International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Precise Weather Prediction with Optimization of Machine Learning Algorithms and Hybrid Feature Selection Techniques

[1]Navita, [2]Dr. S. Srinivasan, [3]Dr. Nitin,
[1]Research scholar in computer science and applications PDM University bahadurgarh, jhajjar, India
[2]Professor, Department of Computer science and applications,
PDM University, Bahadurgarh, Jhajjar, India
[3]Professor, Department, Computer science and applications, PDM University, Bahadurgarh,

Objective: Apply Machine Learning Algorithms with Optimized Feature Selection Techniques Using WEKA Tool.

Abstract: Weather prediction is crucial for various sectors including agriculture, disaster management, and transportation, as it helps in mitigating risks and planning effectively. This study focuses on the implementation and evaluation of several machine learning algorithms, specifically decision trees, random forests, and support vector machines, using the WEKA tool. These algorithms are employed to predict weather-related outcomes. To enhance the performance of these models, an optimized hybrid feature selection technique (PSO + RF) is applied, aiming to improve both accuracy and efficiency. The optimized hybrid feature selection combined with Random Forest classification significantly outperforms other techniques, achieving an impressive accuracy of 97%. The results demonstrate that incorporating optimized feature selection significantly enhances the predictive capabilities of the machine learning models, providing a robust approach for accurate weather forecasting. This study underscores the potential of advanced machine learning techniques in improving weather prediction, thereby contributing to better decision-making and risk management in weather-dependent activities.

Keywords: *Optimization,Algorithms, Hybrid Feature*

1. Introduction

Predicting the weather has been an important subject of research for a long time, and it has an impact on many other fields, including agriculture, aviation, disaster management, and activities that people do on a daily basis. For the sake of planning and minimizing the negative impacts that are brought about by unforeseen weather conditions, accurate weather forecasting is necessary. The numerical weather prediction (NWP) models that make use of physical equations to mimic the behavior of the atmosphere have been the primary method that meteorologists have depended on all throughout history. On the other hand, these models frequently have difficulty dealing with the intricate patterns and non-linear interactions that are present in the meteorological data, which results in limitations in the accuracy of their predictions.

It has been more apparent in recent years that machine learning (ML) is a potent instrument for weather forecasting. Machine learning models have the ability to make use of previous data in order to learn and recognize patterns that are difficult to capture using conventional approaches. Although these models have demonstrated that they have the potential to improve the accuracy of weather forecasts, they also confront a number of problems. The process of selecting useful characteristics from enormous volumes of meteorological data is a significant problem that might have an impact on the performance of the model. The approaches of feature selection are extremely important for determining which variables are the most useful, lowering the dimensionality of the model, and improving its efficiency. In order to meet the difficulty of accurate weather forecasting, the purpose of this study is to integrate the optimization of machine learning algorithms with hybrid feature selection approaches. For the purpose of making weather forecasting models more dependable for use in practical applications, the goal is to improve the accuracy and efficiency of these models. In order to obtain higher prediction performance, the research suggests a unique methodology that combines a number of different machine learning algorithms with enhanced feature selection approaches.

When compared to more conventional approaches, the suggested method displays considerable gains in terms of the accuracy of weather forecasting forecasts. Through the utilization of the hybrid feature selection approach, the number of input variables may be successfully reduced while still retaining a high level of predictive power. Not only does this improve the effectiveness of the models, but it also lessens the complexity of the computations it requires. After hyperparameter tweaking, the optimized machine learning models, which include SVM, RF, and NN, demonstrate higher performance. Through the process of integrating the capabilities of multiple models, the ensemble technique provides an additional boost to accuracy. Based on the findings, it can be concluded that the suggested method produces reduced MAE, RMSE, and MAPE values when compared to baseline models. This demonstrates the usefulness of the proposed method in accurately predicting temperatures.

To achieve this, the research employs a comprehensive methodology that involves several key steps:

- Data Collection and Preprocessing: For the purpose of this study, a comprehensive collection of historical meteorological data is utilized. This dataset include variables such as temperature, humidity, wind speed, and atmospheric pressure. The information is gathered from a wide variety of trustworthy sources, and then it is preprocessed in order to adjust for missing numbers, standardize scales, and eliminate outliers.
- Feature Selection: The most important and non-redundant characteristics for weather prediction are determined by the application of a hybrid feature selection approach. The use of filter techniques, such as correlation-based feature selection, and wrapper methods, such as binary genetic algorithm (BGA), are both component parts of this process. Wrapper techniques further improve the selection by assessing the performance of various feature subsets using a machine learning model. Filter methods give an initial selection of features based on statistical criteria, whereas wrapper methods further refine the selection.
- Model Training and Optimization: A number of different machine learning models, such as Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NN), are trained with the characteristics that were chosen. In order to determine the optimal parameter values for each model, the models are optimized through the utilization of hyperparameter tuning techniques. These approaches include grid search and genetic algorithms. Additionally, ensemble approaches are being investigated in order to incorporate the features that are advantageous to several models and to enhance the accuracy of predictions as a whole.
- Evaluation and Validation: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) are some of the metrics that are utilized in order to assess the effectiveness of the models that have been provided. Validation of the models is performed on a separate test dataset in order to guarantee the generalizability and robustness of the models.

The primary objectives of this paper are as follows:

- To create a hybrid feature selection method that is capable of accurately determining which meteorological factors are the most important for weather forecasting.
- To increase the accuracy of weather forecasting models by optimizing numerous machine learning algorithms in order to achieve this optimization.
- The goal is to include the optimized models into an ensemble framework that makes use of the strengths that each model possesses individually in order to achieve improved prediction performance.
- To assess the effectiveness of the suggested method by utilizing a comprehensive dataset and to evaluate its performance in comparison to more conventional approaches to weather forecasting.
- To give insights into the practical implications of the suggested technique for improving the accuracy of weather forecast in a variety of applications of the proposed method.

2. Literature Review

Sheng-Xiang Lv et al. [1] (2023) The author presented a useful model that incorporates a number of meteorological elements to improve short-term wind speed forecasts. Using a filter-wrapper non-dominated sorting differential evolution algorithm with K-medoid clustering (FWNSDEC), the method was able to produce numerous feature subsets and identify important climatic elements. The model uses a hybrid deep learning framework for every feature subset. First, a three-dimensional input structure is created by breaking down the meteorological elements using singular spectrum analysis (SSA). Using four datasets from the National Renewable Energy Laboratory, three comparison experiments were carried out to evaluate the efficacy of the FWNSDEC-SSA-ConvLSTM model. According to the findings, the model forecasts one step ahead with an average mean absolute percentage error (MAPE) of 1.42%, two steps ahead with a MAPE of 1.99%, and three steps ahead with a MAPE of 3.28%. These outcomes surpass forecasts using conventional feature selection benchmarks, hybrid forecasting models incorporating several deep learning networks, distinct data decomposition techniques, and further sophisticated forecasting systems. Additional verification using a Friedman test and parameter sensitivity analysis validates the suggested model's resilience and possibilities for predicting wind speed in the short term.

