HUF4WP: A Data-Fusion Framework Leveraging High-Utility Patterns for Renewable Energy Classification

M. Saqib Nawaz[a], Philippe Fournier-Viger[a,*], M. Zohaib Nawaz[a,b], Yulin He[c], Unil Yun[d]

[a]College of Computer Science and Software Engineering, Shenzhen University, 518060, Shenzhen, China
[b]Faculty of Computing and IT, Department of Computer Science, University of Sargodha, 40100. Pakistan
[c]Guangdong Laboratory of Artificial Intelligence and Digital Economy, 518107, Shenzhen, China
[d]Department of Computer Engineering, Sejong University, 05006, Seoul, Republic of Korea

Abstract

Accurate classification and forecasting of renewable energy, such as wind and photovoltaic (PV) power, are critical for effective grid management and stable integration of renewable energy sources. However, existing computational methods—particularly time-series-based approaches—often fail to capture complex, latent relationships in renewable energy data and offer limited interpretability. To address these challenges, we propose HUF4WP (High-Utility Framework for Wind/PV power), a high-utility data fusion framework designed to extract and leverage predictive patterns from wind/PV data. HUF4WP transforms continuous energy data into a discretized format and applies high-utility pattern mining algorithms to discover both ordered and unordered patterns, along with high-utility association rules. These patterns are then utilized for wind and solar power classification/prediction by fusing informative feature interactions across various conditions. We evaluate HUF4WP on eight PV and six wind datasets using seven classification models and seven evaluation metrics. Experimental results demonstrate that HUF4WP achieves better classification performance compared to baseline models trained on the raw features. It also provides faster computational performance than baseline and regression-based forecasting approaches. Moreover, the discovered patterns and rules provide interpretable insights into key features and their contributions to renewable energy dynamics.

Keywords: Wind and PV power, High-utility pattern mining, Feature weighting, Renewable energy classification, Discretization

1. Introduction

The global shift toward sustainable energy systems [1] has led to the widespread adoption of renewable sources—particularly wind and photovoltaic (PV) power. These are supported by advanced storage technologies (e.g., batteries, pumped hydro) and smart grids. Renewable energy provides promising alternatives to fossil fuels by providing clean, decentralized, and increasingly cost-effective energy. Recent statistics indicate that renewable sources have supplied a record 32% of global electricity in 2024, up from 30% in 2023, with solar and wind driving the majority

[*] Corresponding author
Email address: philfv@szu.edu.cn (Philippe Fournier-Viger)

of this growth [2]. Furthermore, the International Energy Agency (IEA) projects that renewable energy generation will overtake coal-fired generation by 2025 [3]. Modern energy systems produce large volumes of heterogeneous data from various sources, such as sensors, Internet of Things (IoT) devices, and weather forecasting services. The effective fusion and integration of this multi-source data are critical for computing key grid performance indicators, which are essential for monitoring stability and supporting operational decisions. At the same time, the inherent variability and intermittency of wind/PV power introduce significant challenges for grid stability, operational planning, and energy market operations. This rapid expansion has introduced new complexities in managing and forecasting renewable energy generation.

Addressing these challenges requires sophisticated data-driven methods that can process, fuse, and interpret complex information from diverse renewable energy datasets. Various predictive approaches have been introduced that can be broadly categorized into four frameworks: (1) physical models [4, 5, 6], which rely on meteorological inputs, atmospheric dynamics, and physical principles to simulate wind/PV generation processes; (2) statistical, machine learning (ML), and deep learning (DL) techniques [7, 8, 9, 10, 11, 12, 13], which leverage large amounts of historical data to identify patterns and construct predictive models; (3) meta-heuristic optimization methods, employed to optimize feature selection, model parameters, and ensemble strategies to enhance predictive accuracy and optimize model complexity; and (4) fusion or hybrid methods [14, 15], which combine the strengths of physical, statistical, and heuristic techniques to improve prediction performance (more details are presented in Section 2).

While effective, each of the aforementioned predictive frameworks presents notable limitations in the context of wind/PV power forecasting. Physical models require extensive domain expertise, significant computational resources and often lack robustness [16], making them unsuitable for real-time or large-scale applications. Statistical, ML, and DL methods depend on large volumes of clean historical data. They often struggle to generalize under dynamic weather conditions or when exposed to unseen patterns-especially when trained on imbalanced or noisy datasets [8, 9, 10, 11, 12, 13]. Meta-heuristic methods are useful for optimizing feature selection and model parameters, but they can be computationally intensive, sensitive to hyperparameters and require expert intervention for reliable results [17, 18, 19]. Hybrid and fusion methods often result in overly complex models that are difficult to deploy and maintain in real-world energy systems [14, 15, 20]. Across all four categories, challenges remain in terms of accuracy, generalizability, and scalability, mainly because most methods are evaluated on limited datasets—often confined to a single case study or location.

Pattern discovery techniques—such as frequent itemset mining (FIM) [21] and sequential pattern mining (SPM) [22]—provide a complementary approach to enhance the interpretability of predictive models by identifying frequent patterns and rules in renewable energy datasets. These methods have been applied to uncover recurring sets of features or events in renewable energy data, such as frequent patterns in spatio-temporal wind [23, 24] and solar [25] datasets. However, these methods assume that all features contribute equally to the outcome, neglecting the fact that some features may carry greater importance or relevance than others. For renewable energy forecasting, variables such as wind speed, irradiance, and pressure may vary in significance depending on the operational scenario or forecasting objective. To the best of our knowledge, there is a notable gap in the literature regarding approaches that explicitly incorporate feature utility—the relative importance or contribution of individual energy features—into the pattern mining process, particularly for subsequence analysis and classification.

We propose HUF4WP (High-Utility Framework for Wind/PV), a high-utility data fusion framework for interpretable analysis and classification of wind/PV power. Unlike traditional frequent pattern mining methods, high-utility itemset mining (HUIM) techniques [26] prioritize patterns with higher operational or predictive significance. For example, they emphasize high-impact features like wind speed or irradiance fluctuations. In contrast, high-utility sequential pattern mining (HUSPM) techniques capture ordered patterns. Both ensure that the discovered rules and associations are more relevant for site-specific renewable energy prediction. Furthermore, we incorporate correlated HUIM to identify strongly associated feature value sets that play a critical role in operational decision-making for renewable energy systems. HUF4WP first transforms continuous power data into a discretized format. Then, the importance of each feature is determined using a game theoretic approach. The features with high (low) SHAP (SHapley Additive exPlanations) [27] values are assigned proportional weights. The transformed data, containing features, their values, and assigned utilities, are then processed using various utility mining algorithms to discover high-utility itemsets, high-utility sequential patterns, rules and correlated high-utility itemsets. These frequent high-utility itemsets and sequential patterns are subsequently used to build predictive models that classify/predict wind/PV power. Unlike previous related studies that treat forecasting as a regression problem using continuous outputs, HUF4WP reformulates it as a classification task by discretizing both input features and target variables. This shift enables interpretable, utility-aware modeling and aligns with real-world operational needs where categorical feature levels are often more actionable than precise numerical forecasts. The main contributions of this work are as follows:

- **Discretization of continuous energy data**: A systematic approach is introduced to transform continuous wind/PV data (e.g., wind speed, solar irradiance, temperature, and power output) into categorical representations. This is achieved through binning techniques, where continuous variables are segmented into meaningful intervals based on domain-informed thresholds. The discretized format enhances interpretability, reduces noise sensitivity, and facilitates the application of high-utility pattern mining algorithms.

- **High-Utility Driven Pattern Discovery Framework**: A data fusion framework is developed that integrates multiple high-utility pattern mining techniques to extract interpretable and operationally relevant knowledge from discretized energy datasets. By incorporating utility scores derived from SHAP values, the framework prioritizes feature-value combinations that hold the highest influence on energy outcomes. HUIM identifies co-occurring feature sets with high cumulative utility, while HUSPM captures ordered dependencies critical for understanding energy behavior over time. Correlated HUIM further refines these insights by isolating strongly associated feature-value subsets. Overall, these methods ensure that the discovered patterns are frequent and highly impactful, supporting the development of robust and explainable predictive models for wind/PV power.

- **Discovered Patterns in Classification**: The obtained frequent and high-utility patterns are employed as features for supervised learning tasks. It is found that embedding the patterns into the seven classification models yields improved results while maintaining model interpretability.

- **Experimental Evaluation**: Extensive experiments are performed on eight PV and six wind datasets to evaluate the developed framework using various standard evaluation metrics. The performance of HUF4WP is compared against baseline classification models trained on raw features. Its computational efficiency is also contrasted with that of regression-based forecasting approaches.

The rest of this paper is organized as follows: Section 2 provides a literature review on methods for wind/PV power prediction. Section 3 describes the HUF4WP framework in detail, including the eight PV and six wind datasets, the data preprocessing pipeline, utility-driven pattern mining techniques, and classification strategy. Section 4 presents and discusses the experimental results, including comparative evaluations of HUF4WP. Finally, Section 5 concludes the paper with a summary of findings and directions for future research.

2. Literature Review

Numerous studies have focused on developing computational models for wind/PV power prediction/forecasting. These approaches can be broadly classified into four categories: physical model-based approaches, (2) statistical, ML and DL-based methods, (3) meta-heuristic/ optimization-based techniques, and (4) hybrid approaches. Each category is discussed next.

Physical-based approaches utilize theoretical simulation models to estimate the output power based on well-established physical principles and numerical weather prediction (NWP) data [5, 6]. These methods typically involve multi-step model chains that reflect the physical characteristics of energy systems. For instance, Yang et al. [28] employed a multi-step model chain including irradiance decomposition, transposition from horizontal to tilted surfaces, and PV module performance modeling. Similarly, Lorenz et al. [29] incorporated various models—PEREZ for irradiance transposition, LINEAR for temperature, BEYER for PV performance, and QUADRATIC for inverter behavior. Subsequent studies have refined these pipelines with additional modeling components. For example, Wolff et al. [30] proposed a five-step simulation pipeline combining SKARTVEIT-OLSETH for irradiance separation, KLUCHER for transposition, LINEAR for temperature, BEYER for performance, and QUADRATIC for inverter behavior modeling. Saint-Drenan et al. [31] added the MARTIN-RUIZ angular loss model to the previously outlined steps. Other configurations included DIRINT separation with in-house PV simulations [32] and ENGERER separation with MARTIN-RUIZ loss modeling [33]. A detailed physical model presented in [34] considered spectral, shading, and inverter losses using a combination of ERBS, FAIMAN, EVANS, and QUADRATIC models. A large-scale benchmarking study [4] compared the forecasting accuracy of 32,400 model configurations. These involved various combinations of transposition, separation, temperature, reflection, inverter, and shading models using NWP data. Despite their theoretical rigor, physical models often exhibit rigidity and dependency on accurate input data. To address some of these limitations, statistical models have been introduced as more flexible alternatives that leverage historical data patterns.

Statistical approaches are used due to their simplicity, adaptability, and relatively low computational cost. They rely on historical data to capture the stochastic relationships between meteorological inputs (such as wind speed, solar irradiance, temperature) and power output. Traditional time series forecasting techniques—such as autoregressive integrated moving average (ARIMA) and its variant (ARIMA-GARCH)—were used for power forecasting [35, 36, 37]. Methods like Holt-Winters exponential smoothing [38, 39] are also used to model seasonality and trend in solar/wind generation. Probabilistic models, including probability mass bias [40], probabilistic auto-regression [41], probabilistic forecasting method [11] and Bayesian framework [42] are designed to model uncertainty. ML techniques go a step further by learning complex, non-linear patterns from historical data without relying on explicit physical formulations. Standard ML classifiers such as Support Vector Machines (SVM) [43], Random Forests (RF) and Gradient Boosting Machines (GBMs) [44], k-Nearest Neighbors (k-NN) [45], Decision tree (DT) [46], Extreme Gradient Boosting (XGBoost) [46] and Gradient Boosted Regression Trees (GBRT) [46] were used in forecasting. Several studies [47, 48, 49] have shown that ML models generally outperform physical and statistical approaches in wind/PV power forecasting. For example, Biswas et al. [49] demonstrated that RF consistently outperformed ARIMA across multiple forecasting scenarios and datasets. However, both statistical and ML models face several limitations. They rely heavily on large volumes of cleaned historical data and often struggle to generalize under rapidly changing or chaotic weather conditions. These models also have limited interpretability, especially in the case of complex ML architectures. Additionally, they typically perform less effectively when fusing multi-source or heterogeneous datasets.

On the other hand, DL approaches can effectively capture complex, nonlinear temporal and spatial patterns in energy datasets. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [50] and Convolutional Neural Networks (CNNs) [51, 52], were used for modeling sequential dependencies in wind/PV forecasting. Gated Recurrent Units (GRUs) [53] offer comparable performance but with reduced computational overhead. CNN-LSTM architectures [54, 55] leverage the strengths of both models, CNNs for feature extraction and LSTMs for sequence modeling. Conv-ELSTM [13] is a convolutional ensemble LSTM network for wind energy prediction. Recently, transformer models [12, 56], which rely on attention mechanisms, have gained attention for their scalability and ability to model long-range dependencies in time-series data. Additionally, autoencoders [57] and variational autoencoders (VAEs) [58] have been employed for wind power prediction. Graph-based DL methods—such as CGAE (Convolutional Graph Autoencoder network) [59], GCLSTM (Graph-Convolutional LSTM) and GCTrafo (Graph-Convolutional Transformer) [60]—were proposed for solar irradiance and PV power forecasting. Despite their high accuracy, DL models face several challenges. They require large amounts of high-quality training data, demand significant computational resources and lack interpretability, which can hinder their deployment in real-world operational environments.

Meta-heuristic and optimization algorithms have been widely applied to enhance forecasting models for wind/PV power prediction. Genetic Algorithms (GA) [61] have been used to optimize feature selection in wind speed forecasting models, improving neural network performance. Particle Swarm Optimization (PSO) [62] has been combined with SVM for short-term solar power prediction, yielding better accuracy when integrated with wavelet transforms. Ant Colony Optimization (ACO) [63] has been applied for input selection in solar radiation models, leading to more effective forecasting. Simulated Annealing (SA) [64] has been used for parameter tuning in PV systems, improving simulation fidelity. More recently, the Grey Wolf Optimizer (GWO) [65] has been adopted to fine-tune the hyperparameters of LSTM networks for wind energy forecasting. Yu et al. [18] employed a multi-scale clustering ensemble, similarity matching, and an improved whale optimization algorithm for wind power prediction. They offer flexible and powerful optimization capabilities that significantly enhance the performance of traditional and DL-based forecasting models [19, 17, 66]. Meta-heuristic-based approaches face several limitations. These include high computational cost, sensitivity to selecting hyperparameters and a risk of overfitting due to excessive optimization on training data. They may also suffer from poor generalization to unseen scenarios, face scalability issues as the feature space expands, and often lack inherent interpretability—an increasingly important factor in energy forecasting applications.

Hybrid models have emerged as an effective strategy for wind/PV power forecasting by combining the strengths of the aforementioned approaches. For instance, physical-statistical hybrid models integrated meteorological simulations with statistical corrections for more precise day-ahead PV predictions [67]. Other hybrid methods use meta-heuristic algorithms such as improved whale optimization to fine-tune ML parameters, improving convergence and reducing forecasting error [68]. Ensemble learning approaches further enhance robustness by integrating DL architectures such as LSTM, BiLSTM, and GRU, often optimized using advanced algorithms [69]. Additional hybrid studies for wind/PV forecasting include the following: Li et al. [10] proposed a model that integrated the capture optimization algorithm (CFOA), CNN, BiLSTM, and attention mechanism. A hybrid DL-based neural network [15] comprised a CNN followed by a RBFNN with a double Gaussian function (DGF) as its activation function. The forecasting model [16] was based on weather type, AHA-VMD-MPE decomposition reconstruction, and an improved informer combination. MF-NBEA [9] introduced a multilevel data fusion and neural basis analysis. The model [8] fused sky condition and feature-source information using a multi-task DL architecture based on RNN. Khan et al. [20] employed a hybrid DL model, consisting of CNN, LSTM and a Bi-LSTM. Mirza et al. [14] proposed IEDN-RNET, an inception-embedded deep neural network with ResNet for short/medium-term wind/PV forecasting. While hybrid models demonstrate superior performance across various forecasting tasks, they often entail increased system complexity and reduced model transparency.

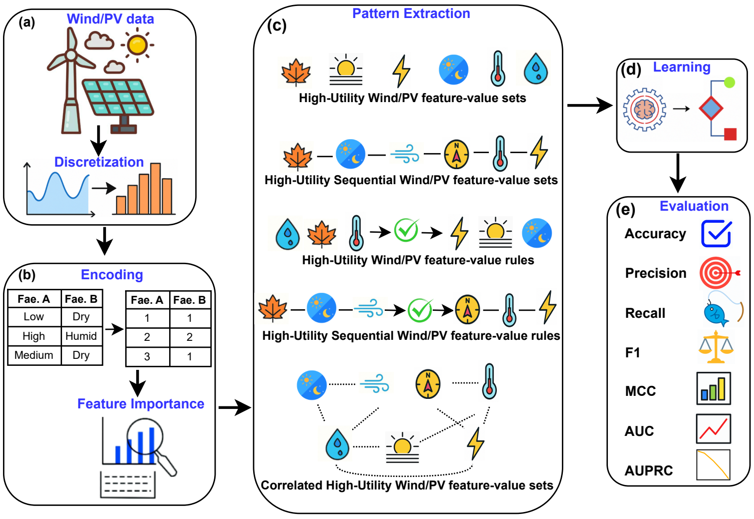
Pattern-based approaches have also been developed in the past. For example, an SPM-based approach [23] simultaneously analyzed wind speed and direction, combining it with novel visualizations to better understand wind behavior over time. The approach of [24] identified and mapped frequent wind profile patterns across space, time and height using multi-dimensional SPM, with a focus on optimizing wind energy harvesting. The Stcop-Miner framework [25] discovered spatiotemporal co-occurrence patterns in large-scale solar event datasets using specialized indexing techniques. Previous works—ranging from statistical techniques to advanced ML, DL and hybrid—focused largely on time-series forecasting using regression-based models to predict continuous values for the dependent features, which is power in wind/PV studies. The performance of models was evaluated through metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). This study adopts a fundamentally different approach by reframing the forecasting problem into a classification task by discretizing the features. This transformation enables not only pattern-based modeling through high-utility pattern mining but also shifts the objective from predicting exact numeric values to identifying characteristic patterns across wind/PV features. Consequently, our evaluation is based on classification metrics such as the Accuracy, Precision, Recall, F1-score, etc.

