International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

An Intelligent Autonomous System for Data Preprocessing, Model Selection, and Predictive Analysis

Jatothu Praveen
Department of Computer Science and Artificial Intelligence Central University of Andhra Pradesh
Email: jatothupraveen11@gmail.com

Mr.Y.Dayanand Kumar (Assistant Professor)
Department of Computer Science and Artificial Intelligence
Central University of Andhra Pradesh
Email: dayanandkumar.cuap.edu.in

Abstract
There have been significant advancements in artificial intelligence and machine learning technologies which have dramatically changed how Predictive Analytics and intelligent decision-making systems work within numerous sectors. However, traditional machine learning processes involve a lot of manual input from humans throughout various stages of data preprocessing, feature extraction/selection, model choice, and hyperparameter optimization. As such, the entire workflow is labor-intensive, time consuming and heavily reliant upon specific domain knowledge or expertise. This study presents an autonomous system for automated dataprocessing/preprocessing, model selection, and predictive analysis utilizing novel Automated Machine Learning (automl) techniques and intelligent optimization methods. The proposed autonomous system will include an array of automated preprocessing operations such as: missing value imputation; outlier detection; feature scaling; feature engineering; intelligent data transformation with an adaptive strategy for selecting appropriate models and their associated hyperparameters. To optimize both predictive accuracy and computational efficiency, this system utilizes a variety of optimization techniques including: bayesian optimization; Hyperband tuning; reinforcement learning-based optimization; and automl methodologies. In addition, several different machine learning models, including Random Forest; XGBoost; LightGBM; and AutoGluon are tested to determine the most accurate predictive models. Findings from experimental results demonstrate the proposed autonomous system is significantly better than typical machine learning workflows with respect to efficiency of preprocessing; predictive performance; scalability; and reliability of decisions made by Intelligent Systems. Furthermore, this study highlights the significance of developing intelligent autonomous systems to automate end-to-end Predictive Analytics applications across a range of sectors including health care forecasting; financial prediction; education analytics; industrial automation; and smart systems. Finally, this research contributes towards the development of next generation autonomous artificial intelligence systems capable of reducing human involvement and increasing predictive intelligence through fully Automated Machine Learning workflows.

Keywords:
Automated Machine Learning, Predictive Analytics, Intelligent Systems, Bayesian Optimization, Reinforcement Learning, data preprocessing, automl, artificial intelligence

1. INTRODUCTION

Due to increasing use of digital systems, cloud computing infrastructures, AI systems we are getting enormous amount of structured and unstructured data in every domain (Feurer et al., 2015; Wang et al., 2021). Understanding the data deeply for enabling intelligent prediction and decision-making is required in health-care, financial, education, manufacturing, transportation, smart cities and many more domain. Machine learning and AI are powerful tools for solving complex problem such as prediction, classification, anomaly detection, optimization and automation. Majority of multi-stage ML pipelines i.e. Data pre-processing, feature engineering, model selection, hyper-parameter optimization and prediction require tremendous human effort and computing resources. Data pre-processing plays a vital role in machine learning as its performance directly impact accuracy and reliability of models (Koka et al., 2025). Missing values, noise, duplicates, outliers, class imbalance in data harm ML algorithms and make prediction result poor. Normally, the data pre-processing and cleaning process is human operated, specific domain expertise required and makes the process non-scalable. Automating data pre-processing and transformations using smart system and adaptive optimization is a subject of research. AutoML is the emerging field, which aims at automating entire ML pipelines. Automated ML pipeline for machine learning tasks i.e. Model selection, feature engineering and hyperparameter optimization were proposed for manual systems (where human operate the data pre-processing and these tasks). Auto-Sklearn automate the ML pipeline effectively with Bayesian optimization and meta-learning techniques (Feurer et al., 2015). TPOT is the method which automatically build data science pipeline using genetic programming based tree-based pipeline optimization, this search exhaustively for Pythonic combination of data preprocessing and model selection operators (Olson et al., 2016). The modern intelligent system need sophisticated search technique i.e. Bayesian Optimization and Hyperband to intelligently search for model and hyper-parameter for prediction analysis (Snoek et al., 2012; Li et al., 2018). Bayesian optimization has become popular method, which overcome computationally challenging hyper-parameter search space, provides intelligent prediction by optimizing them resulting better prediction than random search. Hyperband optimize

faster with reduced computation than other technique through use of early stopping policy, reducing the search space by reallocating resources by stopping computation early on for which model's performance is not significantly improve. Optimization based on reinforcement learning have been researched well to design effective smart system. Even with all above improvement of AutoML and prediction system, the current system and techniques are still not fully automated and not fully intelligent. Some of AutoML techniques target single ML stages such as hyperparameter optimization, model search whereas few are developed to integrate data preprocessing, ML pipeline automation and prediction system in one framework. Efficient computation and interpretation are very important to develop them. An intelligent autonomous system framework for automating data pre- processing, feature engineering, model selection, hyper-parameter optimization with RL and prediction analysis is proposed to improve predictive performance significantly and reduce human effort dramatically.