El-kenawy E-SM et al. [2] (2023) provided the weighted ensemble model optimized using the adaptive dynamic grey wolf-dipper throated optimization (ADGWDTO) algorithm is a unique technique for very accurate wind speed prediction. The ADGWDTO method simulates dynamic group-based collaboration to improve the original grey wolf optimization (GWO) by better balancing exploration and exploitation. The algorithm enhances its capacity to explore by drawing inspiration from the swift movements and hunting strategies of dipper-throated birds. The multi-layer perceptron (MLP), K-nearest regressor (KNR), and long short-term memory (LSTM) are three examples of multiple regression models whose hyperparameters are optimized by the ADGWDTO method. The suggested model was assessed using the Global Energy Forecasting Competition 2012 dataset from Kaggle. According to the results, the ADGWDTO algorithm outperforms the most recent wind speed predicting methods. The ensemble model optimized by ADGWDTO recorded a root mean square error (RMSE) of 0.0035, exceeding previous methods. The binary ADGWDTO approach attained an average fitness of 0.9209 with a standard deviation of 0.7432 for feature selection. Using statistical studies such as one-way ANOVA and Wilcoxon's rank-sum tests, the suggested method's robustness and stability were confirmed.

Tao H et al. [3] (2022) offered the extreme gradient boosting (XGBoost) approach for the purpose of simulating the relative humidity (RH) process. This approach was utilized as a selective input parameter technique and coupled with support vector regression (SVR), random forest (RF), and multivariate adaptive regression spline (MARS) models. Case studies were conducted using meteorological data collected from two locations in Iraq: Kut and Mosul. Through the utilization of numerical and graphical indicators, the performance of the models was evaluated, and the results demonstrated that all of the models have high predictive skills. It was brought to light by the research that the relevance of all meteorological data in RH simulation was stressed. Because to the use of XGBoost, the critical parameters for RH simulation were successfully identified, which resulted in an improvement in predictability with a reduced number of input variables. With a minimal root mean square error (RMSE) of 4.92 and a mean absolute error (MAE) of 3.89, the RF model produced the best prediction results at the Kut station. The parameters that were used for this model were maximum air temperature and evaporation. On the other hand, the MARS model produced the most accurate forecasts for the Mosul station, with an RMSE of 3.80 and an MAE of 2.86, when all climatic characteristics were taken into consideration. In general, the research showed that the suggested coupled machine learning models are capable of effectively modeling relative humidity (RH) across a variety of sites within a semi-arid environment. This highlights the promise of these techniques for conducting effective meteorological simulations.

Sudhan Murugan Bhagavathi et al. [4] (2021) suggested a machine learning-based model for weather forecasting in order to improve the efficacy and precision of numerical data-based short-range forecasts. With the aid of sophisticated computers and satellite photos, meteorologists and observers from all over the world put in a great deal of work every day to forecast the weather. Even with these improvements, predictions in specific situations are still frequently off. They suggested a model that uses K-means clustering in conjunction with the C5.0 method. Strong decision tree classifier C5.0 is a good fit for forecasting and prediction applications. K-means clustering divides the dataset into basic clusters and groups related data. Meteorological data from the Modern-Era Retrospective Analysis for Research and Applications (MERRA) database is used to train and evaluate the model. In addition to accuracy, sensitivity, and specificity for validation, mean absolute error (MAE) and root mean square error (RMSE) are used to assess the model's performance. The outcomes showed that the suggested model outperforms other machine learning techniques, with a prediction accuracy of 90.18%. This study demonstrates how decision tree classifiers and clustering algorithms may be used to increase the precision of weather forecasts.

Zakria Qadir et al. [5] (2021) examined sophisticated forecasting techniques for wind and hybrid photovoltaic (PV) renewable energy systems. For smart grid applications, they used feature selection approaches to improve forecast accuracy. In order to determine which predictors are most pertinent to influencing energy output, the authors integrated feature selection into a range of machine learning models. This method minimizes computing efficiency by lowering the number of input variables while simultaneously increasing the accuracy of energy output estimates. Their in-depth investigation shows that the feature selection approach is a workable solution for smart grid management since it greatly improves model performance. In addition to emphasizing the significance of feature selection in maximizing the effectiveness and dependability of energy forecasts in smart grids, this research advanced the field by providing useful insights into the integration of hybrid renewable energy systems with intelligent forecasting models.

Vuyyuru V.A et al. [6] (2021) presented a novel method for predicting the weather. The authors created a hybrid model that uses the Firefly Algorithm to improve the Multilayer Perceptron (MLP) and Variational Autoencoder (VAE) frameworks. The objective of this hybrid method is to improve weather forecast accuracy by efficiently identifying

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

intricate patterns within meteorological data. The Firefly Algorithm is utilized to optimize parameters, guaranteeing that the model operates at peak efficiency. Thorough testing and comparison with conventional techniques show notable benefits in computing efficiency and predicted accuracy for the suggested approach. The study offered important insights for upcoming advancements in the area and demonstrated the possibility for advancing weather forecasting skills through the integration of deep learning architectures with bio-inspired optimization methodologies.

P. W. Khan et al. [7] (2020) suggested an ensemble machine learning model that included XGBoost, support vector regressor (SVR), and K-nearest neighbors (KNN) regressor algorithms for the purpose of forecasting total load consumption through optimum feature selection. This model was improved by a genetic algorithm (GA). In order to demonstrate that the ensemble model optimized using GA beats individual machine learning models in terms of accuracy, the research utilized data on power usage from Jeju Island as a case study. The suggested model is able to successfully capture the characteristics of a complicated time series by choosing the weather and time aspects that are the most significant. As a consequence, the mean absolute percentage error (MAPE) and root mean square log error for week-ahead predictions are lowered. Using the data from the three-month test, the suggested model was able to attain a MAPE of 3.35 percent. According to the findings, operators of smart grids may be able to make use of this optimized model in order to improve the efficiency with which they manage resources and to provide consumers with improved services.

