AI Driven Wind Energy Forecasting: Case Study for Sustainable Energy in Pakistan

MCS

Author

Muhammad Asif

Registration No

00000327458

Supervisor

Assoc Prof. Dr. Shibli Nisar

A thesis submitted to the faculty of Department of Computer Software Engineering, Military College of Signals, National University of Sciences and Technology, Islamabad, Pakistan, in partial fulfillment of the requirements for the degree of MS in Software Engineering

February 2024

THESIS ACCEPTANCE CERTIFICATE

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr. **Muhammad Asif**, Registration No. **0000327458**, of **Military College of Signals** has been vetted by undersigned, found complete in all respect as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial, fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the student have been also incorporated in the said thesis.

Signature: ______________

Name of Supervisor: Assoc Prof. Dr. Shibli Nisar

Date: 12-2-24

Signature (HoD): ______________ Brig Head of Dept of CSE Mil College of Sigs (NUST)

Date: 20/2/24

Signature (Dean/Principal): ______________

Date: 21/2/24

Brig Dean, MCS (NUST) (Asif Masood, Phd)

FORM TH-4

Form TH-4

NATIONAL UNIVERSITY OF SCIENCES & TECHNOLOGY
MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by **Muhammad Asif** Regn No **00000327458** Titled: **"AI Driven Wind Energy Forecasting : Case Study for Sustainable Energy in Pakistan"** be accepted in partial fulfillment of the requirements for the award of **MS Software Engineering** degree.

Examination Committee Members

1. Name: **Assoc Prof Dr. Ihtesham Ul Islam** Signature: ____________

2. Name: **Asst Prof Dr. Nauman Ali Khan** Signature: ____________

3. Name: **Lec Muhammad Armoghan Khan** (Co-Supervisor) Signature: ____________

Supervisor's Name: **Assoc Prof Dr. Shibli Nisar** Signature: ____________

Date: ____________

Brig
Head of Dept of CSE
Mil College of Sigs (NUST)

Head of Department

20/2/24
Date

COUNTERSIGNED

Brig
Dean, MCS (NUST)
(Asif Masood, Phd)

Dean

Date: 21/2/24

CERTIFICATE OF APPROVAL

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in this thesis, entitled "**AI Driven Wind Energy Forecasting: Case Study for Sustainable Energy in Pakistan**" was conducted by Mr./Ms **Muhammad Asif** under the supervision of **Assoc. Prof. Dr. Shibli Nisar** No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the CSE Department in partial fulfillment of the requirements for the degree of Master of Science in Field of Software Engineering Department of CSE, Military College of Signals, National University of Sciences and Technology, Islamabad.

Student Name: **Muhammad Asif** Signature: ________

Examination Committee:

a) External Examiner 1: Assoc. Prof. Dr. Ihtesham ul Islam (MCS) Signature: ________

b) External Examiner 2: Assoc. Prof. Dr. Nauman Ali Khan (MCS) Signature: ________

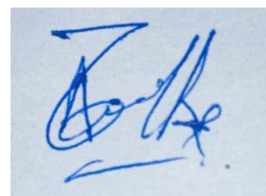
Name of Supervisor: Assoc. Prof. Dr. Shibli Nisar Signature: ________

Name of Dean/HOD: Brig. Adnan Ahmed Khan, PhD Signature: ________

Brig
Head of Dept of CSE
Mil College of Sigs (NUST)

DECLARATION

I Muhammad Asif hereby state that my MS thesis titled " AI Driven Wind Energy Forecasting: Case Study for Sustainable Energy in Pakistan " is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.
At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.



Student Signature

Name: Muhammad Asif

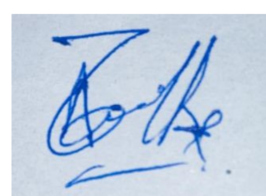
Date: 12th Feb, 2024

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "**AI Driven Wind Energy Forecasting: Case Study for Sustainable Energy in Pakistan**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the University reserves the rights to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.