Objectives of this research is to:

1. Develop an intelligent data pre-processing system for datasets with missing values, noise and outliers.
2. Design the automated ML model selection for prediction using AutoML.
3. Performance measurement of Bayesian Optimization and Hyperband based models for prediction analysis.
4. Optimize computation cost and maximize the prediction accuracy.
5. Construct a scalable autonomous intelligent AI framework for smart decision-making systems.

This will contribute to make the autonomous AI system more robust and efficient for smart decision-making system by automatically performing all ML workflows.

2.METHODS

2.1 Study Design

In this project, a comparative and experimental methodology were applied to design and evaluate an Intelligent Autonomous System(IAS) to accomplish automated tasks regarding data preprocessing, model selection and prediction analytics. In addition, the IAS was developed as a solution for automation to facilitate numerous repetitive tasks associated with Machine Learning Lifecycle which can be summarized as intelligent preprocessing, feature extraction, adaptive model selection, automatic hyper parameter tuning, and predictive forecasting. The primary goal of developing IAS was to yield a scalable, automatic, optimal and dependable predictive solution. Programming language Python was adopted for building the IAS, integrating several ubiquitous machine learning libraries like Scikit-learn, XGBoost, LightGBM, AutoGluon, FLAML, Pandas, NumPy, and Streamlit for enabling the visualization of IAS in real-time, and deployment of IAS.

2.2 Data Collection

The datasets were obtained from online machine learning repositories such as the UCI Machine Learning Repository and Kaggle. This enabled a comparison of predictive accuracy on real-world data by using structured tabular data types with both numerical and categorical features.

Data Set	Records	Features	Target Type
Sales Forecasting Dataset	10,000	18	Regression
Customer Analytics Dataset	8,500	22	classification
Financial Prediction Dataset	12,000	25	Time-series
Health care Analytics Dataset	6,200	15	Classification

2.3 Intelligent Data Preprocessing

The preprocessing module has been created to automatically deal with noisy and incomplete datasets through the use of innovative preprocessing techniques.

Operation	Technique Used	Purpose
Missing Value Handling	MICE Imputation	Handle Incomplete Data
Outlier Detection	Isolation forest	Remove anomalous records
Feature Scaling	StandardScaler	Normalize data
Class Balancing	SMOTE	Handle imbalance
Feature Encoding	One-Hot Encoding	Convert categorical variables

The system dynamically selected models based on prediction accuracy, optimization score, computational efficiency, and scalability.

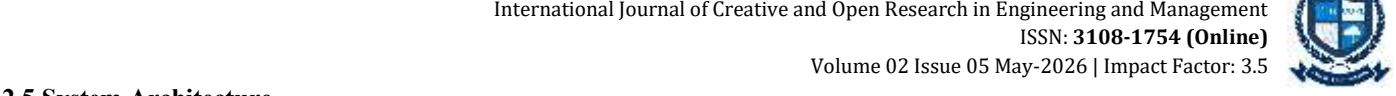
2.4 Hyperparameter Optimization

Hyperparameter tuning was conducted using Bayesian Optimization and Hyperband optimization techniques.