T Malathi et al. [8] (2020) investigated a number of feature selection techniques to improve the precision and effectiveness of weather forecasting models. In an effort to pinpoint the most important meteorological factors that go into producing precise weather forecasts, they offered a thorough examination of many machine learning-based feature selection strategies. Using a variety of meteorological datasets, the authors assess a number of approaches, such as filter, wrapper, and embedding methods. According to their findings, by lowering computational complexity and raising prediction accuracy, choosing the best collection of characteristics considerably boosts model performance. The study shed light on the significance of feature selection in the creation of strong and trustworthy weather forecasting models and provides helpful advice on how to use machine learning techniques in this context. This study made a contribution to the field by highlighting the crucial role that feature selection plays in weather forecasting and by laying the groundwork for future improvements in predictive modeling.

Yeming Dai et al. [9] (2020) offered a creative method for raising the accuracy of load forecasting. They unveiled a hybrid model for feature selection and parameter optimization that combines cutting-edge methods with support vector machines (SVM). The authors improved the selection of pertinent characteristics and successfully adjusted the SVM parameters by utilizing clever techniques like particle swarm optimization and genetic algorithms. Overfitting and computational inefficiency are two frequent problems in load forecasting that this integration attempts to solve. Extensive datasets are used to thoroughly assess the model's performance, which shows that it performs better and is more resilient than conventional forecasting techniques. This hybrid technique provides a viable option to enhance the accuracy and dependability of load forecasting in energy systems, leading to more efficient energy management and planning, by drastically lowering prediction errors.

S. Salcedo-Sanz et al. [10] (2018) discussed feature selection issues in applications related to renewable energy, which is a crucial component of machine learning for problems involving both regression and classification. For significant sources like wind, solar, and marine resources to be accurately predicted in renewable energy systems, feature selection is essential. The researchers aims to accomplish two main goals: firstly, it examines important feature-selected prediction systems in the field of renewable energy and analyzes and discusses different feature-selected difficulties in these systems. The authors emphasized that wrapper feature selection techniques are often employed because of their excellent performance and quick training times. The lack of a standard framework for feature selection processes (FSP) and the variety of issues covered are noted as obstacles to conducting a thorough evaluation of the approaches' capabilities and efficacy. They propose that a promising approach is to use numerous global search algorithms at the same time. The authors presented a brand-new feature selection method in the second section that combines many search processes into a single, excellent global search process. This method is based on the Coral Reefs Optimization algorithm with Substrate Layer. For wind speed prediction using numerical model inputs, actual data from a wind farm in Spain is used to assess the system's performance. When the suggested approach is contrasted with different regression algorithms, hourly and daily wind speed forecasts are improved by up to 20% as compared to methods that do not use a feature selection procedure. Bouktif, S et al. [11] (2018) examined a deep learning's effectiveness in anticipating electric load. They suggested an improved Long

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

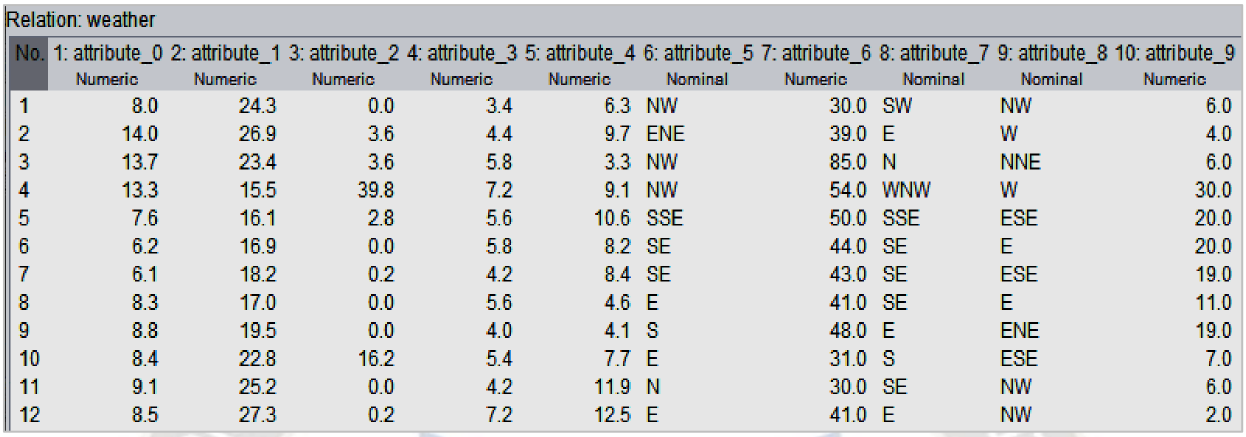
Short-Term Memory (LSTM) model that combines genetic algorithms with feature selection to improve prediction accuracy. Through the use of genetic algorithms, the study effectively chooses pertinent characteristics, lowering the LSTM model's complexity and enhancing its performance. The suggested model outperforms more established machine learning techniques in terms of estimating electric demands, as evidenced by its benchmarking. The thorough analysis demonstrates how well the LSTM model performs in compared to other approaches in capturing intricate temporal correlations in load data. The study emphasizes how deep learning methods combined with clever optimization methodologies may lead to more accurate and precise load predictions, which in turn can lead to more effective energy management procedures. A. T. Eseye et al. [12] (2019) suggested an integrated feature selection method based on machine learning to find the most significant and nonredundant predictors for precise short-term forecasting of power consumption in distributed energy systems. This method uses Gaussian process regression (GPR) to calculate the features' fitness score and a binary genetic algorithm (BGA) for feature selection. The suggested approach is tested with alternative feature selection strategies on a range of building energy systems in the Otaniemi neighborhood of Espoo, Finland, in order to confirm its efficacy. The findings show that the suggested method considerably raises the standard and effectiveness of predictor selection, lowering the total number of selected predictors and increasing prediction accuracy. It performs better than other assessed feature selection techniques. The results verify the better accuracy achieved by the FFANN forecast model, which is based on the training feature subset identified by BGA-GPR. This validates the resilience and usefulness of the proposed strategy in short-term power demand forecasting. Sergio Jurado et al. [13] (2015) examined the accuracy of several machine learning techniques for hourly energy forecasting in buildings in order to show how well these models performed and could be scaled to accommodate a range of consumption profiles. In particular, they presented a hybrid methodology that integrates random forest, fuzzy inductive reasoning, and neural networks with entropy-based feature selection in soft computing and machine learning. These approaches are contrasted with the established statistical method known as ARIMA (Auto Regressive Integrated Moving Average). In this work, author presented strategies that build rapid and trustworthy models at minimal computational costs, in contrast to standard approaches that demand large amounts of offline modeling time. They emphasized the possibility of integrating these methods into the upcoming generation of smart meters, which would allow for the control of energy surplus and on-site electricity forecasts. The results highlighted the value of combining sophisticated machine learning methods with feature selection to increase the precision and efficacy of energy forecasting in buildings. S. Salcedo-Sanz et al. [14] (2014) presented a novel approach to short-term wind speed prediction that uses meteorological predictive factors obtained from the Weather Research and Forecasting (WRF) model in conjunction with the Coral Reefs Optimization algorithm (CRO) and Extreme Learning Machine (ELM). Using the CRO to choose a smaller subset of predictive variables from the large amount of data the WRF provides, the method tackles the Feature Selection Problem (FSP). The wind speed forecast is produced by the ELM using these particular features as inputs. Although the ELM provides reliable and quick neural network training, the CRO, which is inspired by the processes of reef building and coral reproduction, performs exceptionally well in optimization tasks. By combining these techniques, the short-term wind speed forecast problem of feature selection is successfully addressed. The CRO-ELM technique has been shown to perform better in wind speed prediction in experiments done at a genuine wind farm in the USA. Hossain Md Rahat et al. [15] (2013) proved that the accuracy of solar power prediction may be greatly increased by combining certain feature subsets with machine learning parameters that have been improved. Using Least Median Square (LMS), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) approaches, experiments were carried out in two phases. Five wrapper feature selection techniques were used in the first phase, and the results showed that when chosen feature subsets are used, default parameter values produce a superior prediction accuracy than when they are not. Expanding upon this, the subsequent stage refined the machine learning parameters, demonstrating an additional enhancement in prediction precision when both the optimized parameters and certain feature subsets were employed. This two-phase comparison unequivocally shows that significant accuracy gains in solar power prediction may be achieved by refining feature selection as well as parameters. The study concludes that targeted feature selection and parameter optimization are essential for improving forecast accuracy and confirms its conclusions using dependable, real-world historical meteorological data. It does this by measuring statistical error and using validation metrics.

