Ridge Regularized Imputed Scaled Clipping Normalization Based Pre-processing for Marine Weather Forecasting

J. Deepa Anbarasi * and V. Radha

***Abstract*—The prediction of marine weather is an application employed for forecasting atmospheric conditions for a given position and time. Data pre-processing is the first step in the marine weather forecasting process. Data mining method is employed for changing data. The purpose of the data pre-processing is to clean and organize the text for an accurate classification process. Also, the pre-processing of large datasets is used to remove the noisy and missing values encountered during the data collection. Many existing data pre-processing methods are studied to improve forecasting performance. However, the conventional methods of space complexity and time complexity during pre-processing were not reduced. In order to address these problems, Ridge Regularized-Imputed-Scaled Clipping Normalization-based Deep Learnt Data Pre-processing (RRISCN-DLDP) Method is introduced. The key objective of the RRISCN-DLDP method is to remove the noisy data and to fill in the missing values in the database for improving the classification performance. RRISCN Method comprises six layers, namely one input layer, four hidden layers, and one output layer for efficient pre-processing. Initially in RRISCN Method, the number of marine weather data points is collected from the database at the input layer. After that, the input marine weather data is transmitted to hidden layer 1. In that layer, Ridge Regularized data quality is assessed through mismatched data types, mixed data values, and data outliers with higher data quality. Then, the missing data values are filled in hidden layer 2 to perform a data cleaning process using Imputed nearest neighbor interpolation through approximating the feature value for a non-given point in corresponding columns with a lesser error rate. Next, the data duplication is removed in the hidden layer 3 by using pointwise animator correlation analysis to execute the data reduction process for measuring the two marine weather data points. Followed by, the data transformation is performed through the scaled clipping normalization process. In this way, efficient data pre-processing is carried out by using RRISCN Method to minimize time and space consumption. Experimental evaluation is performed using various quantitative metrics namely accuracy, space complexity as well as the time involved in pre-processing. The analyzed results reveal the superior performance of our proposed RRISCN Method with higher pre-processing accuracy by 4% as well as lesser space complexity and pre-processing time by 22% and 13% when compared to using conventional techniques.**

***Index Terms*—Marine weather forecasting, data pre-processing, data mining**

I. Introduction

The prediction of marine weather is issued for forecasting atmospheric conditions to a particular location. The prediction of marine weather is an essential process to evaluate the atmospheric situation. A multi-objective grasshopper optimization was developed in [1] without negative constraints. Ensemble empirical mode decomposition was employed for attaining accurate prediction results. However, the designed method of error rate during pre-processing was not reduced.

U-Net based deep learning architecture was designed in [2] to study complex mapping. The designed architecture employed residual connection, parallel convolution, and asymmetric convolution. But, the pre-processing time was not reduced by U-Net-based deep learning architecture. A new error correction system was introduced in [3] to collect data well as improve prediction ability. Quasi-real-time decomposition approach was built for attaining error to every subseries. However, pre-processing accuracy was not improved by the designed system.

Deep Learning-Based Stacked Sparse Autoencoder (DSSAE) was introduced in [4] for forecasting the marine weather condition of a specific area. It was employed for attaining necessary data for enhancing speed. Scalable Parallelizable Induction of Decision Tree (SPRINT) algorithm was introduced in [5] with decision tree principles. Depending on climate parameters, data gets categorized. However, time consumption during pre-processing was not reduced by the SPRINT algorithm. Marine weather Research and Forecasting (WRF) scheme was introduced in [6] for attaining horizontal effects on small-range of marine weather predictions. But, the computational cost was not reduced by WRF model.

A classifier approach was designed in [7] for marine weather prediction. The designed approach used parameters to predict the marine weather after examining the input information in a database. The zero-mean normalization method was designed in [8] with a Particle Swarm Optimizer of Support Vector Machine (PSO-SVM) for selecting the optimal one. The data pre-processing method was chosen for performing the landslide displacement. But, the pre-processing time was not reduced by the PSO-SVM method.

The Machine Learning (ML) method termed Support Vector Regression (SVR), as well as Gradient Boosting Regression Trees (GBRT), was introduced in [9] for increasing the accuracy of wind power. The data investigation method was employed for visualizing data collected over Supervisory Control and Data Acquisition (SCADA). But, the error rate was not minimized by ML-based method. Deep Gaussian Processes (DGP) were introduced in [10] to examine the forecasting scheme. An optimizer depending on geometric alteration was employed for achieving simulated data. However, the computational cost was not reduced by the learning method.

A hybrid model was introduced in [11] with a

Manuscript received September 2, 2022; revised November 11, 2022; accepted January 28, 2023.

J. Deepa Anbarasi and V. Radha are with Department of Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore, India. E-mail: radhasrimail@gmail.com (V.R.)

*Correspondence: deepaanbu899@gmail.com (J.D.A.)

DOI: 10.7763/IJCTE.2023.V15.1333

decomposition module. Raw wind speed series was developed within many subseries. But, the error rate was not reduced by the designed hybrid model.

A hybrid forecasting framework was designed in [12] with Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN) based convolution Bidirectional Long Short-Term Memory (Bi-LSTM) auto encoder to predict the wind speed. But, the forecasting accuracy was not improved by the designed framework.

A combined forecasting system was designed in [13] to perform deterministic and probabilistic forecasts. Data-denoising algorithms were employed to enhance wind speed forecasting. However, the accuracy was not enhanced by a combined forecasting system.