Student Signature

Name: Muhammad Asif

Date: 12th Feb, 2024

DEDICATION

This thesis is dedicated to

MY FAMILY AND TEACHERS

for their love, endless support and encouragement

ACKNOWLEDGEMENTS

I am grateful to Allah, the Almighty, for His mercy and benevolence who has bestowed me with the strength and the passion to complete this thesis.

I am also thankful to my supervisor Assoc Prof. Dr. Shibli Nisar, Co-supervisor Lect. Muhammad Armoghan Khan EE Department at MCS and committee members who have always guided me with their keen and useful counseling in achieving my research objectives.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

Numerical Weather Prediction	NWP
Auto Regressive Moving Average	ARMA
Extreme Learning Machine	ELM
Kernel Extreme Learning Machine	KELM
Least Squares Support Vector Machine	LSSVM
Artificial Neural Networks	ANNs
Long Short-Term Memory	LSTM
Multi-layer Perceptron	MLP
Deep Belief Network	DBN
Gated Recurrent Unit	GRU
Convolutional Neural Network	CNN
National Electric Power Regulatory Authority	NEPRA
Wind Speed	WS
Wind Power	WP
Mean Variance Estimation	MVE
Extreme Gradient Boosting	XGBoost
Density-Based Spatial Clustering of Applications with Noise	DBSCAN
Recursive Feature Elimination	RFE
Epsilon	EPS
Minimum Points	MIN_PTS

ABSTRACT

Wind power is having a significant effect on the global energy landscape, providing sustainable choices that address environmental, economic, and social concerns. Forecasting wind power is of the utmost importance for the successful incorporation of wind power into the network of electrical power distribution systems. Accurate forecasts empower electric grid operators to foresee variations in wind energy generation. In order to accurately anticipate wind power, meteorological data, weather occurrences, wind turbine performance, and grid limits are all taken into consideration. In this context, we have proposed a novel approach called DBSCAN-RFE-XGBoost a two-stage process. The initial stage, we have proposed a clustering algorithm that uses density-based spatial clustering (DBSCAN) to automate the EPS value that is required for DBSCAN clustering by using K-dist plot and Knee Point Detection Algorithm. Additionally, we have removed outliers from the dataset. In the second stage, we applied two fold scheme, first we enginered temporal features and then applied recursive feature elimination to the preprocessed dataset which identifies best suited features to feed into the XGBoost algorithm to predict wind power. To determine effectiveness of proposed model, we have utilized SCADA dataset obtained from a wind farm in Pakistan. In comparison to existing benchmarking approaches, our suggested model achieves a performance that is 38.89% higher (RMSE), and its effectiveness is demonstrated by an R2 value of roughly 5.10%.

Keywords — Wind Energy Forecasting, DBSCAN, Recursive Feature Elimination, XGBOOST, Deep Learning, Wind energy forecasting in Pakistan

Chapter 1

INTRODUCTION

1.1 Motivation

As the world faces climate change issues and continuously depleting fossil fuels, there is a growing need for sustainable and renewable energy sources. In an effort to address these difficulties, nations are constantly increasing the proportion of renewable energy sources that they include in their overall energy mix [1]. Wind energy has emerged as a potential alternative among the vast portfolio of renewable energy sources due to the fact that it is widely available and has a relatively low impact on the environment. According to the statistics values of installed wind generation capacity of 2023 that were published by the Energy Institute [2], the year 2022 witnessed the installation of 100 GW of new wind power, which represents a 9.1% increase in comparison to the installations that took place in 2021. After analyzing these numbers, it was discovered that the global growth rate for collective installed wind capacity from 2012 to 2022 was observed to be 12.9%. Listed below are the top five countries in terms of the pace of increase of installed wind capacity. In terms of their respective growth rates, Jordan is the country that leads the world with an astounding growth rate of 83.7%, followed by South Africa with a growth rate of 77.2%, the Russian Federation with a growth rate of 71.6%, Uruguay with a growth rate of 39.9%, and Pakistan with a growth rate of 38.3%. Significantly, it should be emphasized that Pakistan exhibits the highest growth rate among the countries in the Asia-Pacific region in this context.There is a lack of standardized datasets for forecasting wind energy in Pakistan. This impedes the progress of developing precise and efficient machine learning based forecasting algorithms for wind energy forecasting specifically tailored for Pakistan.