3. Framework

The HUF4WP framework (Figure 1) consists of five stages: (1) Energy data acquisition—collection and curation of wind/PV records. (2) Discretization and feature weighting—continuous wind/PV features and their values are transformed into an appropriate categorical representation, and the importance (weight) of each feature is determined to quantify its influence. (3) Pattern extraction—mining of (a) high-utility-based co-occurring patterns of wind/PV feature values (unordered) and sequential (ordered) patterns, (b) high-utility association rules for both ordered and unordered wind/PV feature values, and (c) correlated wind/PV feature value sets. (4) Classification—training interpretable models using the extracted co-occurring and sequential patterns as discriminative feature values. (5) Evaluation—performance assessment using standard metrics.

3.1. Wind/PV Datasets

The proposed HUF4WP framework is evaluated on eight PV datasets—abbreviated as PVD1 through PVD8—and six wind datasets—abbreviated as WD1 through WD6. These datasets were


(a)
Wind/PV data
Discretization
(b)
Encoding
Feature Importance
(c)
Pattern Extraction
High-Utility Wind/PV feature-value sets
High-Utility Sequential Wind/PV feature-value sets
High-Utility Wind/PV feature-value rules
High-Utility Sequential Wind/PV feature-value rules
Correlated High-Utility Wind/PV feature-value sets
(d)
Learning
(e)
Evaluation
Accuracy
Precision
Recall
F1
MCC
AUC
AUPRC

Fae. A	Fae. B
Low	Dry
High	Humid
Medium	Dry

Fae. A	Fae. B
1	1
2	2
3	1

Figure 1: Overview of the HUF4WP framework for wind/PV data analysis and classification. It includes five stages: (1) Datasets collection, (2) Feature discretization and weighting, (3) Discovery of (a) high-utility wind/PV feature value sets and sequential patterns of wind/PV feature values, (b) high-utility association rules, and (c) correlated high-utility wind/PV feature sets, (4) Classification by models training using the discovered high-utility feature value sets and sequential patterns along with their associated values, and (5) Framework evaluation using multiple metrics.

obtained from the Chinese State Grid [71]. They include weather-related and power-generation data gathered at 15-minute intervals over a two-year period (2019-2020). The PV and wind datasets contain 8 and 13 features, respectively (Table 1). Dataset characteristics including missing values, mean, maximum and standard deviation for each feature are provided in [71]. All six wind datasets contain 70,176 samples each (total 421,056), while six PV datasets (PVD1, PVD2, PVD4, PVD5, PVD6, and PVD7) contain 70,176 samples and PVD3 and PVD8 contain 52,608 and 69,408 samples, respectively (total 516,072 samples).

Table 1: Features in the wind/PV datasets, with the target features underlined

PV datasets (PVDs)
(1) Time (year-month-day h:m:s), (2) Total Solar Irradiance (W/m^2), (3) Direct Normal Irradiance (W/m^2), (4) Global Horizontal Irradiance (W/m^2), (5) Air Temperature (°C), (6) Atmospheric Pressure (hPa), (7) Relative Humidity (%), (8) Power (MW)
Wind datasets (WDs)
(1) Time (year-month-day h:m:s), (2) Wind Speed at 10 m Height (m/s), (3) Wind Direction at 10 m Height (°), (4) Wind Speed at 30 m Height (m/s), (5) Wind Direction at 30 m Height (°), (6) Wind Speed at 50 m Height (m/s), (7) Wind Direction at 50 m Height (°), (8) Wind Speed at Wheel Hub Height (m/s), (9) Wind Direction at Wheel Hub Height (°), (10) Air Temperature (°C), (11) Atmospheric Pressure (hPa), (12) Relative Humidity (%), (13) Power (MW)

To better capture temporal patterns relevant to energy generation, the original *Time* feature—recorded as a date and 15-minute timestamp—was transformed into two categorical variables, namely *Season* and *Day/Night*. This transformation enhances model interpretability and robustness by encoding essential periodic dynamics. For instance, seasonality captures recurring trends, such as higher PV output in summer and peak wind production in winter, while day/night segmentation reflects the binary nature of solar generation. This process reduces dimensionality, mitigates overfitting risks associated with high-cardinality timestamps, and inherently addresses periodicity, thereby improving generalization across regions/years. Although some temporal granularity is sacrificed, the transformation enhances interpretability and facilitates intuitive modeling of interactions. It is important to point out here that some works such as [15] and [46] divided the datasets into sub-datasets based on seasons and weather conditions, respectively.

The datasets contain various forms of missing or invalid entries (zero, null, ‘NA’, ‘0.001’, ‘-99’, and ‘-’), which were handled using a multi-step cleaning pipeline. This pipeline includes time-based interpolation, median imputation grouped by month and hour, and forward/backward filling. Additional domain-specific logic was also applied, such as replacing negative values, handling zero irradiance during daylight, and clipping outliers to stay within valid physical bounds.

To discover interpretable patterns and reduce sensitivity to noise, all continuous features in the wind/PV datasets were transformed into categories (bins) based on domain-informed discretization rules. Table 2 details the discretization categories and ranges applied to each feature in the wind/PV datasets. For instance, solar irradiance was discretized into levels such as *Low*, *Medium*, *High*, and *Very High*, while wind speed at multiple sensor heights were mapped to qualitative ranges (e.g., *Calm*, *Moderate*, *Strong*). This transformation enables the application of high-utility pattern mining algorithms, which require symbolic inputs. Moreover, discretization also improves model robustness, reduces the risk of overfitting, and enhances the interpretability of discovered rules for renewable energy forecasting.

Table 2: Discretization categories and corresponding ranges for features in WDs and PVDs

Dataset(s)	**Feature**	**Category**	**Range**	**Dataset(s)**	**Feature**	**Category**	**Range**
PV	Solar Irradiance	Low	$v \leq 400$	PV/Wind	Season	Winter	Dec, Jan, Feb
		Medium	$400 < v \leq 700$			Spring	Mar, Apr, May
		High	$700 < v \leq 1000$			Summer	Jun, Jul, Aug
		Very High	$v > 1000$			Autumn	Sep, Oct, Nov
PV/Wind	Pressure	Low	$v \leq 950$	PV/Wind	Humidity	Dry	$v \leq 30$
		Medium	$950 < v \leq 980$			Comfortable	$30 < v \leq 60$
		High	$980 < v \leq 1000$			Humid	$60 < v \leq 80$
		Very High	$v > 1000$			Very Humid	$v > 80$
PV/Wind	Power	Low	$v \leq 5$				
		Medium	$5 < v \leq 15$				
		High	$15 < v \leq 25$				
		Very High	$v > 25$				
Wind	Wind Direction	N-E	$v \leq 90$	PV/Wind	Day/Night	Day	$06{:}00 \leq v < 18{:}00$
		E-S	$90 < v \leq 180$				
		S-W	$180 < v \leq 270$			Night	$18{:}00 \leq v < 06{:}00$
		W-N	$270 < v \leq 360$				
PV/Wind	Temperature	Freezing	$v \leq 0$	Wind	Wind Speed	Calm	$v \leq 2$
		Cold	$0 < v \leq 10$			Breeze	$2 < v \leq 5$
		Mild	$10 < v \leq 20$			Moderate	$5 < v \leq 8$
		Warm	$20 < v \leq 30$			Strong	$8 < v \leq 12$
		High	$v > 30$			Gale	$8 < v \leq 12$

The discretization categories were designed to be consistent for semantically similar features. For instance, the solar irradiance label covers three distinct but related variables—*Total Solar Irradiance*, *Direct Normal Irradiance*, and *Global Horizontal Irradiance*—as they all measure radiative solar input and exhibit comparable physical behavior and value ranges. Similarly, wind speed and wind direction measurements, recorded at multiple altitudes (10 m, 30 m, 50 m, and hub height), were discretized using the same binning strategy due to their consistent semantics and magnitude scales. This uniform categorization enables the application of unified pattern mining models and facilitates interpretability across different sensor heights. The distribution of the discretized power feature (the dependent variable) across six wind and eight PV datasets is presented in Table 3. This distribution highlights the presence of class imbalance in several datasets, particularly the underrepresentation of the *Very High* power class in some PVDs.

Table 3: Distribution of discretized power output classes for the PV and Wind datasets

Dataset	**Low**	**Medium**	**High**	**Very High**
PVD1	42,844	7,212	7,018	13,102
PVD2	17,722	17,386	17,524	17,544
PVD3	36,452	7,526	7,033	1,597
PVD4	42,828	6,949	3,835	16,564
PVD5	43,309	7,234	3,963	15,670
PVD6	45,573	9,832	9,958	4,813
PVD7	47,059	10,995	10,878	1,244
PVD8	49,332	12,484	7,454	138
WD1	17,545	17,544	17,543	17,544
WD2	17,544	17,544	17,544	17,544
WD3	17,558	17,530	17,546	17,542
WD4	17,545	17,543	17,544	17,544
WD5	46,794	11,310	4,764	7,308
WD6	17,539	17,539	17,535	17,538

3.2. Feature Encoding and Weighting

In this step, the features of PVDs and WDs are transformed into a standardized integer-based format [70]. During this transformation, we observed that some discretized feature values—such as *Low* or *High*—could appear among multiple features (e.g., irradiance, temperature, power), leading to ambiguity. To avoid this, each feature is encoded such that its values are mapped to unique, disjoint integers.

This encoding ensures that identical categorical labels from different features do not overlap during the pattern mining process. For example, the label *Low* from the temperature feature and *Low* from the solar irradiance feature are assigned distinct integer representations. Fig. 2(a) shows a sample of the original PV dataset, while Fig. 2(b) and (c) represent its discretized and encoded versions, respectively. This example will serve as a running example throughout the paper.

Formally, let $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ denote the set of features in PVDs and WDs. Each feature f_i is associated with a value set $VS(f_i)$ representing its discretized categories. These value sets are mapped to unique integers satisfying the following condition:

$$\forall f_i, f_j \in \mathcal{F}, f_i \neq f_j \Rightarrow VS(f_i) \cap VS(f_j) = \emptyset$$

The union of all encoded values forms the domain $\mathcal{V} = \bigcup_{f \in \mathcal{F}} VS(f)$.

An encoded dataset is denoted by $\mathcal{D} = \{R_1, R_2, \ldots, R_m\}$, where each record R_x is a total function $R_x : \mathcal{F} \rightarrow \mathcal{V}$ mapping every feature to one of its valid encoded values. For any record R_x, the set of its values is:

$$\text{values}(R_x) = \{v_j \mid (f_j, v_j) \in R_x\}$$

The ordered sequence of values is denoted as $\overline{R_x} = \langle x_1, x_2, ..., x_n \rangle$. For a subset $X \subseteq \mathcal{V}$, the set of contributing features is features$(X) = \{f \mid v \in X \cap VS(f)\}$.

To measure feature importance for forecasting renewable energy, we use SHAP values, which are derived using a linear regression model. SHAP provides a unified framework for quantifying each feature's contribution to the output of a predictive model. The choice of a linear model is motivated by the goals of interpretability and computational efficiency. Since SHAP values are not used for direct prediction but rather for assigning utility scores to discretized feature values, a linear model offers consistent and transparent importance estimates. This facilitates a clear mapping between feature relevance and utility-based pattern discovery, thereby enhancing the interpretability of extracted patterns. Furthermore, linear models reduce the risk of overfitting and keep computational overhead low, which is especially important given the symbolic and pattern-based nature of the downstream analysis.

Let ϕ_i denote the mean absolute SHAP value of feature f_i, which quantifies its global contribution to model performance. A normalized utility score $u(f_i) \in [1, K]$ is assigned to each feature f_i based on a scaling factor K (e.g., $K = 10$):

$$u(f_i) = \left\lceil K \cdot \frac{\phi_i}{\max(\phi_1, \ldots, \phi_n)} \right\rceil \tag{1}$$

Fig. 2(d) displays SHAP-derived utility scores for all features from the PV dataset. Based on this weighting, the local utility of a value set $X \subseteq \mathcal{V}$ for a record R_x, such that $X \subseteq \text{values}(R_x)$ is:

$$u(X, R_x) = \sum_{f_j \in \text{features}(X)} u(f_j) \tag{2}$$

The global utility of a value set X in a dataset $\mathcal{D}$ is defined as:

$$u(X) = \sum_{\substack{R_x \in \mathcal{D} \\ X \subseteq \text{values}(R_x)}} u(X, R_x) \tag{3}$$

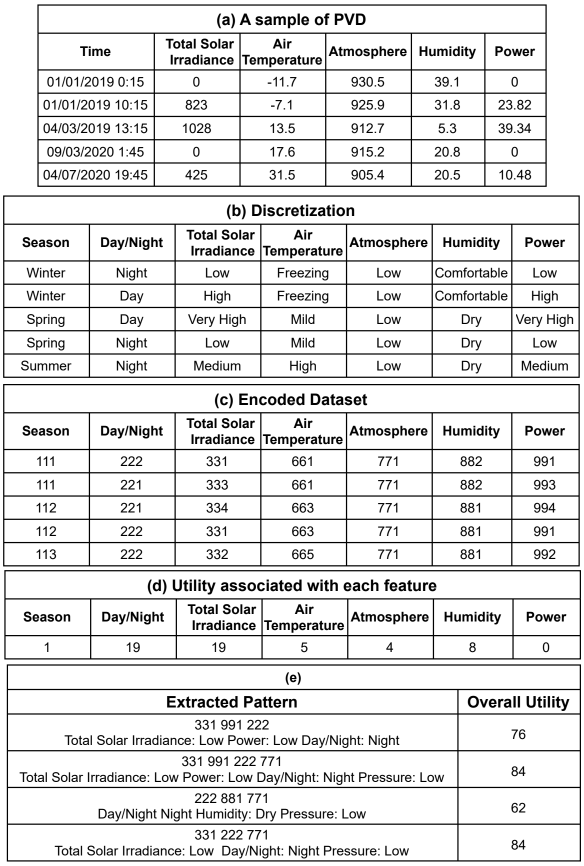
This global utility score $u(X)$ quantifies the predictive relevance of X in the context of wind and solar energy forecasting, by combining both the frequency and utility of the associated features.

3.3. *Pattern Extraction*

After transforming and encoding the WDs and PVDs into a standardized format, we proceed to extract multiple types of high-utility patterns.

High-utility itemsets. We apply HUIM to identify sets of meteorological and power-related features that exhibit high predictive utility. Each 15-minute observation from a wind or PV station is treated as a "transaction" containing discretized feature values such as temperature levels, wind speed categories, irradiance bins, and corresponding power outputs. Utilities are assigned to features according to Equation 3. For a user-defined minimum utility threshold (*minutil*), a feature value set (FVS) is considered a high-utility itemset if $u(FVS) \geq minutil$. Fig. 2(e) illustrates sample high-utility itemsets extracted from the PVD with $minutil = 60$.

High-utility association rules. We also extract association rules from the high-utility itemsets. An association rule $R : X \rightarrow Y$ indicates that the presence of a set of feature values X (e.g., *Summer*, *High* irradiance) strongly suggests the presence of another disjoint set Y (e.g., *Very High* power). The support and confidence of a rule are defined in standard form: $sup(R) = sup(X \cup Y)$ and $conf(R) = \frac{sup(X \cup Y)}{sup(X)}$. In addition, the utility confidence is calculated using SHAP-derived weights as $uconf(R) = \frac{u(X \cup Y)}{u(X)}$. A rule is considered a high-utility association rule if both X and $X \cup Y$ are high-utility itemsets and $uconf(R) \geq minconf$.



(a) A sample of PVD

Time	Total Solar Irradiance	Air Temperature	Atmosphere	Humidity	Power
01/01/2019 0:15	0	-11.7	930.5	39.1	0
01/01/2019 10:15	823	-7.1	925.9	31.8	23.82
04/03/2019 13:15	1028	13.5	912.7	5.3	39.34
09/03/2020 1:45	0	17.6	915.2	20.8	0
04/07/2020 19:45	425	31.5	905.4	20.5	10.48

(b) Discretization

Season	Day/Night	Total Solar Irradiance	Air Temperature	Atmosphere	Humidity	Power
Winter	Night	Low	Freezing	Low	Comfortable	Low
Winter	Day	High	Freezing	Low	Comfortable	High
Spring	Day	Very High	Mild	Low	Dry	Very High
Spring	Night	Low	Mild	Low	Dry	Low
Summer	Night	Medium	High	Low	Dry	Medium

(c) Encoded Dataset

Season	Day/Night	Total Solar Irradiance	Air Temperature	Atmosphere	Humidity	Power
111	222	331	661	771	882	991
111	221	333	661	771	882	993
112	221	334	663	771	881	994
112	222	331	663	771	881	991
113	222	332	665	771	881	992

(d) Utility associated with each feature

Season	Day/Night	Total Solar Irradiance	Air Temperature	Atmosphere	Humidity	Power
1	19	19	5	4	8	0

(e)

Extracted Pattern	Overall Utility
331 991 222 Total Solar Irradiance: Low Power: Low Day/Night: Night	76
331 991 222 771 Total Solar Irradiance: Low Power: Low Day/Night: Night Pressure: Low	84
222 881 771 Day/Night Night Humidity: Dry Pressure: Low	62
331 222 771 Total Solar Irradiance: Low Day/Night: Night Pressure: Low	84

Figure 2: The process of discretizing, transforming, and discovering high-utility itemsets from wind/PV records. (a) A sample of a PVD, (b) its discretization, (c) its encoding, (d) SHAP-based feature utilities, and (e) extracted patterns from the sample PVD for a *minutil* threshold of 60.

High-utility sequential patterns and sequential rules. Given the temporally ordered nature of the datasets, we also extract high-utility sequential patterns. A pattern $P = \langle y_1, y_2, ..., y_n \rangle$ is a time-ordered sequence of feature values (e.g., *Night* → *Strong* wind → *Very High* power). A pattern P is considered high-utility if $u(P) \geq minutil$.

A sequential rule $r : Y \Rightarrow Z$ is an implication between two disjoint feature value sets where Z occurs after Y within the same observation sequence. For example, the rule *Winter* ⇒ *High* wind may capture seasonal wind dynamics. Formally, the rule r occurs in a record $\overline{C_x}$ if there exists an index k such that Y appears in the prefix and Z in the suffix of the sequence. Support and confidence are computed over the ordered sequences, and the utility of r is:

$$u(R) = \sum_{\substack{C_x \in \mathcal{CD} \\ R \sqsubseteq \overline{C_x}}} u(X, C_x)$$

Correlated high-utility itemsets. We additionally extract correlated high-utility itemsets to discover frequent, informative combinations with strong internal coherence. An itemset is considered correlated if it satisfies both high-utility and correlation constraints. The correlation is quantified using the *bond* measure:

$$bond(FVS) = \frac{sup(FVS)}{dissup(FVS)}$$

where $dissup(FVS)$ is the disjunctive support—the fraction of records containing at least one item from the set.