Hybrid ultra-short-term Wind Power Forecasting (WPF) structure was designed in [14] to attain the WPF. Multi-sourced as well as multi-dimensional wind power plant information was pre-processing and Feature Selection (FS) eliminated the unnecessary features. But, the designed method of pre-processing time was not reduced.

The prediction of time series termed as Group Least Square Support Vector Machine (GLSSVM) was designed in [15] to join Least Square Support Vector Machines (LS-SVM) as well as Group Method of Data Handling (GMDH). But, computational complexity was not reduced by GLSSVM.

A forecasting algorithm was introduced in [16] for forecasting the Photovoltaic (PV) power generation with long Short-Term Memory (LSTM) Neural Network (NN). The synthetic prediction of marine weather was carried out to identify PV plant position by combining numerical information with sky forecast. However, the computational cost was not reduced by the forecasting algorithm.

Short-term prediction of wind power method was introduced in [17] through data mining of numerical Marine Weather Prediction (NWP). The short-term weather prediction was based on NWP to contribute to WPF error. But, the pre-processing accuracy was not improved by the designed approach.

An efficient parallelization was carried out [18] through many integrated cores. Intel Many-Integrated Core (MIC) architecture was constructed for achieving the higher-performance of computing. However, the pre-processing time was not reduced by efficient parallelization.

A forecasting framework was introduced in [19] to discover information from NWP with wind as well as solar energy. Gradient boosting tree using feature engineering technique extracted the maximum information from NWP grid. The clearness index was designed in [20] with input data for the LSTM model to increase the prediction accuracy on cloud days. K-means were employed to categorize the marine weather types through data processing. The cloudy days were classified into cloudy and mixed ones. But, the pre-processing accuracy was not minimized by the clearness-index.

Data pre-processing as well as hybrid machines were introduced in [21] for predicting water-level. The designed method of time and error was minimized. However, the accuracy was not enhanced. Yet, another hybrid model was investigated in [22] with higher accuracy for water quality prediction.

Marine weather forecasting is a vital issue to measure the situation of the marine weather for a particular area. Numerous data pre-processing methods were developed to forecast the marine weather data. But, the robustness and accuracy were not superior when considering the vast volume of data. Next, the amount of time and space utilized to pre-process the marine weather data was very higher. In order to overcome the issue, efficient marine weather data pre-processing are required with higher performance of marine weather forecasting. The main objective of the Ridge Regularized-Imputed-Scaled Clipping Normalization-based (RRISCN) Method is to eliminate noisy data and fill in the missing values with minimal minimize time and space. The novelty and contribution of the proposed Ridge Regularized-Imputed-Scaled Clipping Normalization-based Deep Learnt Data Pre-processing (RRISCN-DLDP) Method are given below.

- RRISCN-DLDP Method is used to remove the noisy data and to fill in the missing values for classification performance enhancement. It comprises six layers, namely one input layer, four hidden layers and one output layer for efficient pre-processing.
- The RRISCN Method applies Ridge regularization to perform the data quality assessment through the mismatched data types, mixed data values, and data outliers. The multiple-regression coefficient is identified and multicollinearity issues are avoided also, it minimizes the marine weather forecasting time.
- Imputed nearest neighbor interpolation is utilized in the data cleaning task to fill the missing data (i.e., zonal value) and remove duplicate data in corresponding columns. In this way, the space-time complexity is said to be diminished.
- Pointwise Mutual Tanimoto similarity coefficient is employed in the data reduction to discover the input into two parts (i.e., marine weather data point and neighboring marine weather data point) according to the similarity value. Next, the duplicate marine weather data values are identified and removed. In this way, the time complexity is said to be reduced.
- Scaled clipping normalization applies a data transformation process to normalize the range of independent feature values of marine weather data points for precise forecasting.

The article is categorized: Section II discusses the RRISCN Method for the pre-processing task. Section III provides a detailed analysis of conducted experiment and results. Section IV explains the discussion of proposed and existing methods. Section V describes the conclusion of the current work.

II. Methodology

Forecasting is to create predictions depending on past as well as present data by fashion analysis. Weather forecasting is employed to predict the environment in the assured region with diverse weather metrics. Weather forecasting is performed with information on the present state of the environment. It is a tricky process for meteorologists as well as investigators. To address the issue, the marine weather forecasting approach was introduced in [25] considering dissimilar research by using big data.

Marine forecasting has done more difficult issues with require of information, limited entrée, as well as reporting irregularity. Accessibility is a difficult issue with both detecting enough examination sources as well as access the data. The size of our oceans creates it very tricky to gather data at scale. The incapability to access huge swaths of ocean territory constructs it hard to obtain sufficient data to create precise predictions. Therefore, the three ways such as diversifying data sources, incorporating wave data, and including more data sources used to improve marine weather forecasts.

Data pre-processing is used to transform the raw marine weather data points into structural data points. Raw data is deficient, partial, and not consistent with many errors. Consequently, data pre-processing is required before sending it through the classification process. The data quality is used for training and classification purposes. The marine weather data contains many numbers. The irrelevant one increases the dimensionality and makes the classification process a more difficult one. Consequently, it increases the computational complexity of the classification process. In order to address, these problems, RRISCN Method is introduced.

