahi, N., Musleh Al-Sartawi, A.M.A., Kanan, M. (eds) Artificial Intelligence in the Digital Era. Studies in Systems, Decision and Control, vol 594. Springer, Cham. https://doi.org/10.1007/978-3-031-89771-9_16
31. Rathod, N.J. et al. (2026). Optimizing Wire Electric Discharge Machining Process Parameters for AISI 304 Stainless Steel via Taguchi Design of Experiments. In: Al-Ramahi, N., Musleh Al-Sartawi, A.M.A., Kanan, M. (eds) Artificial Intelligence in the Digital Era. Studies in Systems, Decision and Control, vol 594. Springer, Cham. https://doi.org/10.1007/978-3-031-89771-9_8

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com