The significance of wind power generation in the current energy scenario is substantial, fueled by the worldwide transition towards sustainable energy sources. Nevertheless, the inherent fluctuations and irregularity of wind resources present significant challenges impede the successful incorporation of wind power into the electric grid. The key obstacles to establishing widespread integration of wind energy into the electricity grid are efficient scheduling, effective management, and optimal utilization of wind power. To tackle these problems, it is imperative to develop precise wind energy forecasting systems that may improve the reliability of wind power output. This is essential for maintaining grid stability, maximizing energy production, ensuring economic feasibility, and promoting environmental sustainability.

1.2 Scope

The thesis endeavors to clarify the viability and efficiency of artificial intelligence (AI)-driven models for wind energy prediction in the setting of Pakistan. The research endeavors to create predictive models that are specifically customized to the intricate details of Pakistan's wind patterns by utilizing advanced artificial intelligence techniques, meteorological insights, and historical wind data. The goal goes beyond simple theoretical research; actual wind power generating data will be used to validate and evaluate these models.

1.3 Goals

This work is dedicated to addressing the specific challenges of wind energy forecasting in Pakistan through the development of an AI based Wind Energy forecasting model. The ultimate goal is to play a pivotal role in optimizing the utilization of wind energy in Pakistan, fostering sustainable energy practices and addressing the specific challenges faced by the country in its pursuit of renewable energy sources. Through a comprehensive case study, this research article aims to explore the feasibility and effectiveness of AI-driven wind energy forecasting in the Pakistani context. The study will leverage historical wind data, meteorological information, and AI techniques to construct predictive models customized for the distinct characteristics of wind patterns in Pakistan. By evaluating the accuracy and reliability of these models against real-world wind power generation data, valuable insights can be gained into the capabilities of AI-driven forecasting for enhancing energy planning and management in the country. To achieve the aforementioned goals, We will summarize innovations of our work as follows

- Applied clustering analysis to eliminate outliers and enhance predicting accuracy.
- Proposed knee-point detection based algorithm for Automation of Eps value required for clustering.
- Engineered features that can potentially capture temporal patterns and structures.
- Applied Recursive feature elimination to eliminate redundant features from dataset before feeding it to XGBoost for prediction to improve accuracy of model.
- This work introduces an improved forecasting model called enhanced DBSCAN-RFE-XGBoost. The model utilizes clustering analysis based on enhanced DBSCAN approach and

Recursive Feature Elimination (RFE) to enhance the forecasting capability of the XGBoost model for wind power forecasting. The analysis of the findings demonstrates that the performance of this model outperforms other benchmark models.

1.4 Observations

Wind energy forecasting is a vital component in the contemporary energy landscape, driven by the global transition towards sustainable and renewable energy sources. The inherent variability and intermittency of wind resources pose significant challenges to the seamless incorporation of produced wind power into the electric grid. This work explores the pivotal role of wind energy forecasting in ensuring grid stability, optimizing energy production, enhancing economic viability, and contributing to environmental sustainability. By delving into the technological advancements and innovative methodologies underpinning accurate wind energy forecasts, This study emphasizes the crucial significance of forecasting systems within the context of the ongoing renewable energy revolution. Accurate wind energy forecasting is paramount for maintaining grid stability and ensuring a reliable supply of electricity. In a grid increasingly reliant on intermittent renewable sources, forecasts enable grid operators to anticipate fluctuations in wind power generation, facilitating effective resource allocation and mitigating the risk of power outages [3].