The current RRIPSCN Method is introduced for enhancing the marine weather forecast performance. The main aim of the RRIPSCN Method is to perform efficient marine weather data pre-processing with lesser space complexity and time complexity. On contrary to existing works, Ridge Regularized data quality assessment is employed for mismatched data, mixed data, as well as data outliers. Imputed nearest neighbor interpolation is utilized to fill in missing data values. Next, the Pointwise Mutual Tanimoto similarity coefficient is applied to eradicate duplicate data. Lastly, the scaled clipping normalization process is for performing data transformation. Therefore, the time and space consumption is decreased.

The architecture diagram of the RRISCN Method is given in Fig. 1.

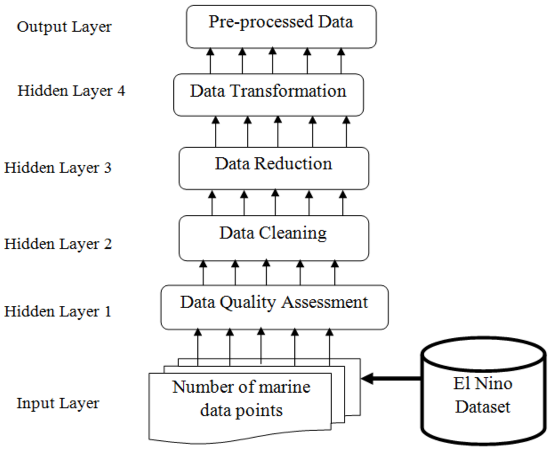

Fig. 1. Structural design of RRISCN method.

Fig. 1 describes the diagrammatic representation of the RRISCN Method. It comprises five layers, namely one input layer, three hidden layers, and one output layer for performing marine weather data pre-processing. The number of marine weather data points is collected from the El Nino dataset at the input layer. The input dataset comprises twelve features. After that, the proposed RRISCN Method performs the pre-processing of features through four different processes namely data quality assessment, data cleaning, data reduction, and data transformation in the four consecutive hidden layers. Finally, the pre-processed marine weather data is sent to the output layer. The block diagram of the RRISCN Method is illustrated in Fig. 2.

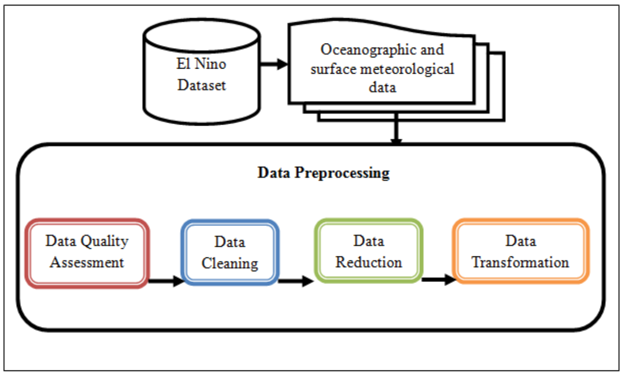

Fig. 2. Block diagram of pre-processing.

Fig. 2 illustrates the block diagram of marine weather data pre-processing. The input marine weather data are collected from the marine weather database and sent to the input layer. In RRISCN Method, the deep neural network includes the number of neurons that are connected from one layer to other consecutive layers in a feed-forward manner with variable weights. The neuron activity at an input layer at time 't' is given as,

$$Input(t) = \sum_{i=1}^{n} MWD_i \times w_{input} + B \qquad (1)$$

From Eq. (1), an input layer joined the input marine weather data 'MWD_i' with initial weight 'w_{input}'. 'B' symbolizes the bias value. After that, input marine weather data is sent to hidden layer 1.

A. Ridge Regularized Data Quality Assessment

In RRISCN Method, data quality assessment is carried out by identifying the mismatched data types, mixed data values, and data outliers. After collecting the marine weather data, it comes in a different format. The objective of the entire process is to reformat marine weather data, to begin with, formatted data. The different marine weather data sources use different descriptors for features. The value descriptors are made uniform. Data outliers have large impacts on data analysis results. Ridge regularization process is used in RRISCN Method for determining the multiple-regression coefficients where linearly independent variables (i.e., marine marine weather data points) are highly associated. Ridge regularization avoids the multicollinearity issues in linear regression with large numbers of parameters. Ridge regularization is formulated as,

$$\alpha_p = \underset{\alpha_p}{argmin} \sum_i \left(y_{MWD_i} - \alpha' x_{MWD_i}\right)^2 \qquad (2)$$

From Eq. (2), the '$argmin$' represent the argument minimum function. 'y_{MWD_i}' denotes the marine weather data points at 'i^{th}' column and 'y^{th}' row. 'x_{MWD_i}' Represent the marine weather data points at 'i^{th}' column and 'x^{th}' row. 'α'' symbolizes the ridge estimator. 'α_p' denotes feature value. Let us consider the input features as zonal winds. Normally, the zonal wind feature value is obtained in the negative value. When the positive value is seen in that particular feature column, it is considered as the mismatched feature value. In this way, mismatched marine weather data is identified and replaced as negative values for enhancing the data quality in the hidden layer 1.

B. Imputed Nearest Neighbor Interpolation Data Cleaning

In the second hidden layer of the RRISCN Method, the data cleaning process is performed. Missing data includes missing data fields, blank spaces in the text, or unanswered questions. The data cleaning task includes the filling of missing values and smoothing and removing the noisy data by resolving the inconsistencies. The missing data is identified and removed the duplicate data in the row and column of the dataset. The imputed Nearest Neighbor Interpolation method in RRISCN Method searches the whole dataset to show the features with missing data. Imputed nearest neighbor interpolation is carried out by approximating the neighbor value of the feature for a non-given point in corresponding columns. Let us consider the feature name as zonal winds. In that particular feature column, some zonal wind data gets missed. So, the missed value is approximated through neighbor values. Likewise, all missing data values are filled in the input database at hidden layer 2. By filling in the missed data values, the error rate during pre-processing process gets minimized. After filling in the missing value, duplicate data in the particular columns are removed through the data reduction process.

