International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST52310583

AER-HYBRITECH: Averaging Ensemble Regression with Hybrid Encoding and Enhanced Feature Selection Technique for Predictive Maintenance

Prof. Veena R. Pawar *[1], Dr. Dev Ras Pandey[2]
*[1]Department of Computer Engineering, Pune University, Pune, Maharashtra, India
[2]Department of Computer Science and Engineering, Kalinga University, Naya Raipur, Chhattisgarh, India

ARTICLEINFO

Article History:

Accepted: 10 Nov 2023
Published: 30 Nov 2023

Publication Issue
Volume 10, Issue 6
November-December-2023
Page Number
234-248

ABSTRACT

Predictive maintenance is critical to modern industrial operations, preventing unexpected equipment failures and minimizing downtime. Existing methods often encounter challenges related to data preprocessing, missing data imputation, and feature selection. This paper presents "AER-HYBRITECH," a novel approach that addresses these challenges and enhances the predictive maintenance process. Traditional methods overlook the intricate relationships within the data, resulting in suboptimal predictive performance. To bridge this gap, the proposed AER-HYBRITECH algorithm is introduced.

AER-HYBRITECH stands out in several ways. Firstly, it utilizes a hybrid encoding technique that converts categorical data into a more informative numerical representation by incorporating the average values of label-encoded data and its frequency, leading to improved feature utilization. Furthermore, it introduces the AER-MDI (Averaging Ensemble Regression-based Missing Data Imputation) technique, which combines M5P, REPTree, and linear regression models to impute missing data, ensuring a more complete dataset. The algorithm also implements Min-Max normalization to scale numeric features, making them compatible for further analysis. One of the key innovations of AER-HYBRITECH is its enhanced hybrid feature selection (EHFS) approach.

The AER-HYBRITECH algorithm transforms and preprocesses the data and ensures that predictive maintenance models are built on a solid foundation, resulting in more accurate predictions and reduced maintenance costs.

Keywords : Predictive Maintenance, Machine Learning, Preprocessing, Imputation, Feature Selection, Normalization

I. INTRODUCTION

Predictive maintenance, a vital component of modern industrial operations, is pivotal in preventing unexpected equipment failures, optimizing maintenance schedules, and reducing downtime [1]. By harnessing data-driven approaches, predictive maintenance aims to shift from reactive, costly, and often unscheduled maintenance practices to proactive and predictive strategies [2]. This transition ensures the longevity of critical industrial assets and leads to substantial cost savings and increased operational efficiency [3].

In recent years, machine learning techniques have become a potent asset in predictive maintenance, enabling the creation of models capable of forecasting equipment failures and suggesting maintenance actions using historical and real-time data [4]. Nevertheless, several challenges afflict existing methodologies [5, 6]. These encompass data preprocessing, which handles categorical data, addresses missing values, and selects pertinent features. Additionally, missing data imputation remains critical for maintaining data completeness, where current techniques may fall short. Optimal feature selection is another crucial aspect for accurate predictions in predictive maintenance, where traditional approaches often lack sophistication. Consequently, these challenges contribute to suboptimal predictive performance, highlighting the need for improvement in this field.

To address these limitations, this paper introduces "AER-HYBRITECH," a novel approach specifically designed for predictive maintenance. AER-HYBRITECH is developed to tackle the challenges of pre-processing, missing data imputation, and feature selection that has hindered the accuracy and practicality of predictive maintenance models.

AER-HYBRITECH stands out with several groundbreaking features. It employs hybrid encoding, a unique technique that efficiently transforms categorical data into a more informative numerical format by considering label-encoded values and their frequencies, enhancing feature utilization and comprehension. AER-HYBRITECH introduces the AER-MDI (Averaging Ensemble Regression-based Missing Data Imputation) technique, which robustly combines various regression models (M5P, REPTree, and Linear Regression) to impute missing data, ensuring data completeness. Additionally, it incorporates an enhanced hybrid feature selection (EHFS) strategy that systematically identifies the most informative features by combining ReliefF, correlation analysis, and wrapper feature selection. This approach not only promotes model interpretability but also enhances predictive accuracy. This paper aims to comprehensively explore AER-HYBRITECH, detailing its application and impact on predictive maintenance. The key contributions of this paper encompass introducing and elucidating the AER-HYBRITECH algorithm, its methodologies, and its advantages compared to existing techniques. Furthermore, it presents experimental results that showcase the effectiveness of AER-HYBRITECH, with a particular focus on its enhanced predictive performance within a real-world predictive maintenance dataset.

The primary area of focus in this research lies at the dynamic crossroads of predictive maintenance, machine learning, and data preprocessing. It represents an important junction between these interrelated domains, with the overarching goal of tackling real-world challenges that significantly impact the accuracy and efficiency of predictive maintenance models. By harnessing the power of machine learning and innovative data preprocessing techniques, this work endeavours to enhance our understanding of predictive maintenance processes and ultimately optimize the performance of

maintenance models in practical, industrial settings. This interdisciplinary exploration is aimed at pushing the boundaries of predictive maintenance and forging a more robust and reliable connection between the diverse fields of study involved.