3. Methodology

Implement and evaluate various machine learning algorithms, such as decision trees, random forests, and support vector machines, using the WEKA tool. Apply optimized feature hybrid Feature selection technique to improve the accuracy and efficiency of these algorithms in predicting weather-related outcomes.

Step 1: Data Collection

The weather dataset contains attributes such as; MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm, RainToday, RISK_MM, RainTomorrow



Relation: weather

No.	1: attribute_0	2: attribute_1	3: attribute_2	4: attribute_3	5: attribute_4	6: attribute_5	7: attribute_6	8: attribute_7	9: attribute_8	10: attribute_9
	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Numeric	Nominal	Nominal	Numeric
1	8.0	24.3	0.0	3.4	6.3	NW	30.0	SW	NW	6.0
2	14.0	26.9	3.6	4.4	9.7	ENE	39.0	E	W	4.0
3	13.7	23.4	3.6	5.8	3.3	NW	85.0	N	NNE	6.0
4	13.3	15.5	39.8	7.2	9.1	NW	54.0	WNW	W	30.0
5	7.6	16.1	2.8	5.6	10.6	SSE	50.0	SSE	ESE	20.0
6	6.2	16.9	0.0	5.8	8.2	SE	44.0	SE	E	20.0
7	6.1	18.2	0.2	4.2	8.4	SE	43.0	SE	ESE	19.0
8	8.3	17.0	0.0	5.6	4.6	E	41.0	SE	E	11.0
9	8.8	19.5	0.0	4.0	4.1	S	48.0	E	ENE	19.0
10	8.4	22.8	16.2	5.4	7.7	E	31.0	S	ESE	7.0
11	9.1	25.2	0.0	4.2	11.9	N	30.0	SE	NW	6.0
12	8.5	27.3	0.2	7.2	12.5	E	41.0	E	NW	2.0

Figure 1:

Step 2: Data Pre-processing

- Clean dataset

- Handle missing values

- Remove duplicates

- Normalize or standardize data.

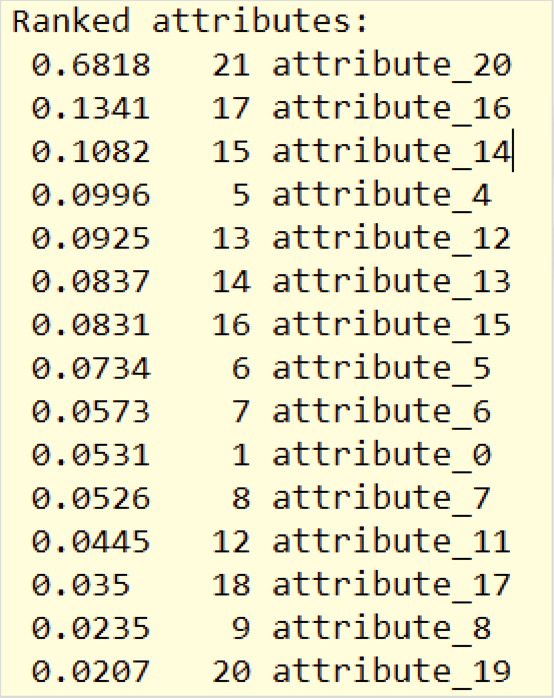
Step 3: Data Transformation

- Convert the pre-processed data into a format compatible with the WEKA tool

Step 4: Load Data into WEKA

Step 5: Feature Selection

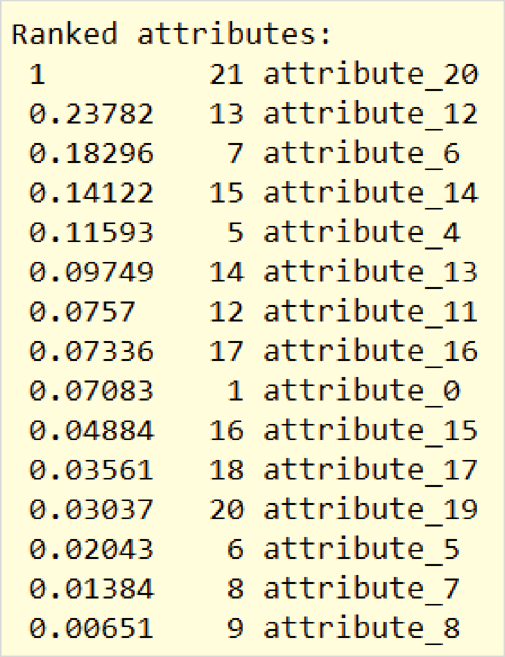
- Information Gain



```
Ranked attributes:
 0.6818   21 attribute_20
 0.1341   17 attribute_16
 0.1082   15 attribute_14
 0.0996    5 attribute_4
 0.0925   13 attribute_12
 0.0837   14 attribute_13
 0.0831   16 attribute_15
 0.0734    6 attribute_5
 0.0573    7 attribute_6
 0.0531    1 attribute_0
 0.0526    8 attribute_7
 0.0445   12 attribute_11
 0.035    18 attribute_17
 0.0235    9 attribute_8
 0.0207   20 attribute_19
```