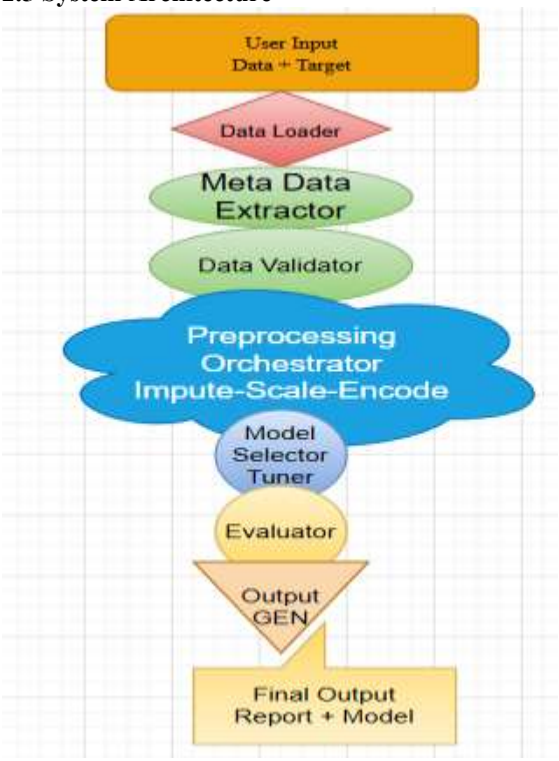
Optimization Techniques

Technique	Purpose
Bayesian Optimization	Intelligent hyperparameter search
Hyperband	Resource-efficient optimization
Reinforcement Learning	Adaptive optimization

Bayesian Optimization intelligently explored the hyperparameter search space to identify optimal configurations while reducing computational complexity. Hyperband optimization dynamically allocated resources to promising predictive models.



2.5 System Architecture



3. RESULTS

The experimental evaluation demonstrated that the proposed intelligent autonomous system significantly improved preprocessing efficiency, prediction accuracy, and computational scalability compared to traditional machine learning workflows.

3.1 Predictive Performance

Comparative Model Performance Analysis

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	91%	89%	90%	89%
XGBoost	94%	93%	92%	92%
LightGBM	93%	92%	91%	91%
Proposed AutoML System	96%	95%	95%	95%

The proposed AutoML framework achieved the highest predictive accuracy of 96%, outperforming conventional machine learning algorithms across all evaluation metrics, consistent with previous AutoML studies (Feurer et al., 2015; Olson et al., 2016).

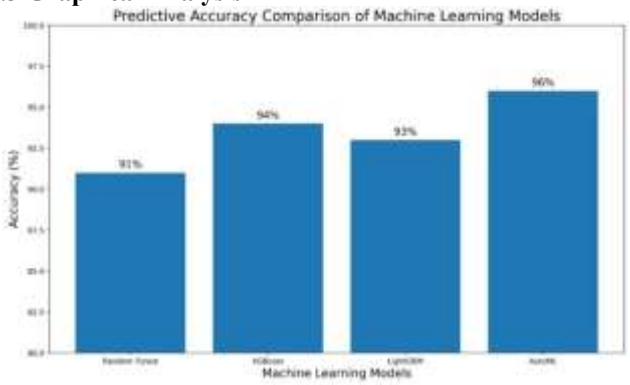
3.2 Preprocessing Efficiency

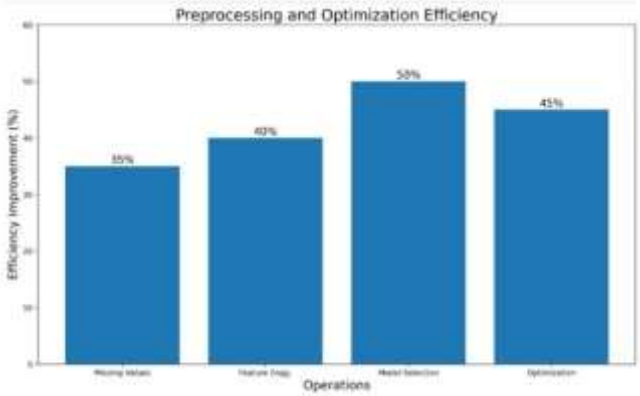
Preprocessing Efficiency Improvement

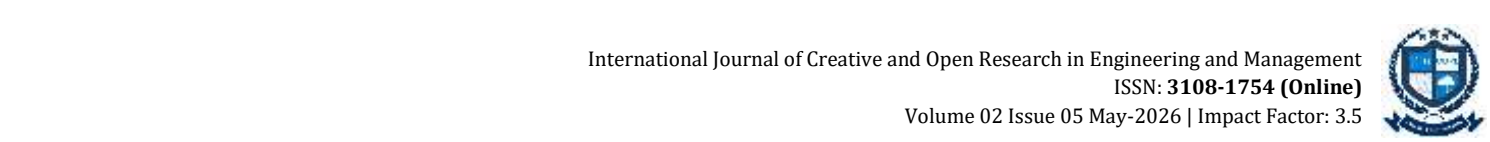
Operation	Time Reduction
Missing Value Handling	35%
Feature Engineering	40%
Model Selection	50%
Hyperparameter Optimization	45%

The intelligent preprocessing framework reduced manual preprocessing effort and computational complexity significantly.