C. Pointwise Mutual Tanimoto Correlated Data Reduction

In the third hidden layer, the correlation function is used in RRISCN Method to find the duplicate marine weather data through Pointwise Mutual Tanimoto similarity analysis. Pointwise Mutual Tanimoto analysis is employed to find the linear relationship between two data points. Pointwise Mutual Tanimoto similarity coefficient [23] is used for finding the relationship between the marine weather data points to perform data reduction. It is formulated as below:

$$\alpha_{PMTCDR} = \frac{\alpha_{p_i} \cdot \alpha_{p_j}}{\sum \alpha_{p_i} + \sum \alpha_{p_j} - \alpha_{p_i} \cap \alpha_{p_j}} \tag{3}$$

From Eq. (3), 'α_{PMTCDR}' denotes the Pointwise Mutual Tanimoto correlated data reduction. 'α_{p_i}' denotes the marine weather data point, 'α_{p_j}' symbolizes the neighboring marine weather data point. '$\sum \alpha_{p_i}$' represents the sum of the score of 'α_{p_i}'. '$\sum \alpha_{p_j}$' represents the sum of the score of 'α_{p_j}'. '$\alpha_{p_i} \cdot \alpha_{p_j}$' represents the mutual dependence between the marine weather data point and neighboring marine weather data point. Tanimoto similarity coefficient provides the similarity value ranges between 0 and 1. Depending on the similarity value, the duplicate marine weather data values are identified and removed in the hidden layer 3. The duplicate marine weather data is removed to reduce space consumption and time consumption for pre-processing.

D. Scaled Clipping Normalized Data Transformation

The data transformation process is performed in the hidden layer 4 through scaled clipping normalization. Scaled clipping normalization is carried out in RRISCN Method for normalizing the range of independent feature values of marine weather data points. The scaled clipping normalization of the marine weather data point [24] is given as below:

$$N_{sc} = \left[\frac{\alpha_{PMTCDR} - min(\alpha)}{max(\alpha) - min(\alpha)}\right] \tag{4}$$

From Eq. (4), 'N_{SC}' represents the scaled clipping normalization process. '$max(\alpha)$' denotes the maximum feature value. '$min(\alpha)$' represents the minimum feature value. In this way, data transformation is carried out for the entire marine weather database in the hidden layer 4. The hidden layer result of the RRISCN Method is formulated as below:

$$Hidden(t) = [\sum_{i=1}^{n} MD_i \times w_{input}] + [w_{ih} \times N_{sc}] \tag{5}$$

From Eq. (5), '$Hidden(t)$' represent an output of a hidden layer. '$Hidden(t-1)$' symbolizes the output from the previously hidden layer. 'w_{ih}' denotes the weight between the hidden layer and the input layer. The pre-processed marine weather data output of the RRISCN Method is obtained at an output layer. It is given as below:

$$Output(t) = [w_{oh} \times Hidden(t)] \tag{6}$$

From Eq. (6), '$Output(t)$' symbolizes the output of the recursive deep neural network. 'w_{oh}' denotes the weight. In this way, marine data get pre-processed in RRISCN Method to perform efficient dimensionality reduction. The algorithmic process of RRISCN is explained as,

Algorithm 1: Ridge Regularized Imputed Pointwise Scaled Clipping Normalization based Deep Learnt Data Pre-processing

Input: El Nino Dataset

Output: Pre-processed marine weather data

```
Begin
Step 1: Collect the number of marine weather data points at the input layer
Step 2:     For each marine weather data in dataset
Step 3:     Perform marine weather data quality assessment through ridge regularization at hidden layer 1
Step 4:         Perform marine weather data cleaning using Imputed nearest neighbor interpolation at the hidden layer 2
Step 5:         Find missing feature of marine weather data
Step 6:             If missing feature value in respective column feature then
Step 7:             Approximate neighbor value of particular feature
Step 8: End if
Step 9:         Perform Data reduction through tanimoto similarity at the hidden layer 3
Step 10:    If duplicate value in the dataset then
Step 11:        Remove the feature value
Step 12:    End if
Step 13:            Apply scaled clipping normalization for data transformation at the hidden layer 4
Step 14:        Return (pre-processed marine weather data) at the output layer
Step 15:  End for
End
```

Algorithm 1 explains data pre-processing. Initially, marine weather data points are collected at an input layer. After that, the data quality is accessed via mismatched data types, mixed data values, and data outliers. Then, the missing data values are filled in the marine weather data cleaning process by approximating the neighbor value in corresponding columns. After that, the data duplication is removed in the data reduction process. Finally, the data transformation is performed through the scaled clipping normalization process. In this way, efficient marine weather data pre-processing is performed by using RRISCN Method to minimize the pre-processing time and space consumption.