To mine these patterns, we employ the EFIM [72] and HGB [73] algorithms for discovering high-utility itemsets and association rules, and USPAN [74] and HUSRM [75] for sequential patterns and rules. Correlated high-utility itemsets are extracted using the FCHM algorithm [76], all adapted for use in renewable energy datasets.

3.4. Classifier Training and Evaluation

In this stage, the frequently occurring patterns extracted in discretized PVDs and WDs are employed for the classification/prediction of power (denoted as *P*), which consist of four categories: $P \in \{$ *Low, Medium, High, Very High*$\}$ (see Table 2). Five classification tasks are defined to comprehensively analyze the predictive models: four binary classification problems—one for each power category—and one multiclass classification encompassing all categories.

Binary classification is used to individually predict each power category. For a given power level $l \in \{$ *Low, Medium, High, Very High*$\}$, a binary classification task is formulated by labeling records as belonging to the specific level l or to other categories. This is mathematically defined as:

$$P_l = \begin{cases} 1, & \text{if } P = l \\ 0, & \text{if } P \in \{Low, Medium, High, VeryHigh\} \setminus \{l\} \end{cases} \tag{4}$$

In this formulation, the positive class ($P_l = 1$) includes all records corresponding to the power level l, while the negative class ($P_l = 0$) includes records from the remaining power categories. For example, when focusing on the *Low* power level, records labeled as *Low* are considered positive, and records from *Medium*, *High*, and *Very High* are considered as *Others* and labeled negative.

In the multiclass (MC) classification setting, each record is labeled with its respective power category from the four possible levels. The objective here is to develop a model that can accurately classify each record into one of the four categories: *Low, Medium, High*, or *Very High*. This task evaluates the model's ability to distinguish between all levels simultaneously. The whole approach allows for a detailed evaluation of each class through binary classification while also assessing the model's ability to handle the complexity of MC prediction.

For classification tasks, seven widely used ML models are evaluated: (1) Gaussian Naive Bayes (GNB), (2) Decision Tree (DT), (3) Random Forest (RF), (4) Multi Layer Perceptron (MLP), (5) Support Vector Machine (SVM), (6) k-Nearest Neighbors (kNN), and (7) Logistic Regression (LR). Seven performance metrics are employed to assess classifiers effectiveness, which are: (1) Accuracy (ACC), the ratio of correctly predicted instances to the total instances; (2) Recall (R), the model's ability to identify all relevant positive instances; (3) Precision (P), the proportion of correctly predicted positive instances among all predicted positives, (4) F1-Score (F1), the harmonic mean of precision and recall; (5) Matthews Correlation Coefficient (MCC), accounts for true and false positives and negatives; (6) Area Under the Curve (AUC), reflects the model's ability to distinguish between classes across different classification thresholds; and

(7) Area Under the Precision-Recall Curve (AUPRC) evaluates the trade-off between precision and recall for different classification thresholds,. These metrics are used because they offer a comprehensive evaluation of classifier performance, and are defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{6}$$

$$Precision(P) = \frac{TP}{TP + FP} \tag{7}$$

$$F - measure = 2 \times \frac{P \times R}{P + R} \tag{8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

$$\mathrm{AUC} = \int_0^1 \mathrm{R}(d\mathrm{FPR}) \tag{10}$$

$$AUPRC = \sum_{i=1}^{n} \frac{(R_i - R_{i-1}) \times (P_i + P_{i-1})}{2} \tag{11}$$

The terms TP, FP, TN, and FN stand for true positive, false positive, true negative and false negative, respectively. $dFPR$ is for the derivative of $FPR = \frac{FP}{FP+TN}$. P_i and R_i in equation 11 represent the values for P and R, respectively, at the i-th decision threshold. The following section details the application of the methodology and the obtained results.

4. Experimental Evaluation

A computing system with 16 GB of RAM and an Intel Core i5-11320H 3.20 GHz processor was utilized to conduct experiments. The SPMF library [77], developed in Java, was employed to extract patterns from the abstracted PVDs and WDs. Several algorithms have been used for the analysis and discovery of patterns, including EFIM, HGB, USPAN, HUSRM, and FCHM. To perform classification, Python was used, employing a variety of libraries, including scikit-learn [78] for ML algorithms, NumPy for numerical computations, and Pandas for data manipulation.

4.1. Important Features

SHAP [27] was first utilized to identify the most influential features for wind/PV power prediction. A linear regression model was first trained on PVDs and WDs, after which SHAP was applied to interpret its predictions. SHAP offers a unified framework to measure the contribution of each feature to the model's output, providing both local and global interpretability.

Figure 3 presents the SHAP summary plots for the combined PVDs and WDs. For the PVDs, the three most influential features were *Day/Night*, *Total Irradiance*, and *Humidity*. Among these features, *Day/Night* emerged as the single most important predictor. This result aligns with the physical properties of solar energy generation, which only occurs during daylight. The binary nature of this feature enables the model to sharply distinguish between energy-producing (day) and non-producing (night) periods. Compared to irradiance values—which may vary even during the day due to cloud cover or angle of incidence—the *Day/Night* feature provides a stable and

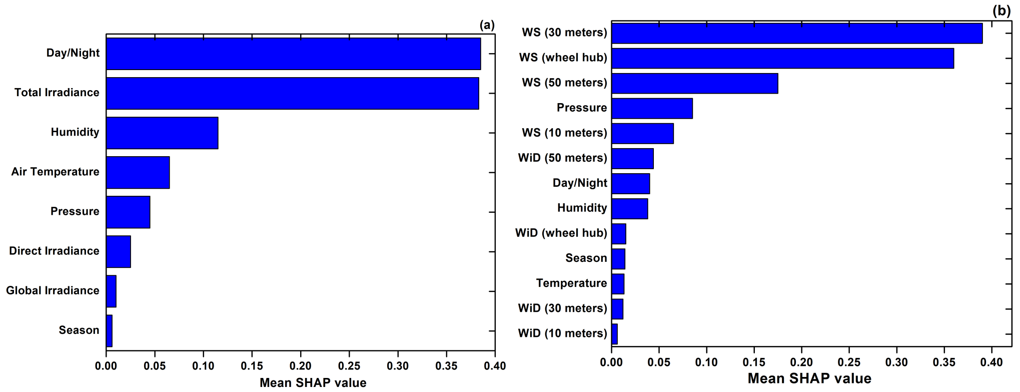


Figure 3: SHAP feature importance for (a) PVDs and (b) WDs. WS and WiD denote wind speed and wind direction, respectively.

highly informative signal, contributing significantly to the model's accuracy. Similarly, for the WDs, the three influential features that contribute significantly to wind power are *Wind Speed at the Height of 30 Meters*, *Wind Speed at the Height of Wheel Hub* and *Wind Speed at 50 Meters*. These results are intuitive, as wind power generation depends directly on wind velocity, particularly at turbine-relevant altitudes. The dominance of wind speed measurements at various heights also underscores the importance of multi-level wind profiling for accurate power prediction. The close SHAP values among these wind speed features indicate some redundancy, yet also reflect the robustness of wind data collected across turbine-relevant altitudes. Interestingly, the *Season* feature showed the lowest SHAP value among all PV features and was ranked 10th out of 13 features in the WDs. This suggests that while seasonality may capture broader trends in weather conditions, it does not provide sufficiently granular information for short-interval (15-minute or 30-minute) forecasting.

SHAP values also not only highlight individual feature contributions but also uncover valuable insights into the intricate interactions between features. For example, while *Day/Night* was identified as a highly influential individual feature, interactions involving *Total Irradiance* and *Humidity* often exhibit synergistic effects that enhance model predictions when considered jointly. This synergy suggests that during the daytime, the level of irradiance influences the expected power output, with cloudy or clear conditions making a significant difference. Previously, a DT model using MSE was used in [46] for feature importance. The most important feature in one WD was *Wind speed at Wheel Hub*, followed by the *Wind speed at 50 m*. Pearson correlation coefficient (PCC) was used in [20] to find the most important feature, and it was found that *Total Solar Irradiance* contributed more to the PV power. Both results align well with our SHAP-based findings for the PVDs and WDs.

4.2. Discovered Patterns

Table 4 and Table 5 present a range of patterns, including high-utility itemsets, high-utility sequential patterns, and correlated high-utility itemsets identified from the PVD1 and WD1 datasets using the EFIM, USPAN, and FCHM algorithms. These patterns capture frequent feature combinations with high predictive utility. They offer insights into prevalent environmental conditions associated with wind/PV power generation.

Table 4: High-Utility patterns found in the PVD1

EFIM	
Itemsets	**U**
Day/Night: Day, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	665567
Day/Night: Night, Power: Low, Humidity: Low, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	1132340
Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	1369564
Temperature: Warm, Day/Night: Night, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	502008
Humidity: Low, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	1046628
Power: Low, Humidity: Low, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	900936
USPAN	
Itemsets	**U**
Total Solar Irradiance: Low, Global Horizontal Irradiance: Low, Temperature: Freezing, Pressure: Low, Humidity: Comfortable	338618
Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low, Humidity: Low	630649
Day/Night: Night, Total Solar Irradiance: Low, Global Horizontal Irradiance: Low, Temperature: Freezing, Pressure: Low, Humidity: Comfortable	331797
Season: Winter, Day/Night: Night, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Temperature: Freezing, Pressure: Low, Humidity: Comfortable	327265
Season: Autumn, Day/Night: Night, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low, Humidity: Low	378784
Total Solar Irradiance: Low, Global Horizontal Irradiance: Low, Temperature: Warm, Pressure: Low, Humidity: Low	411882
FCHM	
Itemsets	**U, B**
Season: Winter, Temperature: Freezing, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	212535, 0.20
Day/Night: Night, Power: Low, Humidity: Low, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	1132340, 0.29
Total Solar Irradiance: High, Day/Night: Day,	347434, 0.26
Temperature: Freezing, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	469227, 0.20
Humidity: Comfortable, Total Solar Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	465102, 0.20
Power: Low, Humidity: Low, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	900936, 0.35

U: Utility, B: Bond.

The EFIM results highlight that the most frequent and high-utility patterns are characterized by low solar irradiance levels across all irradiance types (total, direct normal, and global horizontal), combined with low atmospheric pressure and the daytime temporal feature. Notably, patterns such as *Day/Night: Day, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low* have a utility value of 665,567. Similar high-utility combinations also include *Power: Low* and *Humidity: Low*, indicating a strong relationship between irradiance conditions and power output during nighttime.

SPM via USPAN reveals ordered sequences of feature values, such as *Total Solar Irradiance: Low* → *Global Horizontal Irradiance: Low* → *Temperature: Freezing* → *Pressure: Low* → *Humidity: Comfortable* with a utility of 338,618. This reflects cold, low-pressure periods that still maintain moderate humidity, often observed during winter mornings. These patterns indicate not just co-occurrence, but order, offering predictive cues about how environmental states progress, leading to specific power outputs. Seasonal distinctions emerge as well—for example, *Season:*

Table 5: High-Utility patterns found in the WD1

EFIM	
Itemsets	**U**
Power: High, Wind Speed (Wheel Hub): Moderate, Wind Speed (50 m): Moderate, Wind Speed (30 m): Moderate, Wind Speed (10 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (30 m): S-W, Wind Direction (10 m): S-W, Pressure: Low	401720
Wind Speed (10 m): Strong, Wind Speed (30 m): Strong, Wind Speed (50 m): Strong, Wind Speed (Wheel Hub): Strong, Power: Very High, Pressure: Low	411810
Wind Speed (30 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (30 m): S-W, Wind Direction (10 m): S-W, Pressure: Low	562455
Wind Speed (50 m): Moderate, Wind Speed (30 m): Moderate, Wind Speed (10 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (30 m): S-W, Wind Direction (10 m): S-W, Pressure: Low	534656
Wind Speed (Wheel Hub): Moderate, Wind Speed (50 m): Moderate, Wind Speed (30 m): Moderate, Wind Speed (10 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (30 m): S-W, Wind Direction (10 m): S-W, Pressure: Low	606066
Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Humidity: Comfort, Wind Direction (30 m): S-W, Wind Direction (10 m): S-W, Pressure: Low	350220
USPAN	
Itemsets	**U**
Season: Autumn, Wind Speed (10 m): Moderate, Wind Speed (30 m): Moderate, Wind Speed (50 m): Moderate, Wind Speed (Wheel Hub): Moderate, Pressure: Low	250682
Day/Night: Day, Wind Direction (10 m): S-W, Wind Speed (30 m): Moderate, Wind Direction (30 m): S-W, Wind Speed (50 m): Moderate, Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): Moderate, Wind Direction (Wheel Hub): S-W, Pressure: Low	448983
Wind Speed (10 m): Moderate, Wind Direction (10 m): S-W, Wind Speed (30 m): Moderate, Wind Direction (30 m): S-W, Wind Speed (50 m): Moderate, Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): Moderate, Wind Direction (Wheel Hub): S-W, Pressure: Low	606066
Wind Direction (10 m): S-W, Wind Speed (30 m): Moderate, Wind Direction (30 m): S-W, Wind Speed (50 m): Moderate, Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): Moderate, Wind Direction (Wheel Hub): S-W, Pressure: Low, Humidity: Comfortable	405729
Wind Speed (30 m): Moderate, Wind Direction (30 m): S-W, Wind Speed (50 m): Moderate, Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): Moderate, Wind Direction (Wheel Hub): S-W, Pressure: Low, Humidity: Comfortable	402160
Wind Direction (30 m): S-W, Wind Speed (50 m): Moderate, Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): Moderate, Wind Direction (Wheel Hub): S-W, Pressure: Low	496383
FCHM	
Itemsets	**U, B**
Power: High, Wind Speed (Wheel Hub): Moderate, Wind Speed (50 m): Moderate, Wind Speed (30 m): Moderate, Wind Speed (10 m): Moderate	592658, 0.28
Wind Speed (10 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (10 m): S-W, Pressure: Low	365800, 0.20
Wind Speed (30 m): Moderate, Wind Speed (10 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (30 m): S-W, Wind Direction (10 m): S-W	451000, 0.21
Wind Speed (50 m): Moderate, Wind Speed (30 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (30 m): S-W, Wind direction (10 m): S-W	507346, 0.21
Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind direction (30 m): S-W, Day/Night: Night, Wind Direction (10 m): S-W, Pressure: Low	368368, 0.20
Wind Speed (Wheel Hub): Moderate, Wind Speed (50 m): Moderate, Wind Speed (30 m): Moderate, Wind Direction (50 m): S-W, Wind Direction (Wheel Hub): S-W, Wind Direction (30 m): S-W	615462, 0.21

U: Utility, B: Bond.

Winter → Day/Night: Night → Total Solar Irradiance: Low → Direct Normal Irradiance: Low → Global Horizontal Irradiance: Low → Temperature: Freezing → Pressure: Low → Humidity: Comfortable—indicating that such combinations are not only predictive but temporally ordered. FCHM further identifies patterns that are both high in utility and strongly correlated, as indicated by their bond values. The pattern *Season: Winter, Temperature: Freezing, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low* exhibits a utility of 212,535 with a bond of 0.20, highlighting its strong co-occurrence in the dataset. The highest utility and bond are found in the pattern *Power: Low, Humidity: Low, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low* with a utility 900,936 and a bond of 0.35, further reinforcing the key role of low irradiance and low pressure in determining PV performance under specific environmental conditions.

Results for WD1 highlight the role of wind dynamics across vertical profiles. Synchronized wind speed and direction at different heights influence wind power output. For WD1, EFIM findings reveal that moderate wind speeds—observed at 10 m, 30 m, 50 m, and hub height—combined with southern-to-western (S-W) wind directions form patterns with high utility. For example, the itemset *Wind speed (at all heights): Moderate, Wind direction (at all heights): S-W, Pressure: Low* achieves a utility of 606,066, underlining how aligned wind flows across heights leads to a stable power generation, particularly under low-pressure systems. This also reflects the aerodynamic efficiency achieved when wind flows steadily across the turbine-relevant vertical profile.

USPAN-generated sequential patterns add a temporal dimension. Notably, the appearance of *Season: Autumn* and *Day/Night: Day* in high-utility sequences, alongside S-W wind directions and moderate speeds, indicates that seasonal and diurnal factors modulate the effectiveness of wind flow. These temporal markers enhance the understanding of when consistent wind regimes are most productive.

FCHM reveals high-utility itemsets with strong internal correlation, such as *Wind speed (30 m, 50 m, Wheel Hub): Moderate, Wind direction (30 m, 50 m, Wheel Hub): S-W*, which scores a utility of 615,462 and a bond of 0.21. These correlated configurations reflect real-world aerodynamic consistency, where wind alignment across altitudes leads to high and stable power outputs. For PV systems, the interplay of irradiance, pressure, and temperature directly determines generation potential. Notably, factors like humidity and temperature show non-linear effects, modulating power output even under similar irradiance levels. For wind energy, the vertical synchronization of wind speeds and consistent directions indicates optimal turbine performance scenarios. This provides actionable insights for real-time turbine control and predictive maintenance.

4.3. Discovered Rules

Table 6 summarizes the high-utility rules mined from the PVD1 dataset using both the HGB (association rules) and HUSRM (sequential rules) algorithms, revealing conditional dependencies among PV-related features. For example, an HGB rule reveals that when the conditions *Day/Night: Day* and *Total Solar Irradiance: Low* are satisfied, it is highly likely that *Pressure: Low* will also occur, with perfect utility confidence of 1. Other notable patterns indicate that combinations of low irradiance and low humidity are strong predictors of nighttime conditions, low pressure, and low power outputs—consistent with physical expectations for solar performance. Sequential rules extracted via HUSRM further reinforce these relationships, capturing ordered transitions in environmental states. For example, the sequence starting with *Season: Winter, Day/Night: Day, Low Irradiance, Freezing Temperature* leads to *Low Pressure* with a confidence of 1 and a utility of 215,763, illustrating how seasonal and temporal dependencies influence power generation dynamics. The high-utility rules highlights the dominant influence of irradiance levels, atmospheric pressure, and temporal factors (such as Day/Night and Season) on solar power generation. Moreover, the sequential rules capture dynamic transitions in environmental states. For example, freezing temperatures and low irradiance during winter daytime often precede low atmospheric pressure, which is associated with reduced PV power output.

These temporal insights could be used for real-time forecasting or scheduling maintenance during expected low-generation periods.