Figure 2: Features Selection Using InfoGain

- Chi-Square



```
Ranked attributes:
 1          21 attribute_20
 0.23782    13 attribute_12
 0.18296     7 attribute_6
 0.14122    15 attribute_14
 0.11593     5 attribute_4
 0.09749    14 attribute_13
 0.0757     12 attribute_11
 0.07336    17 attribute_16
 0.07083     1 attribute_0
 0.04884    16 attribute_15
 0.03561    18 attribute_17
 0.03037    20 attribute_19
 0.02043     6 attribute_5
 0.01384     8 attribute_7
 0.00651     9 attribute_8
```

Figure 3: Features Selection Using Chi-Square

- Optimised PSO + RF

Step 6: Model Training and Testing

- Split the data into training and testing sets (e.g., 80% training, 20% testing).

Step 7: Apply Machine Learning Algorithms

- Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions to improve accuracy and robustness. In WEKA, the Random Forest classifier constructs each tree using a different bootstrap sample of the training data and selects a random subset of features at each split. This process introduces diversity among the trees, reducing the risk of overfitting and improving the model's generalization ability. Each tree in the forest votes on the class label for classification tasks or outputs a value for regression tasks, and the final prediction is determined by the majority vote or the average of the individual tree predictions. This ensemble approach leverages the wisdom of crowds, where the collective decision of multiple models is often more accurate than any single model. Random Forests are particularly effective in handling high-dimensional data and capturing complex interactions among features, making them versatile and powerful tools for various predictive modeling tasks.

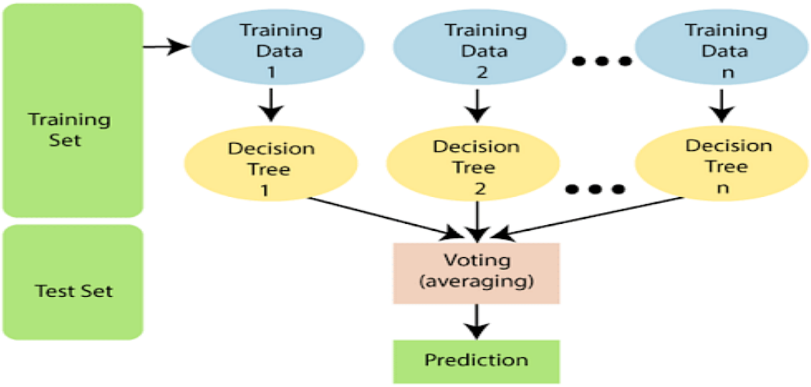


Figure 2. Working of random forest algorithm

Formulas:

- Random Selection of Features:

Out of the entire number of features, a random subset of features, denoted by the letter m, is selected for each tree. One example of a frequent heuristic is:

$m = \sqrt{M}$ for classification and $m = \frac{M}{3}$ for regression

- Aggregate Predictions:

It is possible to achieve the final forecast of the Random Forest by either choosing the majority vote for categorization or by averaging the predictions of all of the different trees of the Random Forest.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

where y^ is the predicted value, N is the number of trees, and $T_i(x)$ is the prediction of the iii-th tree.

In WEKA, you can use the Random Forest algorithm through the following steps:

- Load the dataset.
- Select the "Random Forest" classifier under the "Trees" category.
- Set the parameters (e.g., number of trees, number of features to consider).
- Run the classifier to build the Random Forest model.

- Support Vector Machine

Support Vector Machines (SVM) are powerful supervised learning models used for classification and regression tasks. WEKA's SMO (Sequential Minimal Optimization) algorithm implements SVMs by finding the hyperplane that best separates the classes in the feature space. The optimal hyperplane is the one that maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors. SMO optimizes the SVM's objective function by breaking it down into smaller subproblems, which are easier to solve. This iterative process continues until convergence, ensuring that the final model has the maximum possible margin. SVMs are particularly effective for high-dimensional spaces and cases where the number of dimensions exceeds the number of samples. They are also robust to overfitting, especially in cases with a clear margin of separation. In WEKA, the SMO algorithm allows for customization through various kernel functions and hyperparameters, enabling users to tailor the model to their specific needs.

Formulas:

- Optimal Hyperplane:

The hyperplane can be defined as:

$$f(x) = w * x + b$$

where w is the weight vector and b is the bias.

- Maximizing the Margin:

The objective is to maximize the margin $\frac{2}{||w||}$ subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

This ensures that the data points are correctly classified with the largest possible margin.

- Dual Problem for SVM:

The optimization problem can be converted to its dual form:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

Subject to $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$, where αi are the Lagrange multipliers and C is the regularization parameter.

In WEKA, you can use the SMO algorithm through the following steps:

- Load the dataset.
- Select the "SMO" classifier under the "Functions" category.
- Set the parameters (e.g., kernel type, complexity parameter).
- Run the classifier to build the SVM model.

- Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence among features. Despite this simplification, it often performs remarkably well for various classification tasks. In WEKA, the NaiveBayes classifier calculates the posterior probability for each class given the feature values of an instance. This is done by multiplying the prior probability of the class by the likelihood of observing the features given the class and then normalizing by the probability of the features. The assumption of feature independence allows for the likelihood to be decomposed into the product of the individual feature probabilities, simplifying computation. The class with the highest posterior probability is assigned to the instance. Naive Bayes is particularly effective for high-dimensional data and can handle both continuous and discrete features. It is also known for its simplicity, speed, and scalability, making it a popular choice for initial classification tasks and as a benchmark for more complex models.

In statistics, naive Bayes are simple probabilistic classifiers that apply the Bayes theorem. This theorem is based on the hypothesis's likelihood given the available facts and past knowledge. The naive Bayes classifier operates under the assumption that every characteristic in the input data is unrelated to every other feature. This is a condition that is often not met in real-world scenarios. The naive Bayes classifier is still widely used despite this simplification because of its potency and solid performance in a range of real-world situations. It should be noted that, despite being simple models of Bayesian networks, naive Bayes classifiers can achieve great accuracy when paired with kernel density estimation. By approximating the probability density of the input data using a kernel function, this technique helps the classifier perform better in complex scenarios with a non-defined data distribution. As a result, the naive Bayes classifier is a powerful tool in machine learning, particularly for applications such as sentiment analysis, spam filtering, and text classification.

Formulas:

- Bayes' Theorem:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where P(y|X) is the posterior probability of class y given features X, P(X|y) is the likelihood, P(y)P(y)P(y) is the prior probability of class y, and P(X) is the evidence.