III. Result

In this section, RRISCN Method, existing multi-objective grasshopper optimization [1], and Extended UNet architecture [2] are implemented using JAVA with E1 Nino dataset taken from https://www.kaggle.com/uciml/el-nino-dataset. It comprises the oceanographic as well as surface meteorological readings collected over a sequence of buoys through the equatorial Pacific Ocean. The main aim of the dataset is to forecast the seasonal-to-inter-annual climate variations like air temperature, surface temperatures, and Humidity. The dataset comprises 178,080 instances and 12 different attributes like observation, year, month, day, and so on. The latitude, as well as longitude, represents buoys shifted around dissimilar locations of the equatorial Pacific Ocean. The latitude values are recognized through degrees from an approximate location. The longitude values are gathered with five degrees of approximate location. The zonal and meridional winds were change among −10 m/s as well as 10 m/s. The relative humidity values are normally observed between 70% and 90%. The air temperatures well as sea surface temperature varied between 20 and 30 degrees Celsius. All meteorological readings are collected at the same time of day. To conduct the experiments, the different performance metrics namely pre-processing accuracy, pre-processing time, and space complexity is utilized.

The proposed method is designed by using Eqs. (1)−(6). The accuracy, time, and space complexity metric (i.e., Eqs. (7)−(9) used to analyze the performance of the proposed method. Hence, these equations are tabulated and compared with proposed and existing methods.

A. Analysis on Pre-processing Accuracy

The climate change risk such as greenhouse gases, ocean currents, Non use of Renewable energy sources is considered in the ecosystem. The oceanic risk such as sea level, rise, acidification, critical habitat loss, oil spills, and other non-biodegradables is taken for eco-risk assessment. Atmospheric variability and variability in ocean conditions, such as sea surface temperature, salinity, and sea ice cover and thickness are major effects on human behaviors as well as ecosystems both at sea and on land. Seasonal and sub-seasonal forecasts of ocean temperature were a vital part of managing marine ecosystems. Prediction of marine heat waves at sub-seasonal timescales are permit marine industries as well as managers to fine-tune operational strategies as well as employ strategies to reduce impacts on their businesses and resources.

In our work, weather is tricky to forecast, especially on waterways, however good forecasting are assist ships, as well as their crews, navigate and creating decisions that minimize risks. In E1 Nino dataset, estimating oceanographic as well as the surface meteorological variable is important for increasing prediction through the equatorial Pacific Ocean. Hence, accuracy is most important.

Pre-processing accuracy is referred by proportion of marine weather data that are pre-processed accurately to the total number of marine weather data. Pre-processing accuracy is computed as below:

$$PP_a = \sum_{i=1}^{n} \frac{Numberofmarineweatherdatapre-processedaccurately}{Totalnumberofmarineweatherdata} \times 100 \tag{7}$$

From Eq. (7), ‘ PP_a ’ symbolizes the pre-processing accuracy. It is calculated by percentage (%). The pre-processing accuracy values for proposed and existing methods are calculated and it is shown in Table I.

TABLE I: Comparison of Pre-processing Accuracy for RRISCN Method, Multi-objective Grasshopper Optimization, and Extended UNet Architecture

Number of Marine weather Data (Number)	Pre-processing Accuracy (%)		
	Multi-Objective Grasshopper Optimization	Extended UNet architecture	Proposed RRISCN Method
10,000	84.41	85.48	88.79
20,000	82.74	84.94	85.86
30,000	80.56	82.33	86.29
40,000	79.99	80.54	83.97
50,000	79.40	81.32	83.97
60,000	78.31	79.83	81.60
70,000	77.67	78.94	80.39
80,000	76.44	77.96	79.49
90,000	75.78	77.30	78.30
100,000	74.15	75.69	77.59

Table I explains the performance result of pre-processing accuracy versus the number of marine weather data collected from the input dataset varying from 10,000 to 100,000. The performance of pre-processing accuracy of three different methods namely the proposed RRISCN Method and existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2] are given in Table I. The table values reveal that the pre-processing accuracy of the proposed RRISCN Method is enhanced by two existing techniques. Let us consider that number of marine weather data is 20,000 in the second iteration. Consequently, the pre-processing accuracy of the proposed RRISCN Method is observed as 85.86%, and the pre-processing accuracy of existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2] is 82.74% and 84.94% correspondingly. Ten different pre-processing accuracy results of the proposed RRISCN Method are compared to the existing techniques.

Fig. 3 illustrates pre-processing accuracy versus the number of marine weather data. As described in Fig. 3, the green color bar indicates the pre-processing accuracy of the proposed RRISCN Method. The blue color and red color present the pre-processing accuracy of existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2].

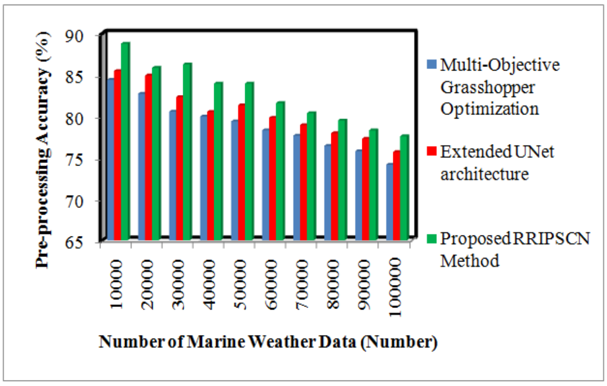

Fig. 3. Measurement of pre-processing accuracy.

Based on the comparison, the pre-processing accuracy of the proposed RRISCN Method is found to be increased. The ridge Regularized data quality is assessed through mismatched data types, mixed data values, and data outliers. The missing data values are filled in the data cleaning process through Imputed nearest neighbor interpolation for non-given points in corresponding columns. After that, the data duplication is eliminated in the data reduction process through Pointwise Tanimoto correlation analysis. The data transformation is carried out through a scaled clipping normalization process. The ten comparison results of the proposed RRISCN Method increase the pre-processing accuracy by 5% when compared to [1] and 3% when compared to [2] respectively.