Similarly, high-utility rules from the WD1 dataset (Table 7) demonstrate strong relationships among wind speed, direction, and power output. Association rules generated by HGB show that moderate or breezy wind speeds at various altitudes are frequently associated with low atmospheric pressure and consistent wind direction (predominantly S-W). For example, the rule *Pressure: Low, Wind Speed (30 m): Breeze* → *Wind Speed (10 m): Breeze* yields a high utility and a utility confidence of 0.844. Sequential rules from HUSRM reveal time-ordered transitions among features such as wind speeds and directions at various heights, leading to outcomes such as *Pressure: Low* or changes in wind speed at the wheel hub. Notably, the strongest sequential rule—with utility 730,485—shows that breezy conditions at 30 meters often precede similar wind conditions at 50 meters and the hub level, underscoring the consistent vertical wind structure in productive scenarios. These results reflect the temporal and spatial coherence of wind patterns, which is crucial for wind power forecasting. It should be noted that the high-utility results reported in Table 7 for the WD1 dataset were obtained using a subset of 50,000 samples. This subset was selected to ensure efficient processing within the available computational constraints while preserving the representative characteristics of the full dataset.

Table 6: High-Utility rules found in the PVD1

HGB				
Antecedents	**Consequents**	**U**	**AU**	**UC**
Day/Night: Day, Total Solar Irradiance: Low	Pressure: Low	807024	666672	1
Total Solar Irradiance: Low, Humidity: Low	Day/Night: Night, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	1189925	797850	0.73
Total Solar Irradiance: Low, Humidity: Low	Day/Night: Night, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low, Power: Low	1132340	797850	0.69
Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Pressure: Low	Day/Night: Night, Total Solar Irradiance: Low, Power: Low	1556217	555345	0.53
Global Horizontal Irradiance: Low, Pressure: Low, Humidity: Low	Day/Night: Night, Total Solar Irradiance: Low, Direct Normal Irradiance: Low	1189925	591892	0.51
Pressure: Low, Humidity: Low	Day/Night: Night, Global Horizontal Irradiance: Low	720159	520884	0.50
HUSRM				
Antecedents	**Consequents**	**SUP**	**C**	**U**
Season: Winter, Day/Night: Day, Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Temperature: Freezing	Pressure: Low	4071	1	215763
Day/Night: Day, Total Solar Irradiance: High	Pressure: Low, Humidity: Low, Power: Very High	4980	0.54	249000
Total Solar Irradiance: Low, Direct Normal Irradiance: Low, Temperature: Comfortable	Pressure: Low, Humidity: Low, Power: Low	5422	0.67	211458
Direct Normal Irradiance: Low, Global Horizontal Irradiance: Low, Temperature: Warm	Pressure: Low, Humidity: Low	13071	0.79	287562
Global Horizontal Irradiance: Low, Temperature: Mild, Pressure: Low	Humidity: Low	10755	0.77	204345
Pressure: Low	Humidity: Low	43407	0.62	520884

U: Utility, AU: Antecedent Utility, UC: Utility Confidence, SUP: Support, C: Confidence.

To gain more insights, the high-utility sequential patterns and rules are visualized in Figure 4. Blue nodes in Figure 4(a) denote discretized PV feature values, with arrows indicating the flow

Table 7: High-Utility rules found in the WD1

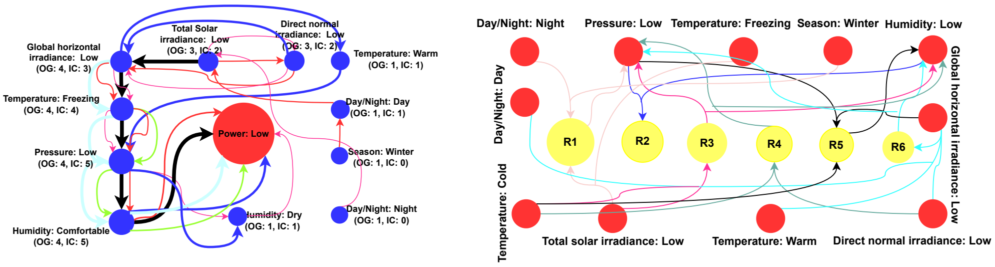
HGB				
Antecedents	**Consequents**	**U**	**AU**	**UC**
Pressure: Low, Wind Speed (30 m): Breeze	Wind Speed (10 m): Breeze	648258	630112	0.84
Wind Speed (10 m): Breeze, Wind Speed (30 m): Breeze	Pressure: Low, Wind Speed (Wheel Hub): Breeze	794774	515282	0.82
Wind Speed (10 m): Moderate, Wind Speed (30 m): Moderate, Wind Speed (50 m): Moderate,	Wind Direction (Wheel Hub): S-W, Pressure: Low, Wind Direction (30 m): S-W, Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): Moderate	602126	692472	0.45
Wind Speed (30 m): Breeze, Wind Speed (Wheel Hub): Breeze	Pressure: Low, Wind Speed (50 m): Breeze	512992	694665	0.46
Wind Speed (50 m): Moderate, Wind Speed (Wheel Hub): Moderate	Pressure: Low, Wind Direction (10 m): S-W, Wind Speed (30 m): Moderate, Wind Direction (30 m): S-W, Wind Direction (50 m): S-W	670140	585466	0.48
Wind Speed (50 m): Breeze, Wind Speed (Wheel Hub): Breeze	Pressure: Low, Wind Speed (30 m): Breeze	542367	564231	0.47
HUSRM				
Antecedents	**Consequents**	**SUP**	**C**	**U**
Wind Speed (50 m): Breeze	Wind Speed (Wheel Hub): Breeze	13390	0.91	415090
Wind Speed (10 m): Breeze, Wind Speed (30 m): Breeze	Wind Speed (50 m): Breeze, Wind Speed (Wheel Hub): Breeze	9814	0.81	608468
Wind Speed (10 m): Breeze	Pressure: Low, Wind Speed (30 m): Breeze, Wind Speed (50 m): Breeze	10739	0.61	547689
Wind Speed (30 m): Breeze	Pressure: Low, Wind Speed (50 m): Breeze, Wind Speed (Wheel Hub): Breeze	11595	0.80	730485
Wind Direction (Wheel Hub): S-W, Wind Speed (10 m): Moderate, Wind Direction (10 m): S-W, Wind Speed (30 m): Moderate, Wind Direction (30 m): S-W, Wind Speed (50 m): Moderate, Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): S-W	Pressure: Low	5081	1	421723
Wind Direction (10 m): S-W, Wind Speed (30 m): Moderate, Wind Direction (30 m): S-W, Wind Speed (50 m): S-W	Wind Direction (50 m): S-W, Wind Speed (Wheel Hub): Moderate	6325	0.82	411125

U: Utility, AU: Antecedent Utility, UC: Utility Confidence, SUP: Support, C: Confidence.

of sequential patterns. Node sizes reflect the number of connections (IC: Incoming connections, OG: Outgoing connections), while arrow thicknesses indicate support values. Each distinct color represents a separate high-utility sequential pattern. Red nodes in Figure 4(b) represent feature-value pairs involved in high-utility sequential rules, and yellow nodes indicate the *Low* power class. Rule nodes (R1–R6) connect antecedent features (via outgoing arrows to the rule node) to consequent features (via arrows from the rule node). The size of yellow nodes is proportional to their support values. One frequent sequential pattern, *Total Solar Irradiance: Low → Global horizontal irradiance: Low → Temperature: Freezing → Pressure: Low → Humidity: Comfortable*, is illustrated in (Figure 4(a)) by a sequence of four black arrows connecting blue nodes representing the respective PV features and their values. The direction and thickness of the arrows indicate the order and strength (support) of the pattern. For high-utility sequential rules, consider rule five, represented by **R5** (Figure 4(b)). The rule involves four red nodes—*Pressure: Low*, *Temperature: Cold*, *Global Horizontal irradiance* and *Humidity: Low*—and a yellow node representing the target class *Power: Low*. In this rule, the antecedents are *Pressure: Low*, *Temperature: Cold* and *Global Horizontal irradiance* and the consequents are *Humidity: Low* and

Power: Low. The antecedent nodes have outgoing arrows toward the **R5** node. Similarly there is an outgoing arrow from **R5** towards the consequent node *Humidity: Low*. In this example, the most important PV features are *Humidity: Comfortable* and *Pressure: Low*, followed by *Temperature: Freezing*, *Global Horizontal Irradiance: Low*, *Global Solar Irradiance: Low*, and *Direct Normal Irradiance*, respectively.

It is important to note that all the high-utility pattern mining algorithms exhibited fast performance, with each completing execution within a few minutes per dataset. This demonstrates the practical viability of the HUF4WP framework for medium-sized renewable energy datasets. While runtime was not a limiting factor in our experiments, we acknowledge that scalability to very large datasets may require further optimization, such as through parallel or distributed implementations. Moreover, the discovery of high-utility patterns and rules is influenced by the choice of algorithm parameters. To ensure that the results were both meaningful and computationally feasible, the parameters for each algorithm were empirically tuned based on the characteristics of each dataset. Consequently, applying HUF4WP to other datasets or domains may require re-tuning to accommodate differences in data distribution or operational conditions.


Global horizontal irradiance: Low (OG: 4, IC: 3)
Total Solar irradiance: Low (OG: 3, IC: 2)
Direct normal irradiance: Low (OG: 3, IC: 2)
Temperature: Warm (OG: 1, IC: 1)
Temperature: Freezing (OG: 4, IC: 4)
Day/Night: Day (OG: 1, IC: 1)
Power: Low
Pressure: Low (OG: 4, IC: 5)
Season: Winter (OG: 1, IC: 0)
Humidity: Dry (OG: 1, IC: 1)
Day/Night: Night (OG: 1, IC: 0)
Humidity: Comfortable (OG: 4, IC: 5)
Day/Night: Night
Pressure: Low
Temperature: Freezing
Season: Winter
Humidity: Low
Day/Night: Day
R1
R2
R3
R4
R5
R6
Global horizontal irradiance: Low
Temperature: Cold
Total solar irradiance: Low
Temperature: Warm
Direct normal irradiance: Low

(a) High-utility sequential patterns of PV features

(b) High-utility sequential rules of PV features

Figure 4: Visual representation of high-utility sequential patterns and rules extracted from PVD1 (Low).

4.4. Classification Results

Before classification, the extracted high-utility patterns from the PVDs and WDs were filtered to retain only those with a minimum length of three elements. To ensure consistency in model training and establish a performance baseline, all classifiers were applied using their default settings.

The classification results for the seven models are summarized in Table 8, where each entry are displayed using the notation $\frac{Low(Medium)}{High(VeryHigh)}MC$. For instance, the GNB result for PVD1, $\frac{0.86(0.778)}{0.702(0.748)}0.41$, indicates that the model achieved 86%, 77.8%, 70.2%, 74.8%, and 41% accuracy for predicting the *Low, Medium, High, Very High*, and MC categories, respectively. The best results for each row are shown in bold in the table. To ensure a thorough evaluation, three different validation methods were tested: (1) 5-fold cross-validation, (2) 10-fold cross-validation, and (3) 80:20 train-test split. Because the 80:20 split consistently yielded superior performance, the reported results are therefore based on this approach. Among the various combinations of classifiers and pattern mining methods (EFIM, USPAN, and FCHM), DT and RF models demonstrated the highest and most consistent accuracies across datasets. Patterns extracted by EFIM generally produced the best classification performance, particularly when paired with tree-based models.

FCHM patterns outperformed those from USPAN in most cases. RF achieved the highest average accuracy with EFIM patterns, while DT slightly outperformed RF when using USPAN and FCHM patterns. GNB consistently showed the lowest performance among all classifiers. When averaging results across all datasets and pattern types, the classifiers ranked as follows: DT (81.7%), RF (81.4%), SVM (72.2%), kNN (71.9%), LR (68.8%), MLP (67%), and GNB (51.9%).

The detailed DT results using EFIM patterns can be found in Table 9. Notably, DT achieved the highest classification accuracy on datasets PVD2 and WD2 using EFIM-derived patterns, indicating strong model alignment with high-utility features in these cases. In contrast, performance on PVD5 and WD5 showed significant variability, highlighting the sensitivity of certain datasets to the quality and structure of the mined patterns. The strong performance of DT and RF can be attributed to their ability to effectively model non-linear relationships and complex feature interactions, which are common in high-utility pattern-based datasets. RF, being an ensemble of multiple DTs, further enhances generalization and reduces overfitting by averaging across many trees, each trained on random data subsets. The ability of tree-based models to generalize well and mitigate overfitting, by aggregating predictions from multiple trees, makes them especially effective for datasets containing complex feature combinations with diverse utility values. On the other hand, simpler models like GNB assume feature independence, which does not align well with the interdependent patterns mined in this study.

While sequential high-utility patterns provided better interpretability, the unordered patterns led to slightly improved classification performance, indicating that for wind/PV power forecasting, the presence of specific feature combinations is more predictive than the sequence in which they occur. The key finding from the classification results is that (1) frequent high- utility-based patterns extracted from wind/PV datasets can be effectively utilized for identification or detection tasks and (2) unordered high-utility features led to superior classification performance compared to sequential high-utility patterns, indicating that in energy-related datasets, the specific order of features holds less significance.

4.5. Comparison

Previous studies–particularly recent DL-based [12, 13, 51, 52, 53, 55, 56, 58, 60], meta-heuristic-based [17, 18, 19, 63, 65, 66] and hybrid-approaches [8, 9, 10, 14, 15, 16, 20]–primarily approached this problem through regression-based time series forecasting, aiming to predict continuous power values. These models are evaluated using metrics such as MAE, MSE, and RMSE. In contrast, our framework redefines the task as a classification problem by discretizing wind/PV features into categorical levels. This problem reformulation shifts the focus from predicting precise numerical values to classifying power output into predefined ranges, with performance assessed using classification metrics such as ACC, P, R F1, etc. To provide a fair comparison and establish a baseline, we trained classification models using all discretized features from the wind/PV datasets. This baseline performance is then compared to models trained on high-utility frequent patterns, enabling an evaluation of the effectiveness of pattern-based feature selection.

Several wind/PV datasets exhibit clear class imbalance (Table 3), with some classes, particularly *Very High* in PVD3 and PVD8, being underrepresented. Class imbalance can potentially skew classification performance by causing models to favor majority classes while underperforming on minority ones. To mitigate this, we applied SMOTE (Synthetic Minority Over-sampling Technique) [79], which generates synthetic examples for minority classes to balance the dataset. The accuracy of classification models on PVDs and WDs, with and without SMOTE, is presented

Table 8: Classification results on high-utility patterns across wind/PV datasets. OA = Overall Average