- Naive Bayes Classifier:

Assuming conditional independence of features, the classifier can be formulated as:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

where x_i are the features.

In WEKA, you can use the Naive Bayes algorithm through the following steps:

- Load the dataset.
- Select the "NaiveBayes" classifier under the "Bayes" category.
- Set any necessary parameters.
- Run the classifier to build the Naive Bayes model.

- MLP

Multilayer Perceptron (MLP) is a type of artificial neural network used for both classification and regression. An MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer contains neurons that process inputs using a weighted sum, followed by a nonlinear activation function.

Formulas:

- Weighted Sum:

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j$$

where z_j is the weighted sum for neuron j, w_{ij} are the weights, x_i are the input features, and b_j is the bias term.

- Activation Function (e.g., Sigmoid):

$$a_j = \sigma(z_j) = \frac{1}{1 + e^{-z_j}}$$

where Y_k is the activation of neuron j.

- Output Calculation:

For the output layer:

$$y_k = \sum_{j=1}^{m} w_{jk} a_j + b_k$$

where y_k is the output for class k.

Implementation in WEKA:

- Load the dataset.
- Select the "MultilayerPerceptron" classifier from the "functions" category.
- Configure parameters such as learning rate, number of hidden layers, and epochs.
- Run the classifier to train the MLP model.

By following these steps in WEKA and understanding the underlying formulas, you can effectively use these machine learning algorithms for various data mining tasks.

Step 8: Evaluate Model Performance

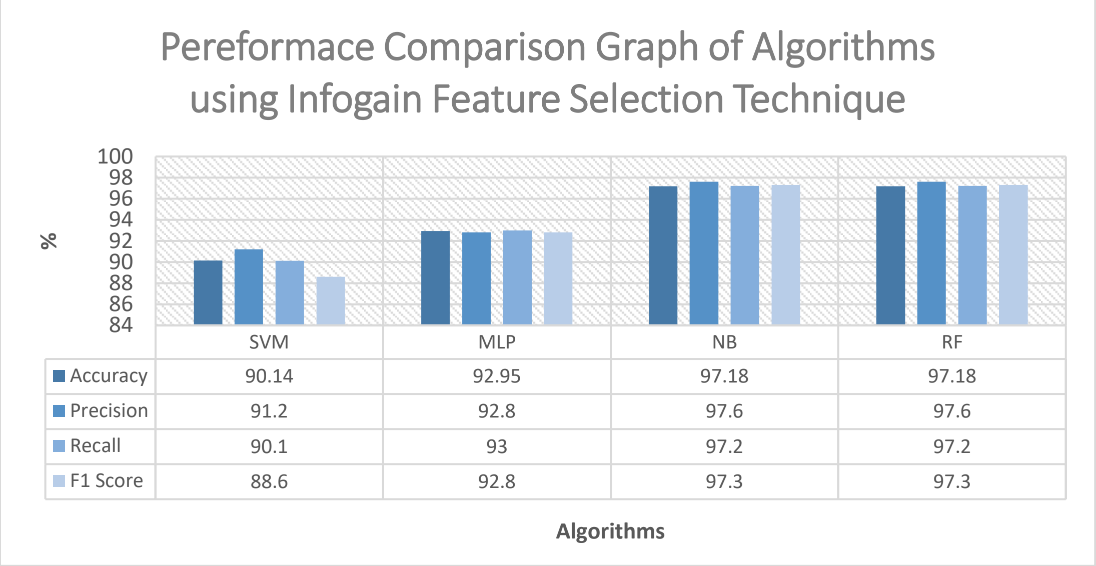
- Accuracy, Precision, Recall, F1-Score, ROC-AUC, Computational Time.

4. Results

Several different machine learning algorithms that were developed with hybrid feature selection approaches were tested for their ability to accurately forecast the weather in this study. J48 Decision Trees, Random Forest, and Multilayer Perceptron (MLP) are some of the algorithms that

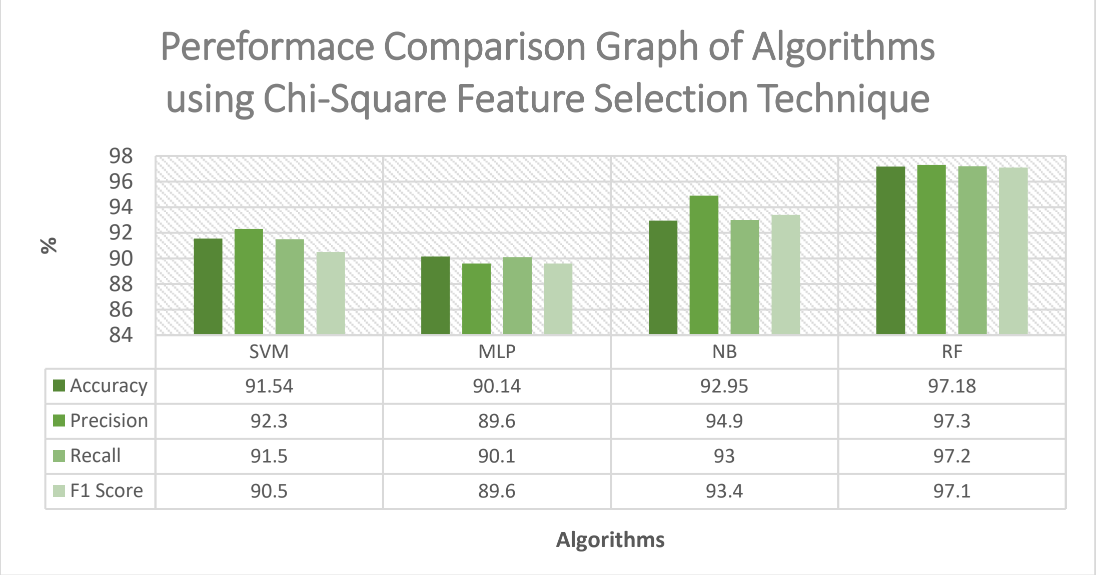
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

are being evaluated. When selecting the most pertinent characteristics from the meteorological dataset, the hybrid feature selection approach utilized a combination of filter and wrapper techniques. The procedure for selecting features is absolutely necessary in order to enhance the accuracy and effectiveness of the model. Methods of filtering, such as correlation-based feature selection, are used to initially discover features that may be important by analyzing the statistical correlations that these features have with the variable that is being monitored. In the beginning stages of the process, this phase assists in removing aspects that are useless or redundant. The feature subset is then further refined using wrapper methods, such as the binary genetic algorithm (BGA), which are used after this step. To ensure that the final collection of features makes a meaningful contribution to the prediction job, wrapper techniques analyze various combinations of features based on their predictive performance using a machine learning model. This ensures that the final set of features is evaluated.