B. Analysis on Pre-processing Time

Pre-processing time is described by the number of time consumed to pre-process marine weather data. It is calculated as below:

$$PP_{time} = \sum_{i=1}^{n} n \times Time[singleMWD_i] \qquad (8)$$

From Eq. (8), 'PP_{time}' symbolizes the pre-processing time. 'n' symbolizes the number of marine weather data. 'MWD_i' represents the marine weather data. It is calculated in milliseconds (ms). The values of pre-processing time for three methods by marine weather data are determined in Table II.

TABLE II: COMPARISON OF PRE-PROCESSING TIME FOR RRISCN METHOD, MULTI-OBJECTIVE GRASSHOPPER OPTIMIZATION, AND EXTENDED UNET ARCHITECTURE

Number of Marine weather Data (Number)	Pre-processing Time (ms)		
	Multi-objective Grasshopper Optimization	Extended UNet architecture	Proposed RRISCN Method
10,000	2500	2300	1900
20,000	2650	2450	2100
30,000	2800	2600	2250
40,000	2950	2785	2350
50,000	3010	2860	2500
60,000	3160	2955	2700
70,000	3250	3100	2920
80,000	3400	3250	3100
90,000	3650	3400	3300
100,000	3850	3700	3450

Table II describes the performance result of pre-processing time versus the number of marine weather data gathered from the input dataset varying from 10,000 to 100,000. The performance of pre-processing time of three different methods namely the proposed RRISCN Method and existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2] are given in Table I. The table values reveal that the pre-processing time of the proposed RRISCN Method is reduced than two existing techniques. Let us consider that number of marine weather data is 40,000 in the fourth iteration. Consequently, the pre-processing time of the proposed RRISCN Method is observed as 2350 ms and the pre-processing time of existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2] is 2950ms and 2785ms correspondingly. Ten different pre-processing time results of the proposed RRISCN Method are evaluated by the existing techniques.

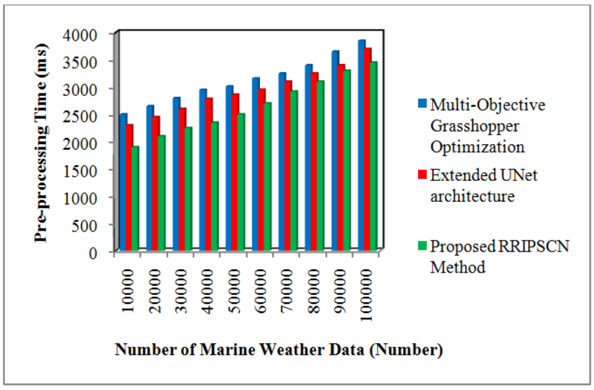

Fig. 4. Measurement of pre-processing time.

Fig. 4 describes pre-processing time versus a number of marine weather data. As described in Fig. 4, the green color bar indicates the pre-processing time of the proposed RRIPSCN Method. The blue color and red color bar represent the pre-processing time of existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2]. Based on the assessment, the pre-processing time of the proposed RRISCN Method is found to be minimized. The ridge Regularized data quality eliminates mismatched data, mixed data, and data outliers. After that, the data cleaning process is carried out through Imputed nearest neighbor interpolation in corresponding columns. The data duplication performed the data reduction process by using Pointwise Tanimoto correlation analysis. Finally, data transformation performs the scaled clipping normalization. In this way, the pre-processing time of the proposed RRISCN Method gets reduced. The ten comparison results of the proposed RRISCN Method reduce the pre-processing time by 16% when compared to [1] and 10% when compared to [2] respectively.

C. Analysis on Space Complexity

Space complexity is defined by number of memory space utilized to marine weather data pre-processing. It is determined by:

$$S_{com} \sum_{i=1}^{n} MWD_i\ Memoryspaceconsumed\ [singleMWD_i] \qquad (9)$$

From Eq. (9), 'S_{com}' denotes the space complexity. 'MWD_i' Denotes the marine weather data. The space complexity is determined by megabytes (MB).

TABLE III: Comparison of Space Complexity for RRISCN Method, Multi-objective Grasshopper Optimization, and Extended UNet Architecture

Number of Marine weather Data (Number)	Space Complexity (MB)		
	Multi-objective Grasshopper Optimization	Extended UNet architecture	Proposed RRISCN Method
10,000	143	120	75
20,000	168	135	102
30,000	189	156	128
40,000	205	180	152
50,000	224	205	180
60,000	271	239	202
70,000	304	258	230
80,000	318	290	248
90,000	339	301	268
100,000	362	358	308

Table III explains the performance result of space complexity versus a number of marine weather data collected from the input dataset ranging from 10,000 to 100,000. The performance of space complexity of three different methods namely the proposed RRISCN Method and existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2]. Table values illustrate the space complexity of the proposed RRISCN Method is reduced than two conventional techniques. Let us consider that number of marine weather data is 50,000 in the fifth iteration. Consequently, the space complexity of the proposed RRISCN Method is attained as 180MB and the pre-processing time of existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2] is 224MB and 205MB correspondingly. Ten different space complexity results of the proposed RRISCN Method are determined with respect to the existing techniques.