EFIM							
Dataset	**GNB**	**DT**	**RF**	**MLP**	**SVM**	**kNN**	**LR**
PVD1	$\frac{0.86(0.778)}{0.702(0.748)}$ 0.41	$\frac{0.975(\mathbf{0.94})}{0.912(0.955)}$ **0.882**	$\frac{\mathbf{0.98}(0.938)}{\mathbf{0.932}(\mathbf{0.962})}$ 0.86	$\frac{0.83(0.732)}{0.688(0.772)}$ 0.505	$\frac{0.902(0.772)}{0.745(0.775)}$ 0.595	$\frac{0.888(0.822)}{0.732(0.842)}$ 0.615	$\frac{0.872(0.75)}{0.75(0.75)}$ 0.375
PVD2	$\frac{0.832(0.75)}{0.788(0.692)}$ 0.562	$\frac{\mathbf{1}(\mathbf{1})}{\mathbf{0.952}(0.94)}$ **0.94**	$\frac{\mathbf{1}(0.995)}{0.948(\mathbf{0.945})}$ 0.938	$\frac{0.87(0.75)}{0.77(0.832)}$ 0.548	$\frac{0.84(0.765)}{0.8(0.815)}$ 0.61	$\frac{0.865(0.748)}{0.782(0.838)}$ 0.572	$\frac{0.75(0.75)}{0.74(0.745)}$ 0.532
PVD3	$\frac{0.902(0.612)}{0.548(0.648)}$ 0.502	$\frac{\mathbf{1}(\mathbf{0.788})}{\mathbf{0.76}(\mathbf{0.925})}$ **0.67**	$\frac{0.998(0.785)}{0.755(0.898)}$ 0.642	$\frac{0.92(0.772)}{0.745(0.625)}$ 0.45	$\frac{0.942(0.75)}{0.758(0.75)}$ 0.515	$\frac{0.98(0.732)}{0.738(0.755)}$ 0.495	$\frac{0.912(0.76)}{0.748(0.742)}$ 0.507
PVD4	$\frac{0.808(0.745)}{0.602(0.81)}$ 0.27	$\frac{0.952(0.932)}{0.87(0.922)}$ 0.838	$\frac{\mathbf{0.962}(\mathbf{0.945})}{\mathbf{0.895}(\mathbf{0.932})}$ **0.878**	$\frac{0.805(0.808)}{0.805(0.815)}$ 0.52	$\frac{0.805(0.818)}{0.838(0.86)}$ 0.59	$\frac{0.83(0.822)}{0.852(0.868)}$ 0.638	$\frac{0.718(0.75)}{0.745(0.792)}$ 0.378
PVD5	$\frac{0.662(0.792)}{0.702(0.69)}$ 0.36	$\frac{0.538(0.888)}{0.91(0.58)}$ 0.428	$\frac{0.538(\mathbf{0.892})}{\mathbf{0.918}(0.57)}$ **0.432**	$\frac{0.71(0.79)}{0.842(0.645)}$ **0.432**	$\frac{\mathbf{0.742}(0.818)}{0.805(0.73)}$ 0.425	$\frac{0.618(0.828)}{0.875(0.65)}$ 0.4	$\frac{0.74(0.77)}{0.74(\mathbf{0.75})}$ 0.352
PVD6	$\frac{0.758(0.74)}{0.678(0.87)}$ 0.53	$\frac{\mathbf{0.988}(\mathbf{0.942})}{\mathbf{0.928}(\mathbf{0.972})}$ **0.928**	$\frac{0.985(0.922)}{0.908(0.968)}$ 0.92	$\frac{0.802(0.64)}{0.67(0.838)}$ 0.485	$\frac{0.785(0.755)}{0.75(0.865)}$ 0.545	$\frac{0.828(0.702)}{0.74(0.858)}$ 0.552	$\frac{0.75(0.75)}{0.758(0.825)}$ 0.488
PVD7	$\frac{0.778(0.772)}{0.748(0.735)}$ 0.44	$\frac{0.97(\mathbf{1})}{\mathbf{0.988}(\mathbf{0.98})}$ **0.988**	$\frac{\mathbf{0.972}(0.975)}{0.985(0.978)}$ 0.942	$\frac{0.835(0.688)}{0.735(0.762)}$ 0.46	$\frac{0.838(0.782)}{0.758(0.77)}$ 0.54	$\frac{0.872(0.768)}{0.735(0.785)}$ 0.572	$\frac{0.73(0.75)}{0.75(0.722)}$ 0.342
PVD8	$\frac{0.765(0.408)}{0.71(0.715)}$ 0.328	$\frac{0.905(0.905)}{\mathbf{0.892}(\mathbf{0.882})}$ 0.808	$\frac{\mathbf{0.94}(\mathbf{0.908})}{0.88(0.862)}$ **0.812**	$\frac{0.832(0.775)}{0.752(0.66)}$ 0.498	$\frac{0.82(0.755)}{0.738(0.755)}$ 0.51	$\frac{0.895(0.768)}{0.748(0.698)}$ 0.562	$\frac{0.768(0.742)}{0.742(0.75)}$ 0.398
WD1	$\frac{0.375(0.3)}{0.748(0.732)}$ 0.325	$\frac{0.945(0.97)}{0.982(0.99)}$ 0.942	$\frac{\mathbf{0.978}(\mathbf{0.988})}{\mathbf{0.99}(\mathbf{0.998})}$ **0.97**	$\frac{0.815(0.895)}{0.928(0.858)}$ 0.715	$\frac{0.848(0.848)}{0.93(0.882)}$ 0.74	$\frac{0.888(0.922)}{0.96(0.932)}$ 0.798	$\frac{0.78(0.76)}{0.805(0.752)}$ 0.507
WD2	$\frac{0.355(0.758)}{0.33(0.415)}$ 0.298	$\frac{\mathbf{1}(0.99)}{0.962(0.962)}$ 0.94	$\frac{\mathbf{1}(\mathbf{0.995})}{\mathbf{0.965}(\mathbf{0.968})}$ **0.958**	$\frac{0.895(0.9)}{0.875(0.872)}$ 0.74	$\frac{0.88(0.865)}{0.888(0.84)}$ 0.728	$\frac{0.935(0.925)}{0.91(0.882)}$ 0.818	$\frac{0.808(0.808)}{0.828(0.75)}$ 0.592
WD3	$\frac{0.35(0.358)}{0.25(0.778)}$ 0.395	$\frac{0.9(\mathbf{0.918})}{0.988(\mathbf{0.995})}$ 0.89	$\frac{\mathbf{0.915}(\mathbf{0.918})}{\mathbf{1}(\mathbf{0.995})}$ **0.902**	$\frac{0.852(0.73)}{0.808(0.875)}$ 0.615	$\frac{0.848(0.755)}{0.795(0.84)}$ 0.585	$\frac{0.86(0.805)}{0.818(0.882)}$ 0.672	$\frac{0.772(0.742)}{0.758(0.768)}$ 0.468
WD4	$\frac{0.765(0.96)}{0.31(0.405)}$ 0.415	$\frac{0.985(\mathbf{0.995})}{\mathbf{1}(0.998)}$ 0.992	$\frac{\mathbf{0.995}(\mathbf{0.995})}{\mathbf{1}(\mathbf{1})}$ **0.995**	$\frac{0.825(0.95)}{0.818(0.89)}$ 0.728	$\frac{0.875(0.975)}{0.84(0.888)}$ 0.778	$\frac{0.872(0.97)}{0.865(0.89)}$ 0.785	$\frac{0.835(0.978)}{0.795(0.792)}$ 0.695
WD5	$\frac{0.782(0.628)}{0.628(0.838)}$ 0.502	$\frac{\mathbf{1}(0.575)}{0.575(\mathbf{0.948})}$ 0.495	$\frac{\mathbf{1}(0.565)}{0.565(0.945)}$ 0.498	$\frac{0.918(0.682)}{0.682(0.865)}$ 0.538	$\frac{0.91(\mathbf{0.742})}{\mathbf{0.742}(0.87)}$ **0.542**	$\frac{0.955(0.612)}{0.612(0.878)}$ 0.495	$\frac{0.79(0.74)}{0.74(0.858)}$ 0.475
WD6	$\frac{0.4(0.308)}{0.258(0.775)}$ 0.385	$\frac{0.94(0.935)}{0.992(\mathbf{1})}$ 0.915	$\frac{\mathbf{0.948}(\mathbf{0.94})}{\mathbf{0.998}(\mathbf{1})}$ **0.94**	$\frac{0.88(0.85)}{0.905(0.895)}$ 0.742	$\frac{0.915(0.885)}{0.908(0.915)}$ 0.812	$\frac{0.908(0.9)}{0.955(0.938)}$ 0.848	$\frac{0.87(0.83)}{0.785(0.775)}$ 0.62
OA	$\frac{0.671(0.636)}{0.571(0.704)}$ 0.409	$\frac{0.936(\mathbf{0.913})}{0.908(\mathbf{0.932})}$ 0.833	$\frac{\mathbf{0.944}(0.911)}{\mathbf{0.91}(0.93)}$ **0.835**	$\frac{0.842(0.783)}{0.787(0.8)}$ 0.57	$\frac{0.854(0.806)}{0.807(0.825)}$ 0.608	$\frac{0.871(0.809)}{0.809(0.835)}$ 0.63	$\frac{0.793(0.777)}{0.763(0.769)}$ 0.481

USPAN							
Dataset	**GNB**	**DT**	**RF**	**MLP**	**SVM**	**kNN**	**LR**
PVD1	$\frac{0.748(0.728)}{0.72(0.282)}$ 0.28	$\frac{0.772(0.742)}{\mathbf{0.78}(\mathbf{0.835})}$ 0.47	$\frac{\mathbf{0.78}(0.73)}{0.762(0.818)}$ **0.492**	$\frac{0.625(0.722)}{0.682(0.695)}$ 0.278	$\frac{0.75(\mathbf{0.75})}{0.75(0.75)}$ 0.315	$\frac{0.74(0.69)}{0.705(0.832)}$ 0.388	$\frac{0.75(\mathbf{0.75})}{0.75(0.725)}$ 0.3
PVD2	$\frac{0.722(0.66)}{0.712(0.658)}$ 0.312	$\frac{\mathbf{0.77}(0.755)}{0.672(\mathbf{0.892})}$ **0.462**	$\frac{0.748(0.75)}{0.652(0.888)}$ 0.452	$\frac{0.672(0.732)}{0.7(0.622)}$ 0.285	$\frac{0.75(0.752)}{\mathbf{0.75}(0.75)}$ 0.34	$\frac{0.722(\mathbf{0.765})}{0.708(0.738)}$ 0.298	$\frac{0.75(0.742)}{\mathbf{0.75}(0.75)}$ 0.378
PVD3	$\frac{0.725(0.715)}{0.722(0.638)}$ 0.298	$\frac{\mathbf{0.785}(0.638)}{0.658(\mathbf{0.782})}$ 0.262	$\frac{0.752(0.612)}{0.612(0.765)}$ **0.305**	$\frac{0.38(0.638)}{0.535(0.75)}$ 0.265	$\frac{0.75(\mathbf{0.75})}{\mathbf{0.75}(0.75)}$ 0.285	$\frac{0.73(0.672)}{0.705(0.742)}$ 0.212	$\frac{0.75(\mathbf{0.75})}{\mathbf{0.75}(0.74)}$ 0.27
PVD4	$\frac{0.728(0.738)}{0.688(0.342)}$ 0.285	$\frac{0.748(0.732)}{\mathbf{0.812}(\mathbf{0.858})}$ 0.46	$\frac{0.742(0.7)}{0.795(\mathbf{0.858})}$ 0.455	$\frac{0.742(0.648)}{0.698(0.722)}$ 0.402	$\frac{\mathbf{0.75}(\mathbf{0.75})}{0.75(0.735)}$ **0.475**	$\frac{0.72(0.725)}{0.762(\mathbf{0.858})}$ 0.402	$\frac{\mathbf{0.75}(\mathbf{0.75})}{0.752(0.725)}$ 0.395
PVD5	$\frac{0.248(\mathbf{0.755})}{0.772(0.38)}$ 0.325	$\frac{0.688(0.688)}{\mathbf{0.805}(\mathbf{0.952})}$ 0.45	$\frac{0.66(0.66)}{0.78(0.948)}$ 0.458	$\frac{0.68(0.678)}{0.782(0.765)}$ 0.312	$\frac{\mathbf{0.755}(0.75)}{0.752(0.805)}$ **0.475**	$\frac{0.73(0.688)}{0.772(0.942)}$ 0.435	$\frac{0.75(0.75)}{0.752(0.712)}$ 0.438
PVD6	$\frac{0.67(0.67)}{0.65(0.73)}$ 0.282	$\frac{\mathbf{0.87}(0.7)}{\mathbf{0.758}(\mathbf{0.85})}$ **0.562**	$\frac{\mathbf{0.87}(0.7)}{0.748(0.83)}$ 0.555	$\frac{0.745(0.695)}{0.62(0.648)}$ 0.238	$\frac{0.75(\mathbf{0.75})}{0.75(0.75)}$ 0.292	$\frac{0.778(0.698)}{0.7(0.762)}$ 0.385	$\frac{0.75(\mathbf{0.75})}{0.75(0.75)}$ 0.315
PVD7	$\frac{0.718(0.572)}{0.73(0.738)}$ 0.258	$\frac{0.75(0.735)}{\mathbf{0.805}(\mathbf{0.832})}$ 0.502	$\frac{\mathbf{0.758}(0.72)}{\mathbf{0.805}(0.815)}$ **0.505**	$\frac{0.608(0.622)}{0.668(0.745)}$ 0.245	$\frac{0.75(\mathbf{0.75})}{0.75(0.75)}$ 0.278	$\frac{0.738(0.715)}{0.708(0.745)}$ 0.398	$\frac{0.75(\mathbf{0.75})}{0.75(0.75)}$ 0.305
PVD8	$\frac{0.66(0.66)}{0.56(0.712)}$ **0.345**	$\frac{0.588(0.588)}{0.695(0.792)}$ 0.295	$\frac{0.565(0.565)}{0.68(\mathbf{0.812})}$ 0.295	$\frac{0.38(0.38)}{0.66(0.702)}$ 0.24	$\frac{\mathbf{0.75}(\mathbf{0.75})}{\mathbf{0.75}(0.75)}$ 0.32	$\frac{0.635(0.635)}{0.618(0.768)}$ 0.295	$\frac{0.74(0.74)}{\mathbf{0.75}(0.745)}$ 0.328
WD1	$\frac{0.318(0.335)}{0.308(0.748)}$ 0.318	$\frac{\mathbf{0.925}(0.828)}{0.915(\mathbf{0.995})}$ 0.83	$\frac{0.922(\mathbf{0.862})}{\mathbf{0.942}(\mathbf{0.995})}$ **0.865**	$\frac{0.835(0.745)}{0.735(0.705)}$ 0.338	$\frac{0.915(0.75)}{0.762(0.75)}$ 0.482	$\frac{0.908(0.762)}{0.75(0.745)}$ 0.488	$\frac{0.735(0.742)}{0.755(0.748)}$ 0.39
WD2	$\frac{0.268(0.692)}{0.742(0.72)}$ 0.28	$\frac{\mathbf{0.995}(0.79)}{0.582(0.775)}$ 0.555	$\frac{\mathbf{0.995}(\mathbf{0.8})}{0.582(\mathbf{0.778})}$ **0.568**	$\frac{0.67(0.718)}{0.738(0.72)}$ 0.288	$\frac{0.75(0.758)}{\mathbf{0.75}(0.75)}$ 0.302	$\frac{0.798(0.75)}{0.69(0.698)}$ 0.348	$\frac{0.75(0.742)}{\mathbf{0.75}(0.75)}$ 0.245
WD3	$\frac{0.268(0.252)}{0.252(0.748)}$ 0.258	$\frac{\mathbf{0.86}(0.602)}{0.602(0.998)}$ **0.422**	$\frac{0.842(0.575)}{0.575(\mathbf{1})}$ 0.42	$\frac{0.74(0.748)}{0.748(0.748)}$ 0.262	$\frac{0.75(\mathbf{0.75})}{\mathbf{0.75}(0.748)}$ 0.282	$\frac{0.74(0.695)}{0.695(0.77)}$ 0.255	$\frac{0.75(\mathbf{0.75})}{\mathbf{0.75}(0.75)}$ 0.292
WD4	$\frac{0.318(0.272)}{0.25(0.742)}$ 0.285	$\frac{\mathbf{0.975}(0.91)}{0.862(0.94)}$ 0.795	$\frac{\mathbf{0.975}(\mathbf{0.928})}{\mathbf{0.875}(\mathbf{0.942})}$ **0.84**	$\frac{0.7(0.722)}{0.75(0.742)}$ 0.288	$\frac{0.75(0.75)}{0.755(0.75)}$ 0.335	$\frac{0.725(0.652)}{0.728(0.74)}$ 0.358	$\frac{0.75(0.75)}{0.75(0.752)}$ 0.278
WD5	$\frac{0.738(0.305)}{0.305(0.728)}$ 0.25	$\frac{\mathbf{0.89}(0.825)}{0.612(0.66)}$ **0.442**	$\frac{0.882(\mathbf{0.835})}{0.612(0.64)}$ 0.435	$\frac{0.718(0.56)}{0.702(0.662)}$ 0.282	$\frac{0.74(0.75)}{\mathbf{0.75}(\mathbf{0.75})}$ 0.332	$\frac{0.725(0.74)}{0.655(0.702)}$ 0.222	$\frac{0.74(0.75)}{\mathbf{0.75}(\mathbf{0.75})}$ 0.252
WD6	$\frac{0.305(0.25)}{0.258(0.74)}$ 0.288	$\frac{\mathbf{0.998}(0.888)}{0.812(\mathbf{0.892})}$ **0.828**	$\frac{0.995(\mathbf{0.898})}{\mathbf{0.82}(0.89)}$ 0.825	$\frac{0.95(0.665)}{0.67(0.725)}$ 0.468	$\frac{0.89(0.75)}{0.75(0.74)}$ 0.452	$\frac{0.962(0.762)}{0.678(0.7)}$ 0.475	$\frac{0.885(0.748)}{0.75(0.745)}$ 0.405
OA	$\frac{0.531(0.543)}{0.548(0.636)}$ 0.29	$\frac{\mathbf{0.829}(0.744)}{0.741(\mathbf{0.861})}$ 0.524	$\frac{0.821(0.738)}{0.732(0.856)}$ **0.534**	$\frac{0.675(0.662)}{0.692(0.711)}$ 0.299	$\frac{0.771(\mathbf{0.751})}{\mathbf{0.751}(0.752)}$ 0.355	$\frac{0.761(0.711)}{0.705(0.767)}$ 0.354	$\frac{0.757(0.747)}{0.751(0.742)}$ 0.328