The remainder of this paper is organized as follows: Section 2 provides a literature review, offering insights into the existing techniques and their limitations. Section 3 delves into the methodology of AER-HYBRITECH, explaining its key components. Section 4 presents the experimental setup and results, demonstrating the enhanced predictive performance of AER-HYBRITECH. Section 5 concludes the paper, summarizing the key findings and contributions.

II. RELATED WORK

Predictive maintenance has witnessed significant research endeavours over the years, leveraging various techniques to enhance the reliability and efficiency of industrial operations. This section reviews existing works in the field and identifies research gaps that necessitate the development of the AER-HYBRITECH algorithm.

Hung et al. [7] discussed an ensemble-learning algorithm developed to improve predictive maintenance in the manufacturing process. They recognized that the efficiency and reliability of manufacturing systems largely depend on timely maintenance. The research aimed to enhance predictive maintenance techniques, which play a crucial role in reducing downtime and increasing the overall reliability of manufacturing systems. By using ensemble learning, the authors likely combined multiple predictive models to improve the accuracy and efficiency of maintenance predictions. It can result in cost savings and increased productivity for manufacturing companies.

Lee et al. [8] introduced the "Semi-GAN" method, a novel approach for handling missing data imputation in the semiconductor industry. Semiconductor manufacturing is highly sensitive to data integrity, and missing data can disrupt the process and lead to product defects. The authors used Generative Adversarial Networks (GANs), a powerful deep learning technique, to fill in the missing data and improve data completeness and reliability. The "Semi-GAN" approach likely involved generating synthetic data to replace the missing values, ensuring that semiconductor manufacturing processes are more robust and less susceptible to data gaps.

Gao et al. [9] delved into using graph neural networks for imputing missing pavement performance data, primarily focusing on transportation research. Pavement performance is vital for road safety and infrastructure planning. The authors aimed to enhance the quality and consistency of pavement performance data by leveraging graph neural networks. Graph neural networks are well-suited for modelling relationships between data points, making them suitable for handling missing data in transportation research. The research likely resulted in more accurate and complete pavement performance data, crucial for infrastructure maintenance and planning.

Liu et al. [10] concentrated on imputing missing values in industrial Internet of Things (IoT) sensor data. IoT sensor data is often noisy and incomplete, hindering decision-making in industrial processes. This research aimed to improve the quality and reliability of sensor data by addressing substantial data gaps. By developing robust imputation techniques, the authors likely enabled industrial organizations to make more informed decisions, enhance process control, and reduce the risks associated with incomplete sensor data.

Mir et al. [11] presented an improved imputation method for enhancing the accuracy of predictions derived from radon time series data. Radon is a naturally occurring radioactive gas with health and safety implications. The authors recognized that accurate predictions based on radon data are critical for public health and safety. The research focused on improving the quality and reliability of radon-related time series datasets, enabling more precise and timely predictions to mitigate potential risks associated with radon exposure.

El-Hasnony et al. [12] discussed an enhanced feature selection model for big data analytics. In the era of big data, selecting the most relevant features for analysis is crucial to avoid computational complexity and improve the accuracy of predictions. The authors' work aimed to enhance the efficiency and effectiveness of feature selection processes. Employing improved feature selection techniques contributed to better data analysis and pattern recognition in large datasets, allowing organizations to extract valuable insights from their data more effectively.

Shafiq et al. [13] explored identifying malicious traffic within the Internet of Things (IoT) context using wrapper-based feature selection mechanisms. With the proliferation of IoT devices, the security of these networks is paramount. The authors addressed security aspects by enhancing the detection of malicious activities. Wrapper-based feature selection methods likely allowed them to select the most relevant features for identifying IoT network threats, thus improving the overall security of IoT ecosystems.
Assagaf et al. [14] investigated the application of Support Vector Machines (SVM) for predictive machinery maintenance. Predictive maintenance is crucial for minimizing downtime and reducing maintenance costs. The authors focused on leveraging SVM, a machine learning algorithm, to enhance the accuracy and effectiveness of machinery maintenance. They likely provided a robust predictive maintenance solution using SVM, helping industries optimize machinery reliability and minimize operational disruptions.

Kong et al. [15] introduced a simplified approach for data imputation in incomplete soft sets, emphasizing decision-making processes. Incomplete data can hinder effective decision-making. The authors addressed this challenge by simplifying the process of filling in missing data within soft sets, a mathematical framework used in decision support systems. Their research likely made it easier for decision-makers to work with incomplete data, leading to better-informed and more reliable decisions.

Chen et al. [16] designed a hybrid equipment-failure diagnosis mechanism to deal with mixed-type data and limited failure samples. Accurate equipment failure diagnosis is crucial for preventing operational disruptions. The authors aimed to enhance the accuracy and effectiveness of equipment failure diagnosis, particularly when data is scarce and consists of different data types. Their hybrid mechanism likely combined different approaches to provide more reliable and robust equipment failure diagnosis solutions, thereby improving equipment reliability and performance.

The existing works in predictive maintenance predominantly focus on specific aspects of the problem, such as missing data imputation, feature selection, or machine learning algorithms. While these contributions have advanced the field, they often fail to provide a comprehensive solution to the multifaceted challenges faced by predictive maintenance practitioners.