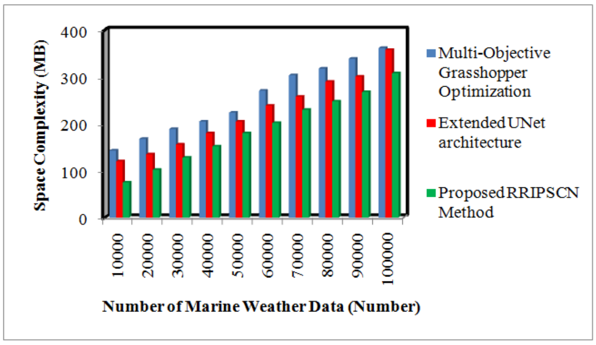

Fig. 5. Measurement of space complexity.

Fig. 5 illustrates space complexity versus number of marine weather data. As described in Fig. 5, the green color bar indicates the space complexity of the proposed RRISCN Method. The blue color and red color bar represent the space complexity of existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2]. Based on the observation, the space complexity of the proposed RRISCN Method is found to be minimized. The ridge Regularized data quality eliminates mismatched data, mixed data, and data outliers. After that, the data cleaning process is carried out through Imputed nearest neighbor interpolation in corresponding columns. This in turn minimizes the space complexity of the proposed RRISCN Method. The ten comparison results of the proposed RRISCN Method reduce the space complexity by 27% when compared to [1] and 17% when compared to [2] respectively.

IV. Discussion

In this section, the result of RRISCN Method and existing multi-objective grasshopper optimization [1] and Extended UNet architecture [2] are discussed with different performance metrics namely pre-processing accuracy, pre-processing time, and space complexity. First, the three pre-processing methods were applied to the E1 Nino dataset. The obtained result of the proposed and existing methods for each of the configurations is shown in Tables I−III. For considering 10,000 marine weather data, the pre-processing accuracy, pre-processing time, as well as space complexity was measured as 85.86%, 2100 ms and 102 MB with the proposed RRISCN method, 82.74%, 2650 ms, and 168 MB by using multi-objective grasshopper optimization and 84.94%, 2450 ms and 135MB with Extended UNet architecture respectively.

The experiments also proved the same situation. With increasing levels of marine weather data, i.e., 10,000, 20,000, 30,000….100,000, the percentages of accuracy decreased. It can be observed that for the three methods the pre-processing time increases as the number of marine weather data decreases. Therefore, the overall results of the RRISCN method perform better than the other conventional methods in almost all the scenarios. The outcome shows that the RRISCN method performs better with an improvement in accuracy by 4%, a reduction of time by 13%, as well as space complexity by 22% for accurate prediction than the existing works.

V. Conclusion

A new method termed RRISCN-DLDP removes the noisy data and fills in missing values for enhancing the classification performance. In RRISCN-DLDP Method, ridge Regularized data quality assesses the mismatched data, mixed data, and data outliers. The missing data values are filled in the data cleaning process with help of Imputed nearest neighbor interpolation through approximating the value for non-given points in corresponding columns. Then, the data duplication is removed by using Pointwise Tanimoto correlation analysis. Finally, the data transformation is carried out by using the scaled clipping normalization process. This, in turn, efficient data pre-processing is carried out to minimize time and space consumption. The performance of the RRISCN Method and existing classification techniques is determined with three different metrics such as pre-processing accuracy, pre-processing time, and space complexity. The observed results demonstrate that the higher pre-processing accuracy is achieved using the RRISCN Method and minimize the time consumption and space complexity when compared to conventional pre-processing methods.In the future, the proposed method is further extended to reduce the dimensionality by using a feature selection process for identifying relevant or redundant features for accurate prediction.

APPENDIX

TABLE A: LIST OF ABBREVIATION

Abbreviation	Description
RRISCN-DLDP	(RRISCN) based Deep Learnt Data Pre-processing
DSSAE	Deep learning-based Stacked Sparse Auto encoder
WRF	Marine weather Research and Forecasting
PSO-SVM	Particle Swarm Optimizer of Support Vector Machine
ML	MACHINE LEARNING
SVR	Support Vector Regression
GBRT	Gradient Boosting Regression Trees
SCADA	Supervisory Control and Data Acquisition
ICEEMDAN	Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
Bi-LSTM	Bidirectional long short-term memory
WPF	Wind Power Forecasting
FS	Feature Selection
GLSSVM	Group Least Square Support Vector Machine
LS-SVM	Least Square Support Vector Machines
GMDH	Group Method of Data Handling
PV	Photovoltaic
LSTM	Long Short-Term Memory
NN	Neural Network
NWP	Numerical Marine weather Prediction
MIC	Many Integrated Core
LSTM	Long Short-Term Memory

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Deepa Anbarasi J. had conducted the research, analyzed the performance of the models, and wrote the paper. Dr. V. Radha had guided towards the research work. Both the authors approve the final version.

REFERENCES

[1] X. Niu and J. Wang, "A combined model based on data pre-processing strategy and multi-objective optimization algorithm for short-term wind speed forecasting," *Applied Energy*, vol. 241, pp. 519–539, May 2019.

[2] J. G. Fernandez, I. A. Abdellaoui, and S. Mehrkanoon, "Deep coastal sea elements forecasting using U-Net based models," *Computer Vision and Pattern Recognition*, pp. 1–12, 2021.

[3] Y. Deng, B. Wang, and Z. Lu, "A hybrid model based on data pre-processing strategy and error correction system for wind speed forecasting," *Energy Conversion and Management*, vol. 212, pp. 1–12, May 2020.