FCHM							
Dataset	**GNB**	**DT**	**RF**	**MLP**	**SVM**	**kNN**	**LR**
PVD1	$\frac{0.745(0.25)}{0.255(0.272)}$ 0.295	$\frac{\mathbf{0.915}(\mathbf{0.85})}{\mathbf{0.832}(\mathbf{0.915})}$ **0.695**	$\frac{0.902(0.832)}{0.82(0.912)}$ 0.69	$\frac{0.698(0.698)}{0.725(0.752)}$ 0.385	$\frac{0.755(0.75)}{0.755(0.785)}$ 0.46	$\frac{0.76(0.738)}{0.755(0.81)}$ 0.388	$\frac{0.775(0.75)}{0.75(0.742)}$ 0.398
PVD2	$\frac{0.755(0.742)}{0.262(0.338)}$ 0.302	$\frac{\mathbf{0.845}(\mathbf{0.775})}{\mathbf{0.765}(\mathbf{0.922})}$ **0.6**	$\frac{0.822(0.75)}{0.735(0.915)}$ 0.598	$\frac{0.675(0.715)}{0.512(0.775)}$ 0.408	$\frac{0.77(0.755)}{0.75(0.81)}$ 0.448	$\frac{0.788(0.708)}{0.682(0.885)}$ 0.37	$\frac{0.748(0.75)}{0.75(0.75)}$ 0.36
PVD3	$\frac{0.765(0.525)}{0.562(0.605)}$ 0.385	$\frac{\mathbf{0.92}(0.73)}{\mathbf{0.805}(\mathbf{0.838})}$ 0.612	$\frac{0.91(0.725)}{0.788(0.83)}$ **0.622**	$\frac{0.762(0.625)}{0.748(0.688)}$ 0.402	$\frac{0.805(\mathbf{0.75})}{0.755(0.8)}$ 0.5	$\frac{0.85(0.688)}{0.735(0.762)}$ 0.468	$\frac{0.758(\mathbf{0.75})}{0.75(0.752)}$ 0.405
PVD4	$\frac{0.752(0.748)}{0.27(0.31)}$ 0.318	$\frac{\mathbf{0.945}(\mathbf{0.775})}{\mathbf{0.778}(\mathbf{0.87})}$ 0.58	$\frac{0.93(0.75)}{0.768(0.845)}$ **0.59**	$\frac{0.74(0.702)}{0.622(0.802)}$ 0.41	$\frac{0.792(0.76)}{0.75(0.812)}$ 0.43	$\frac{0.8(0.745)}{0.72(0.795)}$ 0.472	$\frac{0.755(0.75)}{0.745(0.738)}$ 0.435
PVD5	$\frac{0.732(0.75)}{0.25(0.328)}$ 0.278	$\frac{\mathbf{0.83}(\mathbf{0.812})}{\mathbf{0.808}(0.962)}$ 0.602	$\frac{0.818(0.8)}{0.798(\mathbf{0.965})}$ **0.628**	$\frac{0.748(0.768)}{0.695(0.845)}$ 0.482	$\frac{0.765(0.755)}{0.78(0.888)}$ 0.54	$\frac{0.728(0.78)}{0.768(0.922)}$ 0.522	$\frac{0.75(0.75)}{0.75(0.77)}$ 0.39
PVD6	$\frac{0.26(0.265)}{0.252(0.762)}$ 0.272	$\frac{\mathbf{0.928}(\mathbf{0.825})}{\mathbf{0.828}(\mathbf{0.9})}$ **0.705**	$\frac{0.915(0.815)}{0.798(0.885)}$ 0.67	$\frac{0.705(0.732)}{0.598(0.698)}$ 0.295	$\frac{0.755(0.75)}{0.75(0.792)}$ 0.372	$\frac{0.798(0.742)}{0.685(0.782)}$ 0.428	$\frac{0.75(0.758)}{0.75(0.772)}$ 0.325
PVD7	$\frac{0.25(0.272)}{0.258(0.755)}$ 0.335	$\frac{0.882(\mathbf{0.815})}{\mathbf{0.842}(\mathbf{0.902})}$ **0.735**	$\frac{\mathbf{0.895}(0.788)}{0.812(0.892)}$ 0.685	$\frac{0.805(0.695)}{0.622(0.74)}$ 0.395	$\frac{0.792(0.75)}{0.75(0.805)}$ 0.432	$\frac{0.83(0.722)}{0.698(0.792)}$ 0.475	$\frac{0.752(0.748)}{0.75(0.79)}$ 0.392
PVD8	$\frac{0.74(0.28)}{0.272(0.752)}$ 0.385	$\frac{\mathbf{0.922}(0.792)}{\mathbf{0.842}(\mathbf{0.868})}$ **0.668**	$\frac{0.91(\mathbf{0.798})}{0.818(0.828)}$ 0.642	$\frac{0.838(0.752)}{0.718(0.752)}$ 0.4	$\frac{0.798(0.75)}{0.745(0.765)}$ 0.478	$\frac{0.858(0.722)}{0.712(0.758)}$ 0.442	$\frac{0.77(0.75)}{0.742(0.752)}$ 0.385
WD1	$\frac{0.265(0.755)}{0.255(0.298)}$ 0.258	$\frac{0.898(\mathbf{0.925})}{\mathbf{0.872}(\mathbf{0.875})}$ 0.792	$\frac{\mathbf{0.915}(0.922)}{0.868(\mathbf{0.875})}$ **0.795**	$\frac{0.75(0.682)}{0.658(0.738)}$ 0.42	$\frac{0.75(0.808)}{0.772(0.762)}$ 0.502	$\frac{0.765(0.858)}{0.755(0.772)}$ 0.507	$\frac{0.75(0.76)}{0.75(0.735)}$ 0.428
WD2	$\frac{0.305(0.258)}{0.728(0.75)}$ 0.268	$\frac{0.96(\mathbf{0.91})}{\mathbf{0.76}(0.778)}$ **0.712**	$\frac{\mathbf{0.965}(0.902)}{0.752(\mathbf{0.8})}$ 0.7	$\frac{0.718(0.755)}{0.68(0.685)}$ 0.432	$\frac{0.762(0.858)}{0.738(0.768)}$ 0.548	$\frac{0.838(0.855)}{0.702(0.738)}$ 0.555	$\frac{0.762(0.862)}{0.748(0.778)}$ 0.52
WD3	$\frac{0.335(0.758)}{0.268(0.255)}$ 0.288	$\frac{\mathbf{0.858}(0.788)}{0.822(0.86)}$ 0.652	$\frac{0.845(\mathbf{0.805})}{\mathbf{0.825}(\mathbf{0.875})}$ **0.685**	$\frac{0.732(0.712)}{0.66(0.61)}$ 0.338	$\frac{0.758(0.76)}{0.752(0.762)}$ 0.425	$\frac{0.775(0.752)}{0.73(0.798)}$ 0.445	$\frac{0.755(0.758)}{0.75(0.748)}$ 0.35
WD4	$\frac{0.282(0.725)}{0.745(0.585)}$ 0.26	$\frac{0.922(0.86)}{0.855(0.932)}$ 0.768	$\frac{\mathbf{0.928}(\mathbf{0.868})}{\mathbf{0.872}(\mathbf{0.945})}$ **0.82**	$\frac{0.775(0.802)}{0.615(0.745)}$ 0.472	$\frac{0.792(0.808)}{0.782(0.782)}$ 0.53	$\frac{0.782(0.82)}{0.792(0.802)}$ 0.565	$\frac{0.75(0.758)}{0.76(0.758)}$ 0.375
WD5	$\frac{0.738(0.712)}{0.43(0.43)}$ 0.26	$\frac{\mathbf{0.845}(\mathbf{0.838})}{0.618(0.618)}$ 0.28	$\frac{0.812(0.82)}{0.595(0.595)}$ 0.275	$\frac{0.748(0.75)}{0.505(0.505)}$ 0.352	$\frac{0.772(0.798)}{\mathbf{0.75}(\mathbf{0.75})}$ **0.39**	$\frac{0.8(0.772)}{0.678(0.678)}$ 0.222	$\frac{0.75(0.785)}{\mathbf{0.75}(\mathbf{0.75})}$ 0.298
WD6	$\frac{0.28(0.265)}{0.75(0.738)}$ 0.322	$\frac{\mathbf{0.872}(0.815)}{\mathbf{0.875}(0.892)}$ 0.645	$\frac{0.86(\mathbf{0.822})}{0.872(\mathbf{0.9})}$ **0.67**	$\frac{0.745(0.762)}{0.775(0.735)}$ 0.395	$\frac{0.76(0.785)}{0.76(0.775)}$ 0.472	$\frac{0.795(0.77)}{0.825(0.748)}$ 0.555	$\frac{0.752(0.745)}{0.75(0.752)}$ 0.38
OA	$\frac{0.515(0.522)}{0.397(0.513)}$ 0.302	$\frac{\mathbf{0.896}(\mathbf{0.822})}{\mathbf{0.807}(\mathbf{0.867})}$ 0.646	$\frac{0.888(0.814)}{0.794(0.862)}$ **0.648**	$\frac{0.746(0.725)}{0.652(0.719)}$ 0.399	$\frac{0.773(0.774)}{0.756(0.79)}$ 0.466	$\frac{0.798(0.762)}{0.731(0.789)}$ 0.458	$\frac{0.756(0.762)}{0.75(0.756)}$ 0.389

in Table 10. Contrary to expectations, the results indicate that SMOTE did not consistently improve accuracy across the datasets. In fact, in several cases, accuracy slightly declined after applying SMOTE. In datasets such as PVD3 and PVD8, where class imbalance was most severe,

Table 9: DT results for all metrics on EFIM-derived high-utility patterns

Dataset	ACC	P	R	F1	MCC	AUC	AUPRC
PVD1	$\frac{\mathbf{0.975}(0.94)}{0.912(0.955)}0.882$	$\frac{\mathbf{0.941}(0.896)}{0.822(0.918)}0.885$	$\frac{\mathbf{0.96}(0.86)}{0.83(0.9)}0.882$	$\frac{\mathbf{0.95}(0.878)}{0.826(0.909)}0.883$	$\frac{\mathbf{0.934}(0.838)}{0.767(0.879)}0.844$	$\frac{\mathbf{0.97}(0.917)}{0.903(0.945)}0.931$	$\frac{\mathbf{0.914}(0.811)}{0.747(0.864)}0.91$
PVD2	$\frac{\mathbf{1}(\mathbf{1})}{0.952(0.94)}0.94$	$\frac{\mathbf{1}(\mathbf{1})}{0.893(0.852)}0.94$	$\frac{\mathbf{1}(\mathbf{1})}{0.92(0.92)}0.94$	$\frac{\mathbf{1}(\mathbf{1})}{0.906(0.885)}0.94$	$\frac{\mathbf{1}(\mathbf{1})}{0.875(0.845)}0.92$	$\frac{\mathbf{1}(\mathbf{1})}{0.946(0.938)}0.961$	$\frac{\mathbf{1}(\mathbf{1})}{0.848(0.81)}0.949$
PVD3	$\frac{\mathbf{1}(0.788)}{0.76(0.925)}0.67$	$\frac{\mathbf{1}(0.59)}{0.523(0.898)}0.689$	$\frac{\mathbf{1}(0.49)}{0.45(0.79)}0.67$	$\frac{\mathbf{1}(0.536)}{0.484(0.84)}0.675$	$\frac{\mathbf{1}(0.402)}{0.33(0.794)}0.563$	$\frac{\mathbf{1}(0.715)}{0.657(0.88)}0.79$	$\frac{\mathbf{1}(0.447)}{0.381(0.767)}0.721$
PVD4	$\frac{\mathbf{0.952}(0.932)}{0.87(0.922)}0.838$	$\frac{\mathbf{0.909}(0.892)}{0.75(0.835)}0.842$	$\frac{\mathbf{0.9}(0.83)}{0.72(0.86)}0.838$	$\frac{\mathbf{0.905}(0.86)}{0.735(0.847)}0.838$	$\frac{\mathbf{0.873}(0.817)}{0.649(0.796)}0.784$	$\frac{\mathbf{0.949}(0.921)}{0.845(0.921)}0.9$	$\frac{0.862(0.813)}{0.639(0.777)}\mathbf{0.87}$
PVD5	$\frac{0.538(0.888)}{\mathbf{0.91}(0.58)}0.428$	$\frac{0.062(0.802)}{\mathbf{0.881}(0.041)}0.43$	$\frac{0.06(0.73)}{\mathbf{0.74}(0.03)}0.428$	$\frac{0.061(0.764)}{\mathbf{0.804}(0.034)}0.429$	$\frac{-0.246(0.692)}{\mathbf{0.751}(-0.23)}0.237$	$\frac{0.392(0.836)}{\mathbf{0.869}(0.388)}0.619$	$\frac{0.224(0.662)}{\mathbf{0.74}(0.231)}0.493$
PVD6	$\frac{\mathbf{0.988}(0.942)}{0.928(0.972)}0.928$	$\frac{\mathbf{0.99}(0.881)}{0.845(0.959)}0.931$	$\frac{\mathbf{0.96}(0.89)}{0.87(0.93)}0.928$	$\frac{\mathbf{0.975}(0.886)}{0.857(0.944)}0.928$	$\frac{\mathbf{0.967}(0.847)}{0.809(0.926)}0.904$	$\frac{\mathbf{0.978}(0.925)}{0.908(0.958)}0.957$	$\frac{\mathbf{0.96}(0.812)}{0.767(0.909)}0.944$
PVD7	$\frac{0.97(\mathbf{1})}{0.988(0.98)}0.988$	$\frac{0.958(\mathbf{1})}{0.97(0.951)}0.988$	$\frac{0.92(\mathbf{1})}{0.98(0.97)}0.988$	$\frac{0.939(\mathbf{1})}{0.975(0.96)}0.988$	$\frac{0.919(\mathbf{1})}{0.967(0.947)}0.983$	$\frac{0.953(\mathbf{1})}{0.985(0.981)}0.992$	$\frac{0.902(\mathbf{1})}{0.956(0.937)}0.989$
PVD8	$\frac{\mathbf{0.905}(\mathbf{0.905})}{0.892(0.882)}0.808$	$\frac{0.804(0.798)}{0.794(0.762)}\mathbf{0.807}$	$\frac{0.82(\mathbf{0.83})}{0.77(0.77)}0.808$	$\frac{0.812(\mathbf{0.814})}{0.782(0.766)}0.807$	$\frac{0.748(\mathbf{0.75})}{0.711(0.688)}0.744$	$\frac{0.885(\mathbf{0.892})}{0.867(0.847)}0.882$	$\frac{0.715(0.72)}{0.689(0.649)}\mathbf{0.839}$
WD1	$\frac{0.945(0.97)}{0.982(\mathbf{0.99})}0.942$	$\frac{0.882(0.915)}{0.96(\mathbf{1})}0.943$	$\frac{0.9(\mathbf{0.97})}{\mathbf{0.97}(0.96)}0.942$	$\frac{0.891(0.942)}{0.965(\mathbf{0.98})}0.942$	$\frac{0.854(0.922)}{0.954(\mathbf{0.973})}0.924$	$\frac{0.93(0.975)}{0.978(\mathbf{0.98})}0.962$	$\frac{0.819(0.902)}{0.939(\mathbf{0.97})}0.95$
WD2	$\frac{\mathbf{1}(0.99)}{0.962(0.962)}0.94$	$\frac{\mathbf{1}(0.98)}{0.929(0.929)}0.94$	$\frac{\mathbf{1}(0.98)}{0.92(0.92)}0.94$	$\frac{\mathbf{1}(0.98)}{0.925(0.925)}0.94$	$\frac{\mathbf{1}(0.973)}{0.9(0.9)}0.92$	$\frac{\mathbf{1}(0.987)}{0.948(0.958)}0.966$	$\frac{\mathbf{1}(0.965)}{0.875(0.888)}0.955$
WD3	$\frac{0.9(0.918)}{0.988(\mathbf{0.995})}0.89$	$\frac{0.783(0.832)}{0.99(\mathbf{1})}0.891$	$\frac{0.83(0.84)}{0.96(\mathbf{0.98})}0.89$	$\frac{0.806(0.836)}{0.975(\mathbf{0.99})}0.89$	$\frac{0.739(0.781)}{0.967(\mathbf{0.987})}0.853$	$\frac{0.884(0.915)}{0.978(\mathbf{0.99})}0.933$	$\frac{0.702(0.768)}{0.96(\mathbf{0.985})}0.91$
WD4	$\frac{0.985(0.995)}{\mathbf{1}(0.998)}0.992$	$\frac{0.952(0.98)}{\mathbf{1}(\mathbf{1})}0.993$	$\frac{0.99(\mathbf{1})}{\mathbf{1}(0.99)}0.992$	$\frac{0.971(0.99)}{\mathbf{1}(0.995)}0.993$	$\frac{0.961(0.987)}{\mathbf{1}(0.993)}0.99$	$\frac{0.987(0.997)}{\mathbf{1}(0.995)}0.995$	$\frac{0.945(0.98)}{\mathbf{1}(0.992)}0.994$
WD5	$\frac{\mathbf{1}(0.575)}{0.575(0.948)}0.495$	$\frac{\mathbf{1}(0.039)}{0.039(0.934)}0.497$	$\frac{\mathbf{1}(0.03)}{0.03(0.85)}0.495$	$\frac{\mathbf{1}(0.034)}{0.034(0.89)}0.496$	$\frac{\mathbf{1}(-0.235)}{-0.235(0.857)}0.327$	$\frac{\mathbf{1}(0.391)}{0.391(0.926)}0.663$	$\frac{\mathbf{1}(0.228)}{0.228(0.848)}0.553$
WD6	$\frac{0.94(0.935)}{0.992(\mathbf{1})}0.915$	$\frac{0.896(0.878)}{\mathbf{1}(\mathbf{1})}0.917$	$\frac{0.86(0.86)}{0.97(\mathbf{1})}0.915$	$\frac{0.878(0.869)}{0.985(\mathbf{1})}0.915$	$\frac{0.838(0.826)}{0.98(\mathbf{1})}0.887$	$\frac{0.912(0.909)}{0.985(\mathbf{1})}0.944$	$\frac{0.805(0.79)}{0.978(\mathbf{1})}0.926$
OA	$\frac{\mathbf{0.936}(0.913)}{0.908(0.932)}0.833$	$\frac{\mathbf{0.87}(0.82)}{0.814(0.863)}0.835$	$\frac{\mathbf{0.871}(0.808)}{0.795(0.848)}0.833$	$\frac{\mathbf{0.87}(0.813)}{0.804(0.855)}0.833$	$\frac{\mathbf{0.828}(0.757)}{0.744(0.811)}0.777$	$\frac{\mathbf{0.917}(0.884)}{0.876(0.908)}0.892$	$\frac{0.846(0.778)}{0.768(0.831)}\mathbf{0.857}$

applying SMOTE did not result in any significant accuracy improvement. Despite generating synthetic samples to balance the minority classes, the overall model accuracy either remained unchanged or even decreased slightly. SMOTE works best when minority class examples lie close to each other in feature space. However, if the dataset has complex structures or high-dimensional interactions (such as those involving high-utility patterns), SMOTE might generate synthetic samples that do not reflect true data distributions, leading to lower accuracy. Moreover, some classifers such as RF and DT are inherently robust to moderate class imbalance. In such cases, oversampling may offer little benefit, while possibly making the model more prone to overfitting.

One interesting observation is that DT and RF achieved superior performance (90.4% and 90.6% accuracy) when trained on high-utility pattern features obtained with EFIM compared to training on all features (DT (87.6%), RF(87.8%)) without SMOTE. However, this trend did not extend uniformly to all classifiers. For other models, their accuracy results are higher when using the complete feature set. This again suggests that tree-based models are particularly well-suited to pattern-based feature sets, likely due to their ability to handle discrete, interpretable features and to effectively exploit the hierarchical structure of patterns. In contrast, non-tree-based models appear to benefit from the broader information available in the full feature set. The diversity and richness of data provide more continuous or high-dimensional cues necessary for these models to perform better. Therefore, the choice between pattern-based and full feature-based modeling should consider both the classifier type and computational constraints. For scenarios prioritizing interpretability and efficiency, pattern-based models–especially with DT and RF–are advantageous. However, for maximizing raw predictive power with models like SVM or LR, full feature sets, potentially augmented with techniques like SMOTE, may still offer better results.