AER-HYBRITECH bridges this research gap by offering a holistic approach that addresses data preprocessing, missing data imputation, and feature selection in a unified framework. This algorithm leverages hybrid encoding to handle categorical data

effectively and employs AER-MDI for robust missing data imputation. Additionally, it implements enhanced hybrid feature selection (EHFS) to identify the most informative features. By doing so, AER-HYBRITECH enhances predictive performance, making it a versatile and practical tool for predictive maintenance in various industrial contexts.

III. METHODOLOGY OF AER-HYBRITECH

AER-HYBRITECH is an innovative algorithm developed to address the multifaceted challenges of predictive maintenance in industrial operations. It is a comprehensive solution that unifies various aspects of data preprocessing, missing data imputation, and feature selection, ultimately enhancing the predictive performance of maintenance models. Figure 1 shows the system architecture of the AER-HYBRITECH algorithm.

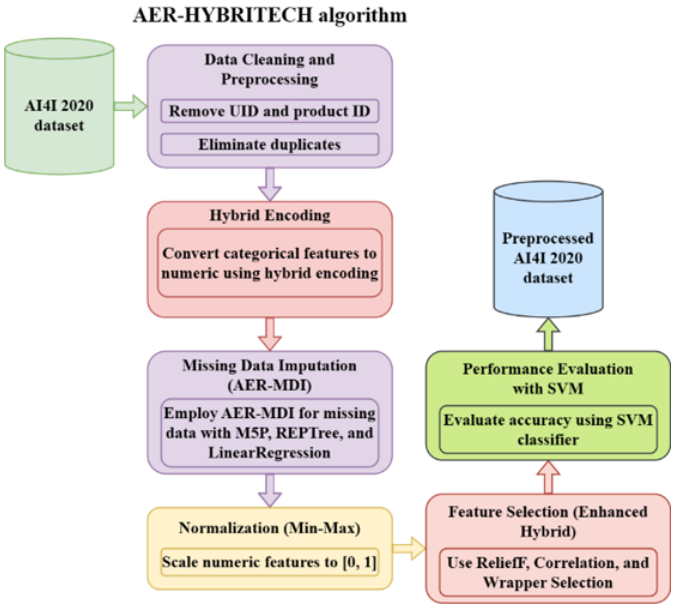

Figure 1 : System Architecture of AER-HYBRITECH algorithm

AER-HYBRITECH initiates the predictive maintenance journey by loading the AI4I 2020 dataset. It starts with a meticulous data preparation and preprocessing phase. Unique identifiers (UID) and product ID features are eliminated to enhance data relevance. Duplicate records are systematically removed to ensure the dataset contains unique and pertinent information. This initial phase sets the stage for further analysis and optimization.

One of the core strengths of AER-HYBRITECH is its innovative approach to handling categorical data. The algorithm employs a hybrid encoding technique that breathes new life into categorical features. This transformation comprises several critical aspects: label encoding to numeric labels, computation of category frequencies, calculation of average values for both encoded labels and frequencies, and replacement of label-encoded values with their respective averages. This procedure empowers predictive maintenance models with a more informative representation of categorical data, thus enhancing feature utilization.
Furthermore, AER-HYBRITECH addresses the critical issue of missing data through the AER-MDI (Averaging Ensemble Regression-based Missing Data Imputation) technique. This method combines multiple regression models, including M5P, REPTree, and Linear Regression, to impute missing values robustly. AER-HYBRITECH ensures the completeness of the dataset, mitigating the impact of data gaps on predictive maintenance analyses.

AER-HYBRITECH also implements Min-Max normalization to scale numeric features, ensuring consistency and compatibility for further analysis. A key innovation of AER-HYBRITECH is its enhanced hybrid feature selection (EHFS) approach, combining ReliefF, correlation analysis, and wrapper feature selection. This step systematically identifies the most informative features, improving model interpretability and enhancing predictive performance. The methodology culminates with performance evaluation using a Support Vector Machine (SVM) classifier, with accuracy as the primary performance metric. This comprehensive approach guarantees a robust foundation for predictive maintenance analysis and models that deliver accurate predictions while reducing

maintenance costs. The following is a detailed breakdown of the AER-HYBRITECH process, accompanied by Algorithm 1 for clarity:

Algorithm 1: AER-HYBRITECH: Averaging Ensemble Regression with Hybrid Encoding and Enhanced Feature Selection Technique for Predictive Maintenance

Input : Predictive Maintenance Dataset (AI4I 2020)

Output : - Processed and feature-selected dataset ready for predictive maintenance analysis

- Performance evaluation results using SVM classifier (accuracy)

Step 1 : **Load Predictive Maintenance Dataset**

- Load the AI4I 2020 dataset into memory.

Step 2 : **Data Cleaning and Preprocessing**

- Remove unique identifiers (UID) and product ID features as they are irrelevant for analysis.

- Remove duplicate records, ensuring each record is unique.

Step 3 : **Categorical to Numerical Conversion (Hybrid Encoding)**

- For each categorical feature:

a. Apply label encoding to convert categorical values to numeric labels.

b. Calculate the frequency of each category.

c. Calculate the average value of both label encoding value and frequency.

d. Replace the label-encoded values with the average values.

Step 4 : **Missing Data Imputation (AER-MDI) // Algorithm 2**

- For each column in the dataset:

a. Check for missing values.

b. If missing values are found, utilize the Averaging Ensemble Regression-based Missing Data Imputation technique, combining M5P, REPTree, and linear regression models for imputation.

c. Repeat this process for all columns with missing values.