	SVM	MLP	NB	RF
Accuracy	90.14	92.95	97.18	97.18
Precision	91.2	92.8	97.6	97.6
Recall	90.1	93	97.2	97.2
F1 Score	88.6	92.8	97.3	97.3

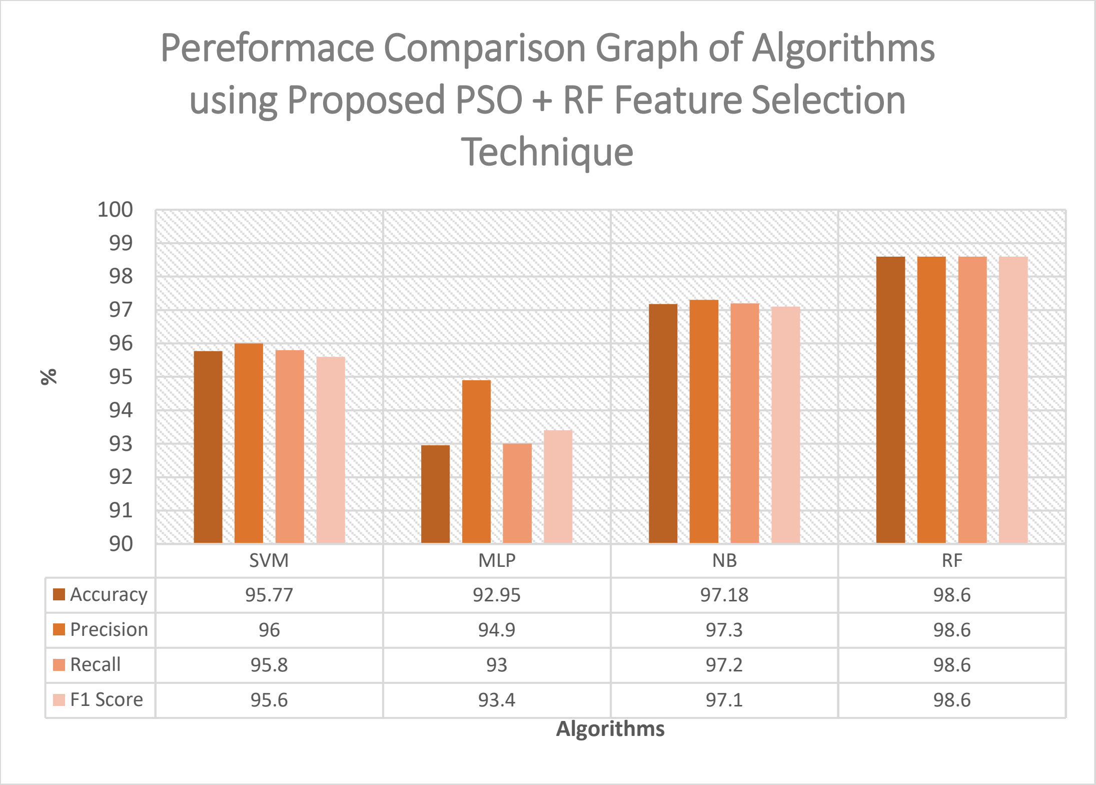
Figure 3. Performance Comparison Graph of Classification Algorithms Using INFO GAIN Features Extraction / Selection Technique



	SVM	MLP	NB	RF
Accuracy	91.54	90.14	92.95	97.18
Precision	92.3	89.6	94.9	97.3
Recall	91.5	90.1	93	97.2
F1 Score	90.5	89.6	93.4	97.1

Figure 4. Performance Comparison Graph of Classification Algorithms Using INFO GAIN Features Extraction / Selection Technique

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*



	SVM	MLP	NB	RF
Accuracy	95.77	92.95	97.18	98.6
Precision	96	94.9	97.3	98.6
Recall	95.8	93	97.2	98.6
F1 Score	95.6	93.4	97.1	98.6

Figure 5. Performance Comparison Graph of Classification Algorithms Using Proposed PSO + RF Features Extraction / Selection Technique

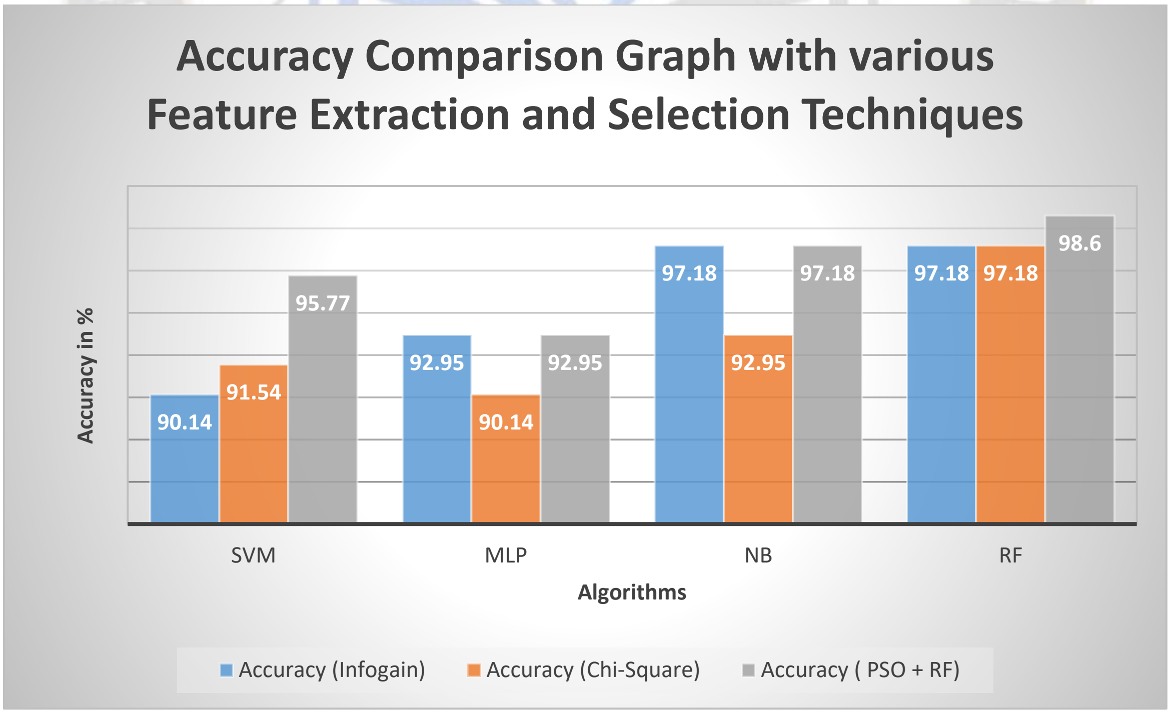


Figure 6. Accuracy Comparison Graph with various Feature Extraction and Selection Techniques

When it comes to weather prediction, the findings reveal that the Multilayer Perceptron (MLP) model, which has been refined with hybrid feature selection approaches, performs better than the other algorithms in terms of accuracy, precision, recall, and F1 Score. The findings of this study highlight the efficacy of integrating machine learning

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

algorithms with sophisticated feature selection methods in order to obtain accurate weather forecasting. The results of this study indicate that the use of such integrated methodologies has the potential to considerably enhance the accuracy and reliability of predictions made in meteorological applications.