In addition to the limited improvements in accuracy, applying SMOTE introduced a substantial increase in computational time. When trained on the original datasets without SMOTE, the classifiers completed execution on all features in approximately 5 hours. However, with SMOTE applied, the execution time increased significantly, requiring more than 17 hours to process the same datasets. This notable rise in computational cost, without clear performance gains, suggests that SMOTE may not be a practical solution for high-dimensional, pattern-rich renewable energy datasets. Particularly when computational efficiency is a priority. In contrast, the proposed HUF4WP framework—which involves first mining high-utility patterns and then using

Table 10: Comparison of classification accuracy when classifiers are trained on the complete set of original features

Without SMOTE							
Dataset	**GNB**	**DT**	**RF**	**MLP**	**SVM**	**kNN**	**LR**
PVD1	$\frac{0.902(0.222)}{0.695(0.905)}0.803$	$\frac{0.917(0.904)}{0.933(0.95)}\mathbf{0.849}$	$\frac{\mathbf{0.917}(\mathbf{0.904})}{0.932(\mathbf{0.95})}\mathbf{0.849}$	$\frac{0.916(0.903)}{0.931(\mathbf{0.95})}\mathbf{0.849}$	$\frac{0.917(0.902)}{\mathbf{0.933}(0.949)}0.849$	$\frac{0.887(0.893)}{0.923(0.937)}0.826$	$\frac{0.914(0.9)}{0.929(0.948)}0.843$
PVD2	$\frac{0.487(0.478)}{0.363(0.924)}0.395$	$\frac{\mathbf{0.82}(0.84)}{0.783(0.927)}\mathbf{0.668}$	$\frac{\mathbf{0.821}(0.84)}{0.783(0.927)}\mathbf{0.668}$	$\frac{\mathbf{0.821}(0.84)}{0.784(\mathbf{0.928})}0.667$	$\frac{\mathbf{0.821}(\mathbf{0.84})}{\mathbf{0.784}(0.927)}0.667$	$\frac{0.803(0.839)}{0.761(0.921)}0.64$	$\frac{0.818(0.84)}{0.777(0.926)}0.659$
PVD3	$\frac{0.858(0.681)}{0.701(\mathbf{1})}0.733$	$\frac{0.873(\mathbf{0.859})}{0.903(\mathbf{1})}\mathbf{0.808}$	$\frac{\mathbf{0.873}(0.858)}{0.903(\mathbf{1})}0.808$	$\frac{\mathbf{0.873}(0.857)}{0.901(\mathbf{1})}0.808$	$\frac{0.872(0.857)}{\mathbf{0.903}(\mathbf{1})}0.808$	$\frac{0.857(0.841)}{0.892(0.999)}0.793$	$\frac{0.87(0.856)}{0.895(\mathbf{1})}0.8$
PVD4	$\frac{0.85(0.307)}{0.769(0.967)}0.798$	$\frac{0.899(0.901)}{0.945(0.969)}0.842$	$\frac{0.899(0.901)}{0.944(0.969)}0.842$	$\frac{0.899(0.902)}{0.944(\mathbf{0.969})}0.84$	$\frac{0.898(\mathbf{0.903})}{\mathbf{0.946}(\mathbf{0.969})}\mathbf{0.842}$	$\frac{0.884(0.894)}{0.935(0.963)}0.82$	$\frac{\mathbf{0.901}(0.901)}{\mathbf{0.946}(\mathbf{0.97})}0.842$
PVD5	$\frac{0.837(0.316)}{0.775(0.994)}0.838$	$\frac{0.895(0.896)}{0.943(\mathbf{0.994})}0.848$	$\frac{\mathbf{0.896}(0.896)}{0.943(\mathbf{0.994})}0.847$	$\frac{0.895(\mathbf{0.896})}{0.943(\mathbf{0.994})}\mathbf{0.849}$	$\frac{0.895(\mathbf{0.896})}{0.944(\mathbf{0.994})}0.848$	$\frac{0.872(0.884)}{0.942(0.992)}0.843$	$\frac{0.894(\mathbf{0.896})}{\mathbf{0.944}(\mathbf{0.994})}0.846$
PVD6	$\frac{0.922(0.319)}{0.932(0.916)}0.796$	$\frac{0.928(0.898)}{\mathbf{0.949}(\mathbf{0.984})}0.879$	$\frac{0.929(0.898)}{\mathbf{0.949}(\mathbf{0.984})}0.879$	$\frac{0.928(\mathbf{0.899})}{\mathbf{0.949}(\mathbf{0.984})}\mathbf{0.88}$	$\frac{\mathbf{0.929}(0.898)}{0.949(\mathbf{0.984})}0.879$	$\frac{0.927(0.888)}{0.945(0.978)}0.868$	$\frac{0.927(0.869)}{0.944(\mathbf{0.984})}0.872$
PVD7	$\frac{0.898(0.27)}{0.9(0.88)}0.774$	$\frac{\mathbf{0.921}(0.902)}{0.957(\mathbf{0.993})}0.884$	$\frac{0.921(0.902)}{0.957(\mathbf{0.993})}\mathbf{0.885}$	$\frac{0.92(0.9)}{\mathbf{0.957}(0.992)}0.884$	$\frac{0.921(\mathbf{0.903})}{0.957(\mathbf{0.993})}0.884$	$\frac{0.907(0.893)}{0.954(0.992)}0.872$	$\frac{0.914(0.873)}{0.957(\mathbf{0.993})}0.874$
PVD8	$\frac{0.866(0.682)}{0.619(0.885)}0.668$	$\frac{0.876(0.84)}{0.933(\mathbf{0.999})}0.822$	$\frac{0.876(0.839)}{0.932(\mathbf{0.998})}0.822$	$\frac{0.877(0.84)}{0.933(\mathbf{0.999})}0.822$	$\frac{0.876(\mathbf{0.841})}{\mathbf{0.933}(0.998)}\mathbf{0.825}$	$\frac{0.859(0.815)}{0.927(0.998)}0.803$	$\frac{\mathbf{0.878}(0.831)}{0.931(0.998)}0.821$
WD1	$\frac{0.846(0.671)}{0.708(0.764)}0.683$	$\frac{0.895(0.818)}{0.845(0.923)}0.74$	$\frac{0.9(0.821)}{0.845(0.924)}0.744$	$\frac{0.9(0.818)}{\mathbf{0.847}(\mathbf{0.926})}0.747$	$\frac{\mathbf{0.902}(\mathbf{0.824})}{0.846(0.925)}\mathbf{0.75}$	$\frac{0.882(0.801)}{0.821(0.921)}0.718$	$\frac{0.897(0.786)}{0.837(0.924)}0.74$
WD2	$\frac{0.878(0.727)}{0.592(0.651)}0.561$	$\frac{0.926(0.814)}{\mathbf{0.778}(0.826)}0.663$	$\frac{0.932(0.817)}{0.777(\mathbf{0.828})}0.669$	$\frac{\mathbf{0.934}(0.818)}{0.777(0.824)}0.67$	$\frac{0.933(\mathbf{0.818})}{0.776(0.826)}\mathbf{0.673}$	$\frac{0.929(0.791)}{0.754(0.811)}0.635$	$\frac{0.93(0.793)}{0.758(0.8)}0.642$
WD3	$\frac{0.76(0.473)}{0.476(0.875)}0.585$	$\frac{0.873(0.789)}{0.813(0.901)}0.682$	$\frac{\mathbf{0.882}(\mathbf{0.799})}{0.824(0.912)}0.701$	$\frac{0.882(0.797)}{0.823(0.913)}0.697$	$\frac{0.877(0.792)}{\mathbf{0.825}(\mathbf{0.914})}\mathbf{0.701}$	$\frac{0.868(0.786)}{0.809(0.907)}0.678$	$\frac{0.852(0.764)}{0.806(0.905)}0.667$
WD4	$\frac{0.707(0.59)}{0.593(0.919)}0.717$	$\frac{0.888(0.815)}{0.855(0.943)}0.754$	$\frac{0.893(0.818)}{0.859(0.946)}\mathbf{0.761}$	$\frac{0.892(\mathbf{0.821})}{0.857(0.946)}0.759$	$\frac{\mathbf{0.896}(0.819)}{\mathbf{0.863}(\mathbf{0.947})}0.759$	$\frac{0.879(0.806)}{0.844(0.942)}0.726$	$\frac{0.887(0.809)}{0.85(0.94)}0.744$
WD5	$\frac{0.876(0.376)}{0.789(0.911)}0.618$	$\frac{0.948(0.915)}{0.934(0.956)}0.876$	$\frac{0.952(0.92)}{0.936(0.958)}0.884$	$\frac{0.953(0.921)}{0.936(\mathbf{0.958})}0.881$	$\frac{\mathbf{0.954}(\mathbf{0.922})}{\mathbf{0.939}(\mathbf{0.958})}\mathbf{0.885}$	$\frac{0.951(0.913)}{0.932(0.955)}0.868$	$\frac{\mathbf{0.954}(0.919)}{0.932(0.956)}0.882$
WD6	$\frac{0.698(0.551)}{0.655(\mathbf{1})}0.669$	$\frac{0.858(0.783)}{\mathbf{0.844}(\mathbf{1})}0.734$	$\frac{\mathbf{0.861}(\mathbf{0.785})}{0.842(\mathbf{1})}\mathbf{0.738}$	$\frac{0.86(0.781)}{0.843(\mathbf{1})}0.735$	$\frac{0.859(0.779)}{0.843(\mathbf{1})}0.734$	$\frac{0.842(0.761)}{0.819(0.998)}0.7$	$\frac{0.84(0.755)}{0.834(\mathbf{1})}0.715$
OA	$\frac{0.813(0.476)}{0.683(0.899)}0.688$	$\frac{0.894(0.855)}{0.887(0.955)}0.789$	$\frac{\mathbf{0.896}(\mathbf{0.857})}{0.888(0.956)}0.793$	$\frac{\mathbf{0.896}(0.857)}{0.888(0.956)}0.792$	$\frac{\mathbf{0.896}(0.857)}{\mathbf{0.889}(\mathbf{0.956})}\mathbf{0.793}$	$\frac{0.882(0.843)}{0.875(0.951)}0.771$	$\frac{0.891(0.842)}{0.881(0.953)}0.782$

With SMOTE							
Dataset	**GNB**	**DT**	**RF**	**MLP**	**SVM**	**kNN**	**LR**
PVD1	$\frac{0.902(0.207)}{0.658(0.905)}0.803$	$\frac{\mathbf{0.916}(0.718)}{0.868(0.915)}0.751$	$\frac{\mathbf{0.916}(0.718)}{0.869(0.915)}0.751$	$\frac{\mathbf{0.916}(0.713)}{0.873(0.913)}0.752$	$\frac{0.916(0.717)}{0.863(0.912)}0.753$	$\frac{0.887(\mathbf{0.892})}{\mathbf{0.922}(\mathbf{0.937})}\mathbf{0.827}$	$\frac{0.907(0.748)}{0.878(0.905)}0.756$
PVD2	$\frac{0.487(0.478)}{0.363(\mathbf{0.924})}0.395$	$\frac{0.742(0.774)}{0.677(0.92)}\mathbf{0.668}$	$\frac{0.742(0.774)}{0.677(0.92)}0.668$	$\frac{0.742(0.748)}{0.677(0.92)}0.666$	$\frac{0.752(0.748)}{0.677(0.921)}0.668$	$\frac{\mathbf{0.803}(\mathbf{0.839})}{\mathbf{0.76}(0.921)}0.64$	$\frac{0.761(0.766)}{0.687(0.92)}0.659$
PVD3	$\frac{\mathbf{0.858}(0.681)}{0.701(\mathbf{1})}0.732$	$\frac{0.851(0.731)}{0.833(0.999)}0.767$	$\frac{0.851(0.731)}{0.833(0.999)}0.767$	$\frac{0.856(0.741)}{0.832(0.999)}0.768$	$\frac{0.851(0.724)}{0.822(0.999)}0.766$	$\frac{0.857(\mathbf{0.84})}{\mathbf{0.892}(\mathbf{1})}\mathbf{0.793}$	$\frac{0.849(0.706)}{0.799(0.999)}0.757$
PVD4	$\frac{0.85(0.307)}{0.769(\mathbf{0.967})}0.798$	$\frac{0.888(0.794)}{0.784(0.955)}0.79$	$\frac{0.888(0.794)}{0.784(0.955)}0.79$	$\frac{0.888(0.791)}{0.784(0.956)}0.791$	$\frac{0.889(0.789)}{0.776(0.953)}0.79$	$\frac{0.884(\mathbf{0.892})}{\mathbf{0.934}(0.963)}\mathbf{0.819}$	$\frac{\mathbf{0.89}(0.797)}{0.77(0.957)}0.786$
PVD5	$\frac{0.837(0.323)}{0.776(\mathbf{0.994})}0.815$	$\frac{0.895(0.804)}{0.786(0.985)}0.803$	$\frac{\mathbf{0.895}(0.804)}{0.786(0.985)}0.803$	$\frac{0.894(0.803)}{0.785(0.989)}0.802$	$\frac{0.894(0.803)}{0.784(0.985)}0.802$	$\frac{0.872(\mathbf{0.884})}{\mathbf{0.941}(0.992)}\mathbf{0.843}$	$\frac{0.893(0.81)}{0.775(0.99)}0.801$
PVD6	$\frac{0.922(0.319)}{0.931(0.916)}0.796$	$\frac{0.925(0.801)}{0.943(0.92)}0.81$	$\frac{0.925(0.801)}{0.943(0.92)}0.82$	$\frac{0.92(0.801)}{0.943(0.914)}0.811$	$\frac{\mathbf{0.927}(0.797)}{0.943(0.913)}0.81$	$\frac{\mathbf{0.927}(\mathbf{0.888})}{\mathbf{0.945}(\mathbf{0.978})}\mathbf{0.868}$	$\frac{0.915(0.799)}{0.931(0.913)}0.818$
PVD7	$\frac{0.898(0.271)}{0.9(0.895)}0.725$	$\frac{\mathbf{0.916}(0.82)}{0.939(0.961)}0.828$	$\frac{\mathbf{0.916}(0.82)}{0.938(0.961)}0.829$	$\frac{0.911(0.837)}{0.941(0.969)}0.83$	$\frac{0.91(0.818)}{0.939(0.961)}0.822$	$\frac{0.908(\mathbf{0.892})}{\mathbf{0.953}(\mathbf{0.992})}\mathbf{0.873}$	$\frac{0.904(0.782)}{0.934(0.968)}0.79$
PVD8	$\frac{\mathbf{0.866}(0.682)}{0.614(0.9)}0.668$	$\frac{0.843(0.746)}{0.91(0.981)}0.765$	$\frac{0.847(0.746)}{0.91(0.981)}0.765$	$\frac{0.844(0.737)}{0.912(0.972)}0.762$	$\frac{0.849(0.741)}{0.912(0.956)}0.76$	$\frac{0.859(\mathbf{0.815})}{\mathbf{0.927}(\mathbf{0.997})}\mathbf{0.802}$	$\frac{0.843(0.718)}{0.9(0.928)}0.739$
WD1	$\frac{0.846(0.671)}{0.703(0.764)}0.683$	$\frac{0.87(0.783)}{0.829(0.919)}0.74$	$\frac{0.875(0.788)}{0.831(0.921)}0.744$	$\frac{0.863(0.764)}{0.833(\mathbf{0.923})}0.747$	$\frac{0.864(0.779)}{\mathbf{0.835}(0.917)}\mathbf{0.75}$	$\frac{\mathbf{0.879}(\mathbf{0.793})}{0.813(0.917)}0.718$	$\frac{0.852(0.722)}{0.808(0.915)}0.741$
WD2	$\frac{0.881(0.713)}{0.59(0.652)}0.561$	$\frac{0.924(0.785)}{0.699(0.789)}0.663$	$\frac{0.928(0.789)}{0.698(0.788)}0.669$	$\frac{\mathbf{0.929}(\mathbf{0.79})}{0.704(0.793)}0.667$	$\frac{0.926(0.783)}{0.685(0.778)}\mathbf{0.673}$	$\frac{0.927(0.783)}{\mathbf{0.748}(\mathbf{0.809})}0.635$	$\frac{0.917(0.729)}{0.604(0.759)}0.642$
WD3	$\frac{0.761(0.473)}{0.473(0.87)}0.585$	$\frac{0.854(0.78)}{0.805(0.891)}0.682$	$\frac{0.862(\mathbf{0.789})}{0.813(\mathbf{0.901})}\mathbf{0.701}$	$\frac{0.843(0.761)}{\mathbf{0.816}(0.893)}0.695$	$\frac{0.843(0.765)}{0.809(0.897)}0.701$	$\frac{\mathbf{0.863}(0.765)}{0.792(0.899)}0.678$	$\frac{0.798(0.711)}{0.793(0.875)}0.667$
WD4	$\frac{0.709(0.59)}{0.592(0.915)}0.717$	$\frac{0.869(0.801)}{0.829(0.939)}0.754$	$\frac{\mathbf{0.875}(\mathbf{0.805})}{0.831(0.94)}\mathbf{0.761}$	$\frac{0.87(0.797)}{0.831(\mathbf{0.941})}0.76$	$\frac{0.873(0.802)}{0.829(0.939)}0.759$	$\frac{0.874(0.797)}{\mathbf{0.839}(0.94)}0.726$	$\frac{0.865(0.79)}{0.812(0.934)}0.744$
WD5	$\frac{0.886(0.383)}{0.791(0.934)}0.562$	$\frac{0.948(0.912)}{0.885(0.955)}0.869$	$\frac{0.952(\mathbf{0.917})}{0.886(\mathbf{0.956})}0.875$	$\frac{0.951(0.907)}{0.887(\mathbf{0.956})}0.872$	$\frac{\mathbf{0.953}(\mathbf{0.917})}{0.856(0.952)}\mathbf{0.876}$	$\frac{0.948(0.901)}{\mathbf{0.917}(0.951)}0.861$	$\frac{0.952(0.911)}{0.81(0.947)}0.874$
WD6	$\frac{0.684(0.551)}{0.652(\mathbf{1})}0.669$	$\frac{0.837(0.722)}{0.805(\mathbf{1})}0.734$	$\frac{0.839(0.724)}{0.804(\mathbf{1})}\mathbf{0.737}$	$\frac{0.819(0.74)}{0.81(\mathbf{1})}0.734$	$\frac{0.837(0.714)}{0.8(\mathbf{1})}0.734$	$\frac{\mathbf{0.842}(\mathbf{0.757})}{\mathbf{0.816}(0.999)}0.7$	$\frac{0.81(0.669)}{0.788(\mathbf{1})}0.713$
OA	$\frac{0.813(0.475)}{0.68(0.903)}0.679$	$\frac{0.877(0.784)}{0.828(0.938)}0.759$	$\frac{0.879(0.786)}{0.829(0.939)}0.763$	$\frac{0.875(0.781)}{0.831(0.938)}0.761$	$\frac{0.877(0.778)}{0.824(0.935)}0.762$	$\frac{\mathbf{0.881}(\mathbf{0.839})}{\mathbf{0.872}(\mathbf{0.95})}\mathbf{0.77}$	$\frac{0.868(0.761)}{0.806(0.929)}0.749$

these patterns as features for classification—proved to be computationally efficient. The pattern discovery and subsequent use in classification typically require only a few minutes per dataset. This highlights both the interpretability and practical advantages of the pattern-based method. It offers a substantial reduction in processing time while maintaining competitive classification performance. Some previous works also reported the computational time for their proposed methods. For example, IEDN-RNET [14] and QT-MARF [12] took more than one hour and 2 hour on average on WDs, and Santa Vitoria and Natal datasets. The method [66] took approximately 45 s on four wind datasets containing 11,520 samples only.

Figure 5 presents the best average values of the classification metrics for three cases: (1) DT trained on EFIM-derived pattern features, (2) RF trained on all features without SMOTE, and (3) RF trained on all features with SMOTE. DT with EFIM achieves superior average performance across several metrics—including ACC, P, F1, MCC and AUPRC—compared to RF when trained on all features, both with and without SMOTE. Case 1 demonstrates a balanced trade-off between interpretability, efficiency, and classification effectiveness, confirming that pattern-based features can lead to robust and high-performing models, especially for tree-based classifiers. Although RF performs reasonably well in bot full-feature cases, the marginal gains observed with SMOTE—particularly in R for minority classes—do not outweigh the previously

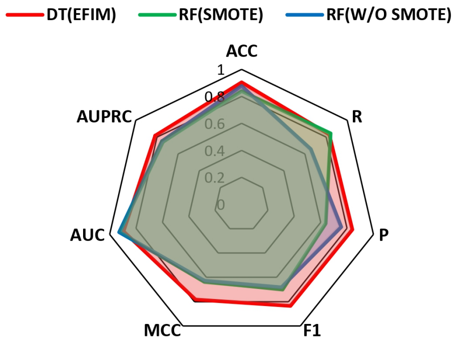


Figure 5: Comparison of the best average classification metrics for three cases: (1) DT with EFIM-derived features (red line), (2) RF with original features without SMOTE (blue line), and (3) RF with original features with SMOTE (green line).

noted computational cost. Notably, DT with EFIM (case 1) achieved approximately a 2.96% improvement in accuracy over RF on all features without SMOTE (case 2) and 7.72% improvement over the RF on all features with SMOTE (case 3).