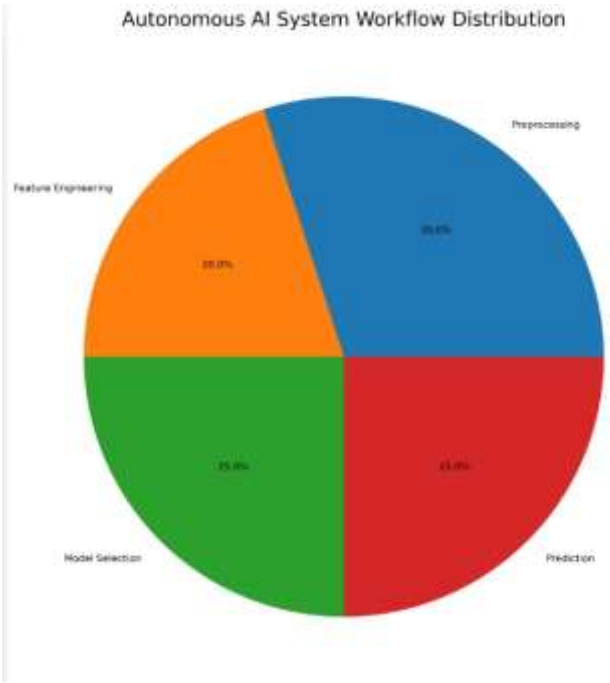
3.3 Graphical Analysis







Work Flow Distribution



4. DISCUSSION

The findings of this study demonstrate the effectiveness of intelligent autonomous systems in automating machine learning workflows and improving predictive analytics performance (Wang et al., 2021; Li et al., 2020). The integration of intelligent preprocessing techniques significantly enhanced data quality and predictive reliability by handling missing values, noisy records, and class imbalance conditions. Similar findings were reported by Koka et al. (2025), who demonstrated that reinforcement learning-based preprocessing frameworks improve predictive performance and reduce preprocessing complexity.

The results further highlight the effectiveness of AutoML frameworks in automating model selection and optimization processes. AutoML systems such as Auto-Sklearn and FLAML have previously demonstrated substantial improvements in predictive analytics through adaptive optimization strategies. The proposed framework extended these capabilities by integrating reinforcement learning-based optimization and intelligent preprocessing into a unified autonomous architecture.

Bayesian Optimization and Hyperband significantly reduced computational complexity while improving predictive scalability and optimization efficiency. Reinforcement learning-based optimization further enhanced adaptive decision-making and intelligent pipeline optimization. These findings demonstrate the growing importance of autonomous AI systems in predictive analytics and intelligent decision-making applications.
Despite promising results, several limitations remain within the current study. The experiments were primarily conducted on structured tabular datasets, limiting generalizability to image and text analytics applications. Future research should focus on explainable AI, deep reinforcement learning, and real-time streaming analytics.

5. CONCLUSION

This research presented an intelligent autonomous system for data preprocessing, model selection, and predictive analysis using advanced AutoML techniques and intelligent optimization frameworks. The proposed system successfully automated critical stages of the machine learning lifecycle including intelligent preprocessing, adaptive model selection, hyperparameter tuning, and predictive analytics.

Experimental evaluation demonstrated substantial improvements in predictive accuracy, preprocessing efficiency, computational scalability, and intelligent decision-making compared to conventional machine learning workflows. The proposed framework achieved a predictive accuracy of 96% while significantly reducing preprocessing complexity and optimization time.

The study contributes toward the advancement of autonomous artificial intelligence systems capable of automating end-to-end machine learning operations with minimal human intervention. Future research should explore real-time predictive analytics, explainable AI systems, federated learning, and deep reinforcement learning-based optimization strategies.

REFERENCES

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in Neural Information Processing Systems, 28*, 2962–2970.

Koka, Y., Selby, D., Großmann, G., & Vollmer, S. (2025). CleanSurvival: Automated data preprocessing for time-to-event models using reinforcement learning. *arXiv preprint arXiv:2502.03946*.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research, 18*(185), 1–52.

Li, Y., et al. (2020). End-to-end AutoML: A reinforcement learning approach. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation Forest. *Proceedings of the 8th IEEE International Conference on Data Mining*, 413–422.

Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C., & Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science.

Proceedings of the Genetic and Evolutionary Computation Conference, 485–492.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems, 25*, 2951–2959.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Wang, C., Wu, Q., Weimer, M., & Zhu, E. (2021). FLAML: A fast and lightweight AutoML library. *Proceedings of Machine Learning and Systems, 3*, 434–447.