Wind farm operators rely on precise forecasts to optimize energy production. By leveraging forecasts to align turbine operations with expected wind conditions, operators can improve energy capture efficiency, reduce maintenance costs, and extend the operational lifespan of equipment. The economic feasibility of wind energy projects hinges on the reliability of forecasting systems. Investors and project developers utilize accurate forecasts to make informed decisions about project locations and financing structures, thereby minimizing financial risks and bolstering investor confidence. Wind energy forecasting plays a pivotal role in the seamless integration of wind power into the electrical grid. Grid operators use forecasts to anticipate wind power output fluctuations and make real-time adjustments to maintain grid stability, reducing the need for backup fossil fuel generation and facilitating the transition to renewable-dominated grids [4].

Precise wind energy forecasts enable wind farm operators to participate effectively in energy markets. By bidding their expected energy production into electricity markets, operators enhance market efficiency and revenue generation potential, while also reducing market price volatility. Accurate wind energy forecasting contributes to environmental sustainability by enabling the replacement of power generation using fossil fuel with clean, renewable sources. This transition results in reducing greenhouse gas emissions, align with global initiatives to address and mitigate climate change. The development of wind energy forecasting technologies stimulates innovation across multiple fields, including meteorology, data science, and renewable energy systems. Continuous research and development efforts lead to more sophisticated forecasting models, benefiting not only the wind energy sector but also scientific and technological advancements at large. Artificial Intelligence (AI) [5] technologies have revolutionized various sectors, and their potential in addressing complex and dynamic challenges in wind energy forecasting is increasingly being recognized. The fusion of advanced machine learning algorithms, big data analytics, and meteorological insights has paved the way for AI-driven wind energy forecasting [6], enabling stakeholders to make informed decisions in the realm of energy management, grid stability, and investment strategies.

Traditional forecasting methods, often based on meteorological models, struggle to capture the complex and non-linear interactions influencing wind patterns. Herein lies the motivation for incorporating AI-driven approaches. By harnessing the power of AI, this research seeks to address the limitations of conventional forecasting methods and contribute to the realization of a sustainable energy future in Pakistan. However, making the most of wind power requires accurate predictions of wind conditions, a task that has been greatly improved through employing AI/ML techniques, including Neural Networks, Support Vector Machines, and Random Forests, are increasingly instrumental in refining wind energy forecasts [7]. These algorithms can capture intricate relationships between various meteorological variables, historical wind patterns, and their impact on wind power generation. By processing vast datasets encompassing historical weather records and wind turbine performance, machine learning models can discern complex patterns and optimize forecasting accuracy [8].

1.5 National Interests and Benefits

Wind energy forecasting is highly relevant to national needs as it supports the efficient assimilation of wind power into the energy mix. Accurate forecasts aid in optimizing grid management, ensuring a reliable and stable electricity supply. They contribute to energy security, cost reduction, and carbon emissions reduction goals. Wind energy forecasting also supports the growth of renewable energy industries, job creation, and economic development, while helping nations transition towards a sustainable and low-carbon future. Conducting research and improving wind energy forecasting capabilities can benefit Pakistan by reducing dependency on foreign expertise and fostering local expertise and knowledge in wind energy management. It will make ways to save costs, to be more self-sufficient and the development of a skilled workforce within the country's renewable energy sector. It aligns with national objectives of promoting domestic resources and sustainable development in the energy sector.

1.6 Wind Power Installations

The data from the Global Wind Report [9] over the past decade reveals a clear upward trajectory in global wind power installations. Notably, the figures show consistent year-on-year growth in installed wind power capacity, with only minor fluctuations. The surge from 36 GW in 2013 to 77.6 GW in 2022 reflects a doubling of global wind power capacity over this period. This impressive growth is indicative of several factors driving the adoption of wind energy. First and foremost is the growing awareness of climate change and the need to reduce carbon emissions. It is anticipated that this trend will continue to gain rise in the coming years as the world strives to meet its sustainability and environmental targets.