5. Conclusion

This study introduced HUF4WP, a novel high-utility data fusion framework designed for the analysis and efficient classification/prediction of wind/PV power. Unlike traditional regression-based forecasting models, HUF4WP reformulates the prediction of wind/PV power as a classification problem, enabling more actionable insights through categorical levels. By integrating high-utility pattern mining algorithms and feature weighting, HUF4WP effectively identified not only frequent but operationally significant patterns in renewable energy datasets. Extensive experimental results across multiple wind/PV datasets demonstrated that HUF4WP achieves superior classification performance compared to baseline classifiers trained directly on the raw features, while also offering faster computational performance than both baseline and regression-based forecasting approaches. Moreover, the discovered high-utility patterns and rules provide transparent and interpretable insights into key environmental and temporal drivers of renewable energy generation. These insights expose meaningful dependencies among features such as irradiance, wind speed, atmospheric pressure, and diurnal cycles. This provides valuable guidance for energy management, grid stability, and predictive maintenance. The adaptability of the HUF4WP framework allows it to be extended to other domains where interpretability and high-utility pattern discovery are essential.

While HUF4WP demonstrated strong classification performance and interpretability, certain limitations remain. Firstly, HUF4WP relies on the discretization of continuous features, which—while improving interpretability—may lead to a loss of fine-grained information. Secondly, the discretization thresholds for features are statically defined based on domain knowledge and may require adaptation when applied to different regions or evolving operational conditions. Thirdly, although HUF4WP is computationally efficient for the studied datasets, its scalability to ultra-large energy data streams requires further investigation. Fourthly, the performance of classification models was partially dependent on the accuracy of SHAP-based feature importance estimations. Future work will focus on (1) dynamic thresholds adjustment to accommodate changing environmental conditions; (2) integration with real-time data streams for continuous learning; (3) extending HUF4WP to multi-source heterogeneous energy systems; (4) incorporating domain-specific constraints and exploring multi-objective optimization in high-utility pattern mining to further enhance the operational relevance of discovered insights; (5) using non-linear models such as RF or XGBoost for SHAP value estimation to capture more complex feature interactions and improve the granularity of importance assessments; and (6) discovering negative or contrasting patterns in energy data and integrating them into the classification framework.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Code Availability: The code(s) used in this study is available at: github.com/saqibdola/HUF4WP.

CRediT author statement

M. Saqib Nawaz: Methodology, Conceptualization, Data Curation, Software, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing. **Philippe Fournier-Viger:** Supervision, Resources, Methodology, Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing. **M. Zohaib Nawaz**: Formal Analysis, Validation, Visualization, Methodology, Writing - Review & Editing. **Yulin He:** Investigation, Validation, Writing - Review & Editing. **Unil Yun:** Validation, Writing - Review & Editing.

Conflict of Interest: Authors declare no conflict of interest.

References

[1] F. Kreith, C. F. Kutscher, J. B. Milford, Principles of Sustainable Energy Systems, Third Edition, CRC Press, 2018.

[2] S. Twidale, Renewables provided record 32% of global electricity in 2024, available at: reuters.com/sustainability/climate-energy/renewables-provided-record-32-global-electricity-2024-ember-says-2025-04-07.

[3] E. Çam, M. Casanovas, J. Moloney, Electricity 2025 - Analysis and forecast to 2027, International Energy Agency, available at: https://iea.blob.core.windows.net/assets/0f028d5f-26b1-47ca-ad2a-5ca3103d070a/Electricity2025.pdf, 2025.

[4] M. J. Mayer, G. Gróf, Extensive comparison of physical models for photovoltaic power forecasting, Applied Energy, 283 (2021) 116239.

[5] D. Niu, L. Sun, M. Yu, K. Wang, Point and interval forecasting of ultra-short-term wind power based on a data-driven method and hybrid deep learning model, Energy 254 (2022) 124384.

[6] J. Yan, Y. Liu, S. Han, Y. Wang, S. Feng, Reviews on uncertainty analysis of wind power forecasting. Renewable Sustainable Energy Reviews, 52 (2015) 1322–1330.

[7] D. Markovics, J. Martin, Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. Renewable Sustainable Energy Reviews, 161 (2022) 112364.

[8] G. Terrén-Serrano, M. Martínez-Ramón, Deep learning for intra-hour solar forecasting with fusion of features extracted from infrared sky images, Information Fusion, 95 (2023) 42-61.

[9] S. Almaghrabi, M. Rana, M. Hamilton, M. S. Rahaman, Multivariate solar power time series forecasting using multilevel data fusion and deep neural networks, Information Fusion, 104 (2024) 102180.

[10] B. Li, H. Wang, J. Zhang, Short-term power forecasting of photovoltaic generation based on CFOA-CNN-BiLSTM-Attention. Electrical Engineering, (2025) https://doi.org/10.1007/s00202-025-03031-9.

[11] H. Fan, Z. Zhen, N. Liu, Y. Sun, X. Chang, Y. Li, F. Wang, Z. Mi, Fluctuation pattern recognition based ultra-short-term wind power probabilistic forecasting method, Energy, 266 (2023) 126420,

[12] A. F. Mirza, Z. Shu, M. Usman, M. Mansoor, Q. Ling, Quantile-transformed multi-attention residual framework (QT-MARF) for medium-term PV and wind power prediction, Renewable Energy, 220 (2024) 119604.

[13] G. Wang, X. Huang, Y. Li, H. Wang, X. Zhang, J. Qiu, Conv-ELSTM: An ensemble deep learning approach for predicting short-term wind power. IET Renewable Power Generation, 18 (2024) 4084–4096.

[14] A. F. Mirza, M. Mansoor, M. Usman, Q. Ling, Hybrid Inception-embedded deep neural network ResNet for short and medium-term PV-Wind forecasting, Energy Conversion and Management, 294 (2023) 117574.

[15] Y.-Yi. Hong, C. L. P. P. Rioflorido, A hybrid deep learning-based neural network for 24-h ahead wind power forecasting, Applied Energy, 250 (2019) 530-539.

[16] S. Cui, S. Lyu, Y. Ma, K. Wang, Improved informer PV power short-term prediction model based on weather typing and AHA-VMD-MPE, Energy, 307 (2024) 132766,

[17] A. F. Mirza, M. Mansoor, M. Usman, Q. Ling, A comprehensive approach for PV wind forecasting by using a hyperparameter tuned GCVCNN-MRNN deep learning model, Energy, 283 (2023) 129189.

[18] C. Yu, A comprehensive wind power prediction system based on correct multiscale clustering ensemble, similarity matching, and improved whale optimization algorithm—A case study in China, Renewable Energy, 243 (2025) 122529.

[19] C. Stoean, M. Zivkovic, A. Bozovic, N. Bacanin, R. Strulak-Wójcikiewicz, M. Antonijevic, R. Stoean, Metaheuristic-based hyperparameter tuning for recurrent deep learning: application to the prediction of solar energy generation. Axioms, 12 (2023) 266.

[20] U. A. Khan, N. M. Khan, M. H. Zafar, Resource efficient PV power forecasting: Transductive transfer learning based hybrid deep learning model for smart grid in Industry 5.0, Energy Conversion and Management: X, 20 (2023) 100486.

[21] J. M. Luna. P. Fournier-Viger, S. Ventura, Frequent itemset mining: a 25 years review, WIREs Data Mining and Knowledge Discovery, 9 (2019) e1329.

[22] P. Fournier-Viger et al, A survey of sequential pattern mining, Data Science and Pattern Recognition, 1 (1) (2017) 55-74.

[23] N. Yusof, R. Zurita-Milla, M-J. Kraak, B. Retsios, Mining Frequent spatio-temporal patterns in wind speed and direction, in Connecting a Digital Europe Through Location and Place, Springer, 2014.

[24] N. Yusof, R. Zurita-Milla, Mapping frequent spatio-temporal wind profile patterns using multi-dimensional sequential pattern mining, International Journal of Digital Earth, 10(3) (2016) 238–256.

[25] B. Aydin, D. Kempton, V. Akkineni, R. Angryk, K.G. Pillai, Mining spatiotemporal co-occurrence patterns in solar datasets, Astronomy and Computing, 13 (2015) 136-144.

[26] P. Fournier-Viger, J. C-W. Lin, T. Truong-Chi, P. Nkambou, R. A survey of high utility itemset mining, In: High-Utility Pattern Mining. Studies in Big Data, vol 51. Springer, 2019.

[27] E. Mosca, F. Szigeti, S. Tragianni, D. Gallagher, G. Groh, SHAP-based explanation methods: a review for NLP interpretability," in Proc. of ACL (20220) 4593–4603.

[28] D. Yang E. Wu J. Kleissl, Operational solar forecasting for the real-time market. International Journal of Forecasting, 35 (2019) 1499–519.

[29] E. Lorenz, T. Scheidsteger, J. Hurka, Regional PV power prediction for improved grid integration. Prog Photovoltaics Res Appl., 19 (2011) 757–71.

[30] B. Wolff, J. Kühnert, E. Lorenz, O. Kramer, D. Heinemann, Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. Solar Energy, 135 (2016) 197–208.

[31] Y. N. Saint-Drenan, S. Bofinger, R. Fritz, S. Vogt, G.H. Good, J. Dobschinski, An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation, Solar Energy, 120 (2015) 479–93.

[32] N. Holland, X. Pang, W. Herzberg, S. Karalus, J. Bor, E. Lorenz, Solar and PV forecasting for large PV power plants using numerical weather models, satellite data and ground measurements in Proc. of PVSC, (2019) 1609–1614.

[33] R. A. Silva, M. C. Brito, Spatio-temporal PV forecasting sensitivity to modules' tilt and orientation. Applied Energy, 255 (2019) 113807.

[34] Almeida MP, Munoz ~ M, de la Parra I, Perpin~ ´ an O. Comparative study of PV power forecast using parametric and nonparametric PV models. Sol Energy 2017;155: 854–66

[35] H. Ye, B. Yang, Y. Han, Q. Li, J. Deng, S. Tian, Wind Speed and Power Prediction Approaches: Classifications, Methodologies, and Comments. Front. Energy Res. 10:901767, 2022.

[36] A. Lau, P. McSharry, Approaches for multi-step density forecasts with application to aggregated wind power, arXiv (2010) 1003.0996.

[37] M. -S. Ko, L. Lee, J. -K. Kim, C. W. Hong, Z. Y. Dong, K. Hur, Deep concatenated residual network with bidirectional LSTM for one-hour-ahead wind power forecasting, IEEE Transactions on Sustainable Energy, 12(2) (2020) 1321–1335.

[38] A. Dairi, F. Harrou, Y. Sun, S. Khadraoui, Short-term forecasting of photovoltaic solar power production using variational auto-encoder driven deep learning approach, Applied Sciences, 10 (2020) 8400.

[39] M. Ferreira, A. Santos, P. Lucio, Short-term forecast of wind speed through mathematical models, Energy Reports, 5 (2019) 1172-1184.

[40] C. Wan, J. Wang J, J. Lin, Y. Song, Z. Dong, Nonparametric prediction intervals of wind power via linear programming, IEEE Transactions on Power Systems, 33(1) (2018), 1074–1076.

[41] J. Dowell, P. Pinson, Very short-term probabilistic wind power forecasts by sparse vector autoregression. IEEE Transactions on Smart Grid, 7(2) (2016) 763–70.

[42] W. Xie, P. Zhang, R. Chen, Z. Zhou Z, A nonparametric bayesian framework for short-term wind power probabilistic forecast. IEEE Transactions on Power Systems, 34(1) (2019)371–379.

[43] H. S. Dhiman, D. Deb, Machine intelligent techniques for ramp Event prediction in offshore and onshore wind farms, arxiv, (2020) 14220.

[44] Y. Gala, Á. Fernández, J. Díaz, J. R. Dorronsoro, Hybrid machine learning forecasting of solar radiation values, Neurocomputing, 176 (2016) 48-59.

[45] N. E. Benti, M. D. Chaka, A. G. Semie, Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects. Sustainability, 15 (2023) 7087.

[46] D. C. Arati, P. S. Menon, J. Velayudhan, P. Poornachandran, A. K. Raj, O. K. Sikha, Enhancing Wind Power Prediction through Machine Learning, in Proc. pg AISP, (2024) 1-5.

[47] M. X. Mabel, E. Fernandez, Analysis of wind power generation and prediction using ANN: A case study. Renewable Energy, 33 (2008) 986–992.

[48] J. Zeng, W. Qiao, Short-term wind power prediction using a wavelet support vector machine, IEEE Transactions on Sustainable Energy, 3(2) (2012) 255–264.

[49] A. K. Biswas, S. I. Ahmed, T. Bankefa, P. Ran-Ganathan, H. Salehfar, Performance analysis of short and mid-term wind power prediction using ARIMA and hybrid models, in Proc. of PECI, (2021) 1–7.

[50] E. López, C. Valle, H. Allende, E. Gil, H. Madsen, Wind power forecasting based on echo state networks and long short-term memory. Energies, 11 (2018) 536.

[51] S. Kazmi, A concurrent CNN-RNN approach for multi-step wind power forecasting, Masters Thesis, Toronto Metropolitan University (2022).

[52] J. Yu, X. Li, L. Yang, L. Li, Z. Huang, K. Shen, X. Yang, X.; Yang, Z. Xu, D. Zhang, Deep Learning Models for PV Power Forecasting: Review. Energies, 17 (2024) 3973.

[53] G. Gong, K. Lou, J. Yin, D. Li, Forecast of photovoltaic Power Generation Based on GRU. In Proc. of EITCE'22, (2022) 283–286.

[54] K. Wang, X. Qi, H. Liu, Photovoltaic power forecasting based LSTM-convolutional Network, Energy, 189 (2019) 116225.

[55] O. Rubasinghe, X. Zhang, T. K. Chau, Y. H. Chow, T. Fernando, H. H. -C. Iu, A novel sequence to sequence data modelling based CNN-LSTM algorithm for three years ahead monthly peak load forecasting, IEEE Transactions on Power Systems, 39 (1) (2024) 1932-1947.

[56] S. Huang, C. Yan, Y. Qu, Deep learning model-transformer based wind power forecasting approach. Frontiers in Energy Research, 16 (2023).

[57] S. Tasnim, A. Rahman, A. M. Than Oo, Md. E. Haque, Autoencoder for wind power prediction. Renewables 4 (2017) 6.

[58] F. Harrou, A. Dairi, A. Dorbane, Y. Sun, Enhancing wind power prediction with self-attentive variational autoencoders: A comparative study, Results in Engineering, 23 (2024) 102504.

[59] M. Khodayar, S. Mohammadi, M.E. Khodayar, J. Wang, G. Liu, Convolutional graph autoencoder: a generative deep neural network for probabilistic spatiotemporal solar irradiance forecasting, IEEE Transactions on Sustainable Energy, 11 (2) (2019) 571–583.

[60] J. Simeunović, B. Schubnel, P. -J. Alet, R. E. Carrillo, Spatio-temporal graph neural networks for multi-site PV power forecasting, IEEE Transactions on Sustainable Energy, 13 (2) (2022) 1210-1220.

[61] R. A. Gupta, R. Kumar, A. K. Bansal, Selection of input variables for the prediction of wind Speed in wind farms based on genetic algorithm, Wind Engineering, 35 (2011) 649-660.

[62] A. T. Eseye, J. Zhang, D. Zheng, Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information, Renewable Energy, 118 (2018) 357-367,

[63] M. S. Iqbal, M. M. Kabir, A. S. Surja, A. Rouf, Solar radiation prediction using ant colony optimization and artificial neural network, European Journal of Engineering and Technology Research, 7 (2022).

[64] K.M. El-Naggar, M.R. AlRashidi, M.F. AlHajri, A.K. Al-Othman, Simulated annealing algorithm for photovoltaic parameters identification, Solar Energy, 86 (2012) 266-274.

[65] M. El-Dosuky, R. Alowaydan, B. Alqarni, Wind power forecasting using grey wolf Ootimized long short-term memory based on numerical weather prediction, Journal of Power and Energy Engineering, 12 (2024) 1-16.

[66] P.Lu, L. Ye, M. Pei, Y. Zhao, B. Dai, Z. Li, Short-term wind power forecasting based on meteorological feature extraction and optimization strategy, Renewable Energy, 184 (2022) 642-661,

[67] C. Timplalexis, N. Bezas, A. D. Bintoudi, L. Zyglakis, V. Pavlopoulos, A. C. Tsolakis, S. Krinidis, D. Tzovaras,A hybrid physical/statistical day-ahead direct PV forecasting engine, IET Conference Proceedings, 5 (2020) 258-263.

[68] S. Syama, J. Ramprabhakar, R. Anand, V. P. Meena, J. M. Guerrero, A novel hybrid methodology for wind speed

and solar irradiance forecasting based on improved whale optimized regularized extreme learning machine, Scientific Reports, 14 (2024) 31657.

[69] A. A. Ali , F. A. Kadhim , A. Abdelaziz A, E-K. El-Sayed M, I. Abdelhameed, K. D. Sami, Optimized ensemble model for wind power forecasting using hybrid whale and dipper-throated optimization algorithms, Frontiers in Energy Research, 11 (2023).

[70] M. S. Nawaz, M. Z. Nawaz, P. Fournier-Viger, J. M. Luna, Analysis and classification of employee attrition and absenteeism in industry: A sequential pattern mining-based methodology, Computers in Industry, 159-160, (2024) 104106.

[71] Y. Chen, J. Xu, Solar and wind power data from the Chinese state grid renewable energy generation forecasting competition, Scientific Data, 9 (2022) 57.

[72] S. Zida, P. Fournier-Viger, J. Chun-W. Lin, C-W. Wu, V. S. Tseng, EFIM: A highly efficient algorithm for high-utility itemset mining, in Proc. of MICAI, (2015) 530-546.

[73] J. Sahoo, A. Das, A. Goswami, An efficient approach for mining association rules from high utility itemsets, Expert Systems with Applications, 42 (2015) 5754-5778.

[74] J. Yin, Z. Zheng, L. Cao, USpan: an efficient algorithm for mining high utility sequential patterns, in Proc. of KDD'12, (2012) 660-668.

[75] S. Zida, P. Fournier-Viger, C-W. Wu, J. C-W. Lin, V. S. Tseng, Efficient mining of high utility sequential rules, in Proc. of MLDM, (2015) 157-171.

[76] P. Fournier-Viger, Y. Zhang, J. C-W. Lin, D-T. Dinh, H. B. Le, Mining correlated high-utility itemsets using various correlation measures, Logic Journal of IGPL, 28 (2020) 19032.

[77] P. Fournier-Viger, J. C. -W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, H. T. Lam, The SPMF open-source data mining library version 2, in Proc. of ECML PKDD, (2016) 36-40.

[78] O. Kramer, Scikit-learn, in Machine Learning for Evolution Strategies, Springer, 2016 45–53.

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