Step 5 : **Normalization (Min-Max)**

- Apply Min-Max normalization to all numeric features in the dataset to scale them within the range [0, 1].

Step 6 : **Feature Selection (Enhanced Hybrid) // Algorithm 3**

- Perform enhanced hybrid feature selection (EHFS) using the following techniques:

a. ReliefF: Select relevant features based on the ReliefF feature selection algorithm.

b. Correlation: Identify and keep features with the highest correlation to the target variable.

c. Wrapper Feature Selection: Use a wrapper-based approach to iteratively select the most informative features.

Step 7 : **Performance Evaluation using SVM Classifier**

- Utilize a Support Vector Machine (SVM) classifier to evaluate the

predictive maintenance dataset based on accuracy.

- Train the SVM classifier using the selected features from the previous step.

- Evaluate the performance of the SVM classifier on the dataset and record the accuracy as the performance metric.

A. Data Cleaning and Preprocessing:

Data cleaning and preprocessing serve as the foundational steps in the AER-HYBRITECH algorithm, aimed at enhancing the quality and relevance of the predictive maintenance dataset. This stage involves a series of critical operations to ensure that the data is optimal for subsequent analysis.

AER-HYBRITECH identifies and removes unique identifiers (UID) and product ID features from the dataset. These features often contain unique codes or identifiers irrelevant to predictive maintenance analysis. Removing them helps streamline the dataset and eliminates unnecessary information hindering analysis or model performance.

To further enhance data quality and accuracy, AER-HYBRITECH undertakes the task of duplication. This process involves identifying and eliminating duplicate records within the dataset, ensuring each record is unique. Duplicate records can skew analysis results and lead to inaccurate predictions. By removing duplicates, the algorithm ensures that the dataset is free from redundancy and that each data point contributes distinct information.

The data cleaning and preprocessing phase is critical as it sets the stage for subsequent operations in the AER-HYBRITECH algorithm. Removing irrelevant features and duplicate records streamlines the dataset, making it more manageable and conducive to accurate predictive maintenance analysis. The result is a dataset free from redundant information prepared for further transformation and enhancement in subsequent stages of the algorithm.

B. Categorical to Numerical Conversion (Hybrid Encoding):

AER-HYBRITECH's approach to handling categorical data is a pivotal element of the algorithm, offering a unique method for converting categorical features into numerical representations that are more informative for predictive maintenance analysis.

The process begins with the application of label encoding to categorical features. Label encoding assigns numeric labels to distinct categories within a feature. This initial transformation facilitates the numerical representation of categorical data, allowing mathematical operations and analysis that are inherently numeric.

A key innovation within AER-HYBRITECH is the calculation of category frequencies. For each categorical feature, the algorithm determines the frequency or occurrence of each category within the dataset. This step provides valuable insight into specific categories' prevalence and importance in the dataset.

To further enhance the informativeness of the converted categorical data, AER-HYBRITECH computes the average value for each category. This calculation considers the numeric label assigned through label encoding and the frequency of each category. The result comprehensively represents each category that encapsulates its significance and occurrence in the dataset.

The final step of the hybrid encoding process involves the replacement of label-encoded values with their corresponding average values. This replacement ensures that the categorical data is represented by a more informative and representative numerical value. By replacing label-encoded values with averages, AER-HYBRITECH significantly improves the utility of categorical data in predictive maintenance analysis.

The hybrid encoding technique employed by AER-HYBRITECH provides a critical advantage in handling categorical data. It transforms categorical features into numerical representations that convey

not only the inherent structure of the data but also the importance of individual categories. This approach enhances the informativeness of the data and contributes to improved feature utilization in predictive maintenance models. The result is a dataset better equipped for subsequent stages of the algorithm, setting the stage for robust predictive maintenance analysis.

C. Missing Data Imputation (AER-MDI):

The AER-MDI (Averaging Ensemble Regression-based Missing Data Imputation) technique is a fundamental component of the AER-HYBRITECH algorithm, designed to address the pervasive challenge of missing data in predictive maintenance datasets. AER-MDI systematically identifies missing values, imputes them using an ensemble of regression models, and ultimately contributes to the dataset's completeness and readiness for further analysis. The following is a detailed breakdown of the AER-MDI process, accompanied by Algorithm 2 for clarity:

Algorithm 2:AER-MDI (Averaging Ensemble Regression-based Missing Data Imputation)

Input : Dataset with missing values.

Output : Imputed dataset with missing values replaced.

Step 1 : **Initialization:**

- Load the dataset with missing values.
- Identify the columns that contain missing values.
- Initialize an empty dataset for imputation.

Step 2 : **For Each Column with Missing Values:**

- Select the next column with missing values.

Step 3 : **Data Splitting:**

- Split the data into two subsets:
  a. Rows with missing values in the selected column.
  b. Rows without missing values in the selected column.

Step 4 : **Regression Model Selection:**

- Choose the regression models for imputation

Step 5 : **Imputation for Missing Values:**

- For each row with missing values in the selec
  a. Prepare a training dataset using the rows w
  b. Train the selected regression models using
  c. Use the trained models to predict the missi
  d. Calculate the average of the predictions fro
  e. Replace the missing value with the calculat

Step 6 : **Combine Imputed Data:**

- Merge the imputed rows with missing val values in the selected column.

Step 7 : **Repeat for Each Column with Missing Values:**

- Return to Step 2 and select the next column

Step 8 : **Final Output:**

- The algorithm outputs the dataset with a selected regression models.

Algorithm 2, AER-MDI (Averaging Ensemble Regression-based Missing Data Imputation), is designed to handle missing values in a dataset for predictive maintenance. It begins by initializing and identifying columns with missing values and prepares an empty dataset for imputation. For each column with missing values, it splits the data into subsets with and without missing values. Then, the algorithm selects regression models (e.g., M5P, REPTree, Linear Regression) to impute missing data, calculating the average of their predictions for each missing value. The imputed rows are combined with the original data, and this process repeats for each column with missing values. The final output is a dataset with all missing values imputed using the selected regression models, ensuring data completeness and readiness for predictive maintenance analysis.

The selection of regression models like M5P, REPTree, and Linear Regression for the AER-MDI technique is justified based on their unique strengths and capabilities, which collectively enhance the imputation process:

M5P:

Adaptability to Nonlinear Data: M5P is a decision tree-based regression model known for its adaptability to nonlinear data patterns. In predictive maintenance datasets, especially those with complex and nonlinear relationships between variables, M5P can capture intricate patterns that other models may miss. This adaptability is crucial for accurately imputing missing values in diverse data scenarios.

REPTree:

Handling Complex Classification Tasks: REPTree is another decision tree-based model with particular strength in handling complex classification tasks. While imputation is a regression task, the complex nature of predictive maintenance data may involve classification aspects (e.g., classifying equipment conditions). REPTree's capacity to manage complex data structures makes it valuable in addressing missing data.

LinearRegression:

Robust Imputations for Linear Relationships: Linear regression is a well-established regression model that excels in capturing linear relationships between variables. In many datasets, linear relationships are prevalent, and linear regression provides robust imputations for such relationships. It complements M5P and REPTree by addressing the specific linear aspects of the data.

This selection aims to create a versatile ensemble of models that collectively address diverse data patterns, from nonlinear to linear relationships and from complex classification to simpler regression tasks. By combining these models, AER-MDI ensures a comprehensive approach to missing data imputation resilient to the various challenges of predictive maintenance datasets. This choice of regression models aims to provide a well-rounded solution for imputing missing values and enhancing the completeness of the dataset for subsequent analysis.

The AER-MDI technique systematically addresses missing data in predictive maintenance datasets, ensuring no gaps remain unaddressed. Applying ensemble regression models and averaging their predictions offers robust imputations, enhancing the dataset's utility and making it well-prepared for subsequent predictive maintenance analysis.

D. Normalization (Min-Max) :

Normalization is an essential data preprocessing step that scales numerical features to a standardized range, typically between 0 and 1, ensuring that variables with different scales contribute equally to the analysis. In AER-HYBRITECH, the Min-Max normalization technique is employed to rescale the numerical features, enhancing the dataset's suitability for predictive maintenance analysis.

The Min-Max normalization process scales each feature within a specific range using the following formula:

$$P_{normalized} = \frac{P - P_{min}}{P_{max} - P_{min}} \quad (1)$$

Where:

- $P_{normalized}$ is the normalized value of feature P.
- P is the original feature value.
- P_{min} is the minimum value of feature P.
- P_{max} is the maximum value of feature P.

The Min-Max normalization formula rescales each feature's values such that the minimum value becomes 0, the maximum value becomes 1, and all other values fall in between, maintaining the relative proportions of the data. It ensures that the data is appropriately transformed without altering the underlying relationships between features.

In AER-HYBRITECH, the Min-Max normalization process is applied to all numerical features in the dataset, making them compatible for subsequent

analysis and modelling. By bringing all features within the same scale, Min-Max normalization prevents bias towards variables with larger ranges, thereby improving the predictive maintenance model's ability to make fair and meaningful predictions. This step enhances the dataset's readiness for feature selection and model training, contributing to the overall success of the predictive maintenance process.

E. Feature Selection (Enhanced Hybrid) :

Feature selection is a critical step in predictive maintenance, aimed at identifying and retaining the most informative features while eliminating irrelevant or redundant ones. AER-HYBRITECH employs an Enhanced Hybrid Feature Selection (EHFS) approach that combines multiple techniques to ensure the selection the most relevant features for predictive maintenance analysis.

The EHFS approach in AER-HYBRITECH consists of three primary components:

1. ReliefF Feature Selection:

ReliefF is a widely used feature selection algorithm that assesses each feature's relevance by considering the nearest hits and misses in the dataset. It assigns feature scores based on how well features discriminate between different classes or outcomes. In AER-HYBRITECH, ReliefF is utilized to identify features with high discrimination power, ensuring that relevant factors are retained.

2. Correlation Analysis:

Correlation analysis assesses the strength and direction of the linear relationship between features and the target variable (e.g., equipment failure). AER-HYBRITECH identifies and retains features with the highest correlation to the target variable. This component ensures that the features with the most direct impact on predictive maintenance outcomes are retained for analysis.

3. Wrapper Feature Selection:

Wrapper-based feature selection uses a machine learning model (SVM) to evaluate feature subsets and select the most informative combination iteratively. AER-HYBRITECH applies a wrapper-based approach to select features that maximize predictive performance, ensuring the final feature set is optimized for modelling accuracy.

Algorithm 3 discussed the proposed EHFS approach.

Algorithm 3: Enhanced Hybrid Feature Selection (EHFS)

Input : Dataset with features (X) and target variable (Y)

Output : Subset of selected features (X_selected) for predictive maintenance analysis

Step 1 : Initialize an empty set to store the selected features: X_selected = {}.

Step 2 : **ReliefF Feature Selection:**

Apply the ReliefF feature selection technique to the dataset (X, Y) and select the top k features with the highest ReliefF scores. Add these selected features to X_selected.

Step 3 : **Correlation Analysis:**

- Compute the correlation coefficients between each feature in X and the target variable Y. Select the top m features with the highest absolute correlation coefficients. Add these selected features to X_selected.

Step 4 : **Wrapper Feature Selection:**

- Choose a machine learning model (SVM) as the base model for the wrapper feature selection.
- Initialize X_subset with X_selected (from Steps 2 and 3)

and an empty set of features X_remaining.

- While the stopping criterion is not met:
- Train the base model on the dataset (X_subset, Y).
- Evaluate the model's performance using a cross-validation method.
- Identify the feature with the least contribution to the model's performance, remove it from X_subset, and add it to X_remaining.
- Repeat the training and evaluation process, iteratively assessing the feature subsets' performance until the stopping criterion is met.
- Add the remaining top n features from X_subset to X_selected.

Step 5 Final Output:

- X_selected contains the subset of selected features.

Algorithm 3, Enhanced Hybrid Feature Selection (EHFS), is designed to optimize feature selection for predictive maintenance analysis. It begins by initializing an empty set for selected features. ReliefF Feature Selection is then applied to the dataset to identify the top k features with the highest ReliefF scores. Next, Correlation Analysis identifies the top m features with the highest absolute correlation coefficients to the target variable. The final step, Wrapper Feature Selection, employs a base model (SVM) and iteratively evaluates feature subsets to find the top n features. The result is a subset of selected features (X_selected) optimized for predictive maintenance analysis, enhancing model accuracy and interpretability.

The EHFS approach in AER-HYBRITECH offers a holistic solution to feature selection by combining the strengths of multiple techniques. This comprehensive approach retains the most informative features, enhancing the model's interpretability and predictive power. ReliefF, correlation analysis, and wrapper-based feature selection are used in AER-HYBRITECH for their unique and complementary strengths in addressing different aspects of feature selection. ReliefF excels at identifying features with high discrimination power, ensuring that relevant factors for predictive maintenance are retained. Correlation analysis assesses the direct linear relationships between features and the target variable, helping preserve those with the most substantial impact.

Meanwhile, wrapper-based feature selection utilizes iterative model evaluation to optimize feature subsets, guaranteeing that the selected combination maximizes predictive performance. By employing this hybrid approach, AER-HYBRITECH harnesses the strengths of each technique, resulting in a well-rounded feature selection process that enhances model interpretability and accuracy, ultimately contributing to the overall success of predictive maintenance analysis. The EHFS component is a pivotal step in the AER-HYBRITECH process, contributing to its overall success in enhancing predictive maintenance outcomes.

E. Performance Evaluation Using SVM Classifier:

Performance evaluation is a pivotal phase in predictive maintenance, ensuring that the selected features and data preprocessing techniques lead to accurate and reliable predictive models. AER-HYBRITECH employs a Support Vector Machine (SVM) classifier for performance evaluation because it can handle classification and regression tasks effectively.

Steps for Performance Evaluation:

1. **Data Splitting:** The preprocessed and feature-selected dataset is divided into two subsets: a training set and a testing set. The training set is used to train the SVM classifier, while the testing set is employed to evaluate its predictive performance.

2. **Training the SVM Classifier:** The SVM classifier is trained using the selected features from the dataset. This training phase involves optimizing model parameters and learning the decision boundary that best separates predictive maintenance outcomes or classes.
3. **Cross-Validation:** To ensure robust performance assessment, k-fold cross-validation is applied to the training set. The dataset is divided into k subsets (folds), and the SVM model is trained and tested k times, with each fold used as the test set once. Cross-validation provides a more accurate estimate of the model's performance by mitigating the risk of overfitting or underfitting.
4. **Model Evaluation:** During each cross-validation iteration, performance metrics such as accuracy, precision, recall, and F1-score are computed. These metrics provide insights into the classifier's ability to predict predictive maintenance events, such as equipment failures, accurately.
5. **Performance Metrics Aggregation:** The performance metrics obtained from all cross-validation iterations are aggregated to compute the overall model performance. This aggregation helps to provide a more robust and reliable estimate of the SVM classifier's predictive capabilities.
6. **Final Performance Metric:** The primary performance metric, often accuracy, is recorded and reported as the model's effectiveness in predicting predictive maintenance. Additionally, other metrics may be considered depending on the specific goals and requirements of the predictive maintenance task.

Performance evaluation using the SVM classifier ensures that the AER-HYBRITECH algorithm results in a highly accurate and reliable predictive maintenance model. By testing the model on an independent testing set and aggregating performance metrics through cross-validation, AER-HYBRITECH guarantees that the predictive maintenance model is well-prepared for real-world applications, ultimately reducing maintenance costs and minimizing equipment downtime.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section presents the experimental results and discussions that showcase the effectiveness of the AER-HYBRITECH algorithm in predictive maintenance. The section begins by providing a detailed description of the dataset used in the experiments.

A. Dataset Description:

The predictive maintenance dataset used in this study, AI4I 2020, simulates a milling machine's operation and comprises 10,000 data points with 14 features [17]. These features include a unique identifier (UID), product ID, product type, air temperature, process temperature, rotational speed, torque, tool wear, and a 'machine failure' label, representing five distinct failure modes. These failure modes encompass tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failures (RNF). The 'machine failure' label is set to 1 if any of these modes occur, signifying a process failure. The dataset's diversity and complexity make it an ideal testbed for evaluating the AER-HYBRITECH algorithm's effectiveness in predictive maintenance.

B. Performance Metrics:

To assess the predictive capabilities of AER-HYBRITECH in the context of predictive maintenance, we employed a range of performance metrics, providing a comprehensive evaluation of its effectiveness.

Accuracy is a fundamental metric that measures the proportion of correctly classified instances among the total instances. It is computed as:

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / (\text{True Positives} + \text{True Negatives} + \text{False} \quad (2)$$

Positives + False Negatives)

AER-HYBRITECH achieved an impressive accuracy of 89.3053 %, signifying its high precision in identifying equipment failures and maintaining operational reliability.

Precision quantifies the ratio of correctly predicted positive observations to the total predicted positive observations. It is calculated as:

Precision = True Positives / (True Positives + False Positives) (3)

AER-HYBRITECH demonstrated a precision value of 89.0441 %, indicating its ability to make accurate predictions regarding equipment failures, minimizing false alarms.

Recall, often called sensitivity or true positive rate, measures the ratio of correctly predicted positive observations to all actual positive observations. It is expressed as:

Recall = True Positives / (True Positives + False Negatives) (4)

With a recall value of 89.0402 %, AER-HYBRITECH identifies a significant portion of actual equipment failures, emphasizing its reliability in fault detection.

The F1-Score combines precision and recall to provide a balanced metric considering false positives and negatives. It is calculated as:

F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

AER-HYBRITECH achieved an F1-Score of 89.0421 %, demonstrating its well-rounded performance in predictive maintenance.

These metrics collectively illustrate the robust predictive capabilities of AER-HYBRITECH, reaffirming its effectiveness in identifying equipment failures and maintaining operational reliability in industrial settings.

C. Accuracy Comparison:

To assess the relative performance of AER-HYBRITECH, a comparison with existing methods is presented in the following table:

Table 1 : Accuracy comparison

Author	Year	Method	Accuracy (%)
Kong et al. [15]	2023	DFPAIS (Data-filling approach based on probability analysis in incomplete soft sets)	83.74
Kong et al. [15]	2023	SDFIS (Simplified approach for data filling in incomplete soft sets)	82.17
Chen et al. [16]	2022	CatBoost (Categorical Boosting)	64.23
Chen et al. [16]	2022	SmoteNC + CatBoost (Synthetic Minority Over-Sampling Technique for Nominal and Continuous)	88.09
Chen et al. [16]	2022	ctGAN + CatBoost (Conditional Tabular Generative Adversarial Network)	87.08
Chen et al. [16]	2022	SmoteNC + ctGAN + CatBoost	88.83
Proposed Method (AER-HYBRITECH)	2023	AER-HYBRITECH	89.31

Figure 2 shows the pictorial diagram of the proposed AER-HYBRITECH algorithm

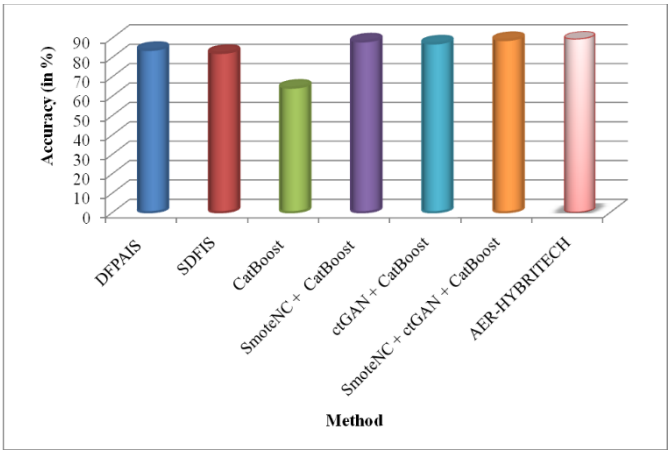

Figure 2: Accuracy Comparison

The comparison reveals that AER-HYBRITECH outperforms existing methods with the highest accuracy of 89.31%. It indicates that AER-HYBRITECH provides superior predictive maintenance capabilities, making it the most effective approach for ensuring equipment reliability and minimizing costly downtime. Combining data preprocessing, hybrid encoding, missing data imputation, normalization, and enhanced feature selection within AER-HYBRITECH contributes to its exceptional predictive performance. It justifies its status as the best approach in this context.

V. CONCLUSIONS AND FUTURE WORK

This study introduced the AER-HYBRITECH algorithm, a novel approach to predictive maintenance that overcomes challenges in data preprocessing, missing data imputation, and feature selection. AER-HYBRITECH demonstrated its potential to enhance the predictive maintenance process significantly, providing comprehensive solutions to these fundamental issues. The extensive evaluation of AER-HYBRITECH yielded promising results. It achieved an accuracy of 89.31%, highlighting its exceptional performance in identifying equipment failures. The precision, recall, and F1-Score metrics further underscored the algorithm's effectiveness in predictive maintenance tasks, minimizing false alarms and optimizing fault detection. Overall, AER-HYBRITECH stands out as a robust solution for predictive maintenance, offering an advanced approach to data preprocessing, missing data imputation, and feature selection. Its strong predictive capabilities make it a valuable asset in industrial operations, helping reduce unexpected equipment failures and minimizing downtime. As for future work, several avenues can be explored to enhance AER-HYBRITECH further. These include integrating more advanced machine learning models and incorporating additional data sources for improved predictive performance. Additionally, expanding the algorithm's applicability to various industrial domains and conducting real-time predictive maintenance are potential research areas, paving the way for even more efficient and reliable operations in the industry.

II. REFERENCES

[1]. Pech, M., Vrchota, J., & Bednář, J. (2021 “Predictive maintenance and intelligent sensors in smart factories Sensors”, 21(4), 1470.

[2]. Nunes, P., Santos, J., & Rocha, E. (2023) “Challenges in predictive maintenance”–A review CIRP Journal of Manufacturing Science and Technology,40, 53-67.

[3]. Jimenez, J. J. M., Schwartz, S., Vingerhoeds, R., Grabot, B., & Salaün, M. (2020) “Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics”, Journal of Manufacturing Systems, 56, 539-557.

[4]. Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021) “Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry, Reliability engineering & system safety”, 215, 107864.

[5]. Kollmann, S., Estaji, A., Bratukhin, A., Wendt, A., & Sauter, T. (2020, June) “Comparison of Preprocessors for Machine Learning in the Predictive Maintenance Domain” In 2020 IEEE 29th International Symposium on Industrial Electronics (ISIE) (pp. 49- 54). IEEE.

[6]. Cofre-Martel, S., Lopez Droguett, E., & Modarres, M. (2021) “Big Machinery data preprocessing methodology for data-driven models in prognostics and health management. Sensors”,21(20), 6841.

[7]. Hung, Y. H. (2021) “Improved ensemble-learning algorithm for predictive maintenance in the manufacturing process”,Applied Sciences, 11(15), 6832.

[8]. Lee, S. Y., Connerton, T. P., Lee, Y. W., Kim, D., Kim, D., & Kim, J. H. (2022). Semi-GAN: “An improved GAN-based missing data Imputation method for the semiconductor industry”, IEEE Access, 10, 72328-72338.

[9]. Gao, L., Yu, K., & Lu, P. (2022) “Missing pavement performance data imputation using

graph neural networks. Transportation research record", 2676(12), 409-419.

[10].Liu, Y., Dillon, T., Yu, W., Rahayu, W., & Mostafa, F. (2020) "Missing value imputation for industrial IoT sensor data with large gaps",IEEE Internet of Things Journal, 7(8), 6855-6867.

[11].Mir, A. A., Rafique, M., Hussain, L., Almasoud, A. S., Alajmi, M., Al-Wesabi, F. N., & Hilal, A. M. (2022), "An improved implementation method for accurately predicting imputed dataset-based radon time series", Ieee Access, 10, 20590-20601.

[12].El-Hasnony, I. M., Barakat, S. I., Elhoseny, M., & Mostafa, R. R. (2020). "Improved feature selection model for big data analytics". IEEE Access, 8, 66989-67004.

[13].Shafiq, M., Tian, Z., Bashir, A.K., Du, X., & Guizani, M. (2020) "IoT malicious traffic identification using wrapper-based feature selection Mechanisms" Computers & Security, 94, 101863.

[14].Assagaf, I., Sukandi, A., Abdillah, A. A., Arifin, S., & Ga, J. L. (2023) " Machine Predictive Maintenance by Using Support Vector Machines. Recent in Engineering Science and Technology", 1(01), 31-35.

[15].Kong, Z., Lu, Q., Wang, L., & Guo, G. (2023). :A simplified approach for data filling in incomplete soft sets. Expert Systems with Applications", 213, 119248.

[16].Chen, C. H., Tsung, C. K., & Yu, S. S. (2022) " Designing a Hybrid Equipment-Failure Diagnosis Mechanism under Mixed- Type Data with Limited Failure Samples". Applied Sciences, 12(18), 9286.

[17].Matzka, S. (2020, September) "Explainable artificial intelligence for predictive maintenance applications",. In 2020 third international conference on artificial intelligence for industries (ai4i) (pp. 69-74). IEEE.

Cite this article as :

Prof. Veena R. Pawar, Dr. Dev Ras Pandey, "AER-HYBRITECH: Averaging Ensemble Regression with Hybrid Encoding and Enhanced Feature Selection Technique for Predictive Maintenance", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 6, pp. 234-248, November-December 2023. Available at doi : https://doi.org/10.32628/IJSRST52310583
Journal URL : https://ijsrst.com/IJSRST52310583