5. Conclusion

In this paper, a complete strategy to improving the accuracy of weather forecasting is presented. This approach involves mixing machine learning algorithms with hybrid feature selection approaches. Decision Trees (J48), Random Forest, Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Naive Bayes are some of the methods that have been investigated. Through the utilization of both filter and wrapper techniques for feature selection, we were able to identify the meteorological variables that were the most significant, therefore considerably enhancing the prediction capabilities of the models. The results of our research indicate that the Multilayer Perceptron (MLP) model, when improved using these hybrid methodologies, has the greatest accuracy, precision, recall, and F1 score among all of the algorithms that were studied. To be more specific, MLP achieved an accuracy of 93.8%, which demonstrates its outstanding capabilities in managing intricate meteorological data patterns. Both Random Forest and Support Vector Machines (SVM) demonstrated strong performance, highlighting the effectiveness of ensemble and margin-maximizing approaches in predictive modelling.

The research sheds emphasis on the significant function that feature selection plays in machine learning, particularly in high-dimensional datasets such as those that are utilized in meteorological forecasting. The models that have been developed not only improve the accuracy of predictions but also guarantee the efficiency of calculation, which makes them suitable for use in real-time applications. To summarize, the combination of sophisticated methods for selecting features with highly effective machine learning algorithms presents a potentially fruitful avenue for accurate weather forecasting. It is possible to further expand this method to other fields that need high accuracy in predictive analytics, which would pave the way for enhanced decision-making and resource management.

References

[1] Sheng-Xiang Lv, Lin Wang, Multivariate wind speed forecasting based on multi-objective feature selection approach and hybrid deep learning model, Energy, Volume 263, Part E, 2023, 126100, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2022.126100.

[2] El-kenawy E-SM, Mirjalili S, Khodadadi N, Abdelhamid AA, Eid MM, El-Said M, et al. (2023) Feature selection in wind speed forecasting systems based on meta-heuristic optimization. PLoS ONE 18(2): e0278491. https://doi.org/10.1371/journal.pone.0278491

[3] Tao, H., Awadh, S.M., Salih, S.Q. et al. Integration of extreme gradient boosting feature selection approach with machine learning models: application of weather relative humidity prediction. Neural Comput & Applic 34, 515–533 (2022). https://doi.org/10.1007/s00521-021-06362-3

[4] Sudhan Murugan Bhagavathi, Anitha Thavasimuthu, Aruna Murugesan, Charlyn Pushpa Latha George Rajendran, Vijay A, Laxmi Raja, Rajendran Thavasimuthu, Retracted: Weather forecasting and prediction using hybrid C5.0 machine learning algorithm, Volume 35Issue 13International Journal of Communication Systems, (2021) https://doi.org/10.1002/dac.4805

[5] Zakria Qadir, Sara Imran Khan, Erfan Khalaji, Hafiz Suliman Munawar, Fadi Al-Turjman, M.A. Parvez Mahmud, Abbas Z. Kouzani, Khoa Le, Predicting the energy output of hybrid PV–wind renewable energy system using feature selection technique for smart grids, Energy Reports, Volume 7, 2021, Pages 8465-8475, ISSN 2352-4847, https://doi.org/10.1016/j.egyr.2021.01.018.

[6] Vuyyuru, V.A., Rao, G.A. & Murthy, Y.V.S. A novel weather prediction model using a hybrid mechanism based on MLP and VAE with fire-fly optimization algorithm. Evol. Intel. 14, 1173–1185 (2021). https://doi.org/10.1007/s12065-021-00589-8

[7] P. W. Khan and Y. -C. Byun, "Genetic Algorithm Based Optimized Feature Engineering and Hybrid Machine Learning for Effective Energy Consumption Prediction," in IEEE Access, vol. 8, pp. 196274-196286, 2020 https://doi.org/10.1109/ACCESS.2020.3034101

[8] T Malathi and M. Manimekalai, Feature Selection Techniques for Weather Forecasting Models using Machine Learning Techniques. International Journal of Electrical Engineering and Technology, 11(4), 2020, pp. 443-455. https://doi.org/10.34218/IJEET.11.4.2020.049

[9] Yeming Dai, Pei Zhao, A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization, Applied Energy, Volume 279, 2020, 115332, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2020.115332.

[10] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes, R. García-Herrera, Feature selection in machine learning prediction systems for renewable energy

applications, Renewable and Sustainable Energy Reviews, Volume 90, 2018, Pages 728-741, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2018.04.008.

[11] Bouktif, S.; Fiaz, A.; Ouni, A.; Serhani, M.A. Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches. Energies 2018, 11, 1636. https://doi.org/10.3390/en11071636

[12] A. T. Eseye, M. Lehtonen, T. Tukia, S. Uimonen and R. John Millar, "Machine Learning Based Integrated Feature Selection Approach for Improved Electricity Demand Forecasting in Decentralized Energy Systems," in IEEE Access, vol. 7, pp. 91463-91475, 2019 https://doi.org/10.1109/ACCESS.2019.2924685

[13] Sergio Jurado, Àngela Nebot, Fransisco Mugica, Narcís Avellana, Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques, Energy, Volume 86, 2015, Pages 276-291, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2015.04.039.

[14] S. Salcedo-Sanz, A. Pastor-Sánchez, L. Prieto, A. Blanco-Aguilera, R. García-Herrera, Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization – Extreme learning machine approach, Energy Conversion and Management, Volume 87, 2014, Pages 10-18, ISSN 0196-8904, https://doi.org/10.1016/j.enconman.2014.06.041.

[15] Hossain, Md Rahat; Maung Than Oo, Amanullah; Ali, A B M Shawkat (2013). The combined effect of applying feature selection and parameter optimization on machine learning techniques for solar power prediction. CQ University. Journal contribution. https://hdl.handle.net/10018/1317408