[4] R. G. Madhukar and R. Dharavath, "DSSAE-BBOA: Deep learning-based marine weather big data analysis and visualization," *Multimedia Tools and Applications*, vol. 80, pp. 27471–27493, 2021.

[5] N. Krishnaveni and A. Padma, "Marine weather forecast prediction and analysis using sprint algorithm," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4901–4909, 2021.

[6] P. Kumar, S. P. Ojha, R. Singh, C. M. Kishtawal, and P. K. Pal, "Performance of marine weather research and forecasting model with variable horizontal resolution," *Theoretical and Applied Climatology*, vol. 126, pp. 705–713, 2016.

[7] M. Biswas, T. Dhoom, and S. Barua, "Marine weather forecast prediction: An integrated approach for analyzing and measuring marine weather data," *International Journal of Computer Applications*, vol. 182, issue 34, pp. 20–24, December 2018.

[8] Z. Zou, Y. Yang, Z. Fan, H. Tang, M. Zou, X. Hu, C. Xiong, and J. Ma, "Suitability of data pre-processing methods for landslide displacement forecasting," *Stochastic Environmental Research and Risk Assessment*, vol. 34, pp. 1105–1119, 2020.

[9] U. Singh and M. Rizwan, "Analysis of wind turbine dataset and machine learning based forecasting in SCADA-system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2022, pp. 1–12, 2022.

[10] L. Coulibaly, C. A. K. A. Kounta, B. Kamsu-Foguem, and F. Tangara, "Learning with deep Gaussian processes and homothety in marine weather simulation," *Neural Computing and Applications*, vol. 2022, pp. 1–15, 2022.

[11] Y. Deng, B. Wang, and Z. Lu, "A hybrid model based on data pre-processing strategy and error correction system for wind speed forecasting," *Energy Conversion and Management*, vol. 212, pp. 1–18, May 2020.

[12] V. Kosana, K. Teeparthi, and S. Madasthu, "Hybrid wind speed prediction framework using data pre-processing strategy based autoencoder network," *Electric Power Systems Research*, vol. 206, pp. 1–15, May 2022.

[13] M. Lv, J. Li, X. Niu, and J. Wang, "Novel deterministic and probabilistic combined system based on deep learning and self-improved optimization algorithm for wind speed forecasting," *Sustainable Energy Technologies and Assessments*, vol. 52, Part B, pp. 1–15, August 2022.

[14] D. Niu, L. Sun, M. Yu, and K. Wang, "Point and interval forecasting of ultra-short-term wind power based on a data-driven method and hybrid deep learning model," *Energy*, pp. 1–15, May 2022.

[15] M. Malvoni, M. G. Giorgi, and P. M. Congedo, "Forecasting of PV power generation using marine weather input data-pre-processing techniques," *Energy Procedia*, vol. 126, pp. 651–658, September 2017.

[16] M. S. Hossain and H. Mahmood, "Short-term photovoltaic power forecasting using an LSTM neural network and synthetic marine weather forecast," *IEEE Access*, vol. 8, pp. 172524–172533, September 2020,

[17] Q. Xu, D. He, N. Zhan, C. Kang, Q. Xia, J. Bai, and J. Huang, "A short-term wind power forecasting approach with adjustment of numerical marine weather prediction input by data mining," *IEEE Transactions on Sustainable Energy*, vol. 6, issue 4, pp. 1283–1291, October 2015.

[18] M. Huang, B. Huang, and H.-L. A. Huang, "Acceleration of the WRF Monin–Obukhov–Janjic surface layer parameterization scheme on an MIC-based platform for marine weather forecast," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, issue 10, pp. 4399–4408, October 2017.

[19] J. R. Andrade and R. J. Bessa, "Improving renewable energy forecasting with a grid of numerical marine weather predictions," *IEEE Transactions on Sustainable Energy*, vol. 8, issue 4, pp. 1571–1580, October 2017.

[20] Y. Yu, J. Cao, and J. Zhu, "An LSTM short-term solar irradiance forecasting under complicated marine weather conditions," *IEEE Access*, vol. 7, pp. 145651–145666, October 2019.

[21] S. J. Mohammed, S. L. Zubaidi, S. Ortega-Martorell, N. Al-Ansari, S. Ethaib, and K. Hashim, "Application of hybrid machine learning models and data pre-processing to predict water level of watersheds: Recent trends and future perspective," *Cogent Engineering*, vol. 9, no. 1, 2022.

[22] Z. S. Khudhair, S. L. Zubaidi, S. Ortega-Martorell, N. Al-Ansari, S. Ethaib, and K. Hashim, "A review of hybrid soft computing and data pre-processing techniques to forecast freshwater quality's parameters: Current trends and future directions," *Environments*, vol. 9, no. 7, p. 85, 2022.

[23] A. Sharma and S. P. Lal, "Tanimoto based similarity measure for intrusion detection system," *Journal of Information Security*, vol. 2, pp. 195–201, 2011.

[24] D. D. Noel, K. G. A. Justin, A. K. Alphonse, L. H. Désiré, D. Dramane, and D. N. G. Malerba, "Normality assessment of several quantitative data transformation procedures," *Biostat Biom Open Access J.*, vol. 10, no. 3, 555786, 2021.

[25] J. D. Anbarasi and V. Radha, "Review on marine weather forecasting with big data," in *Proc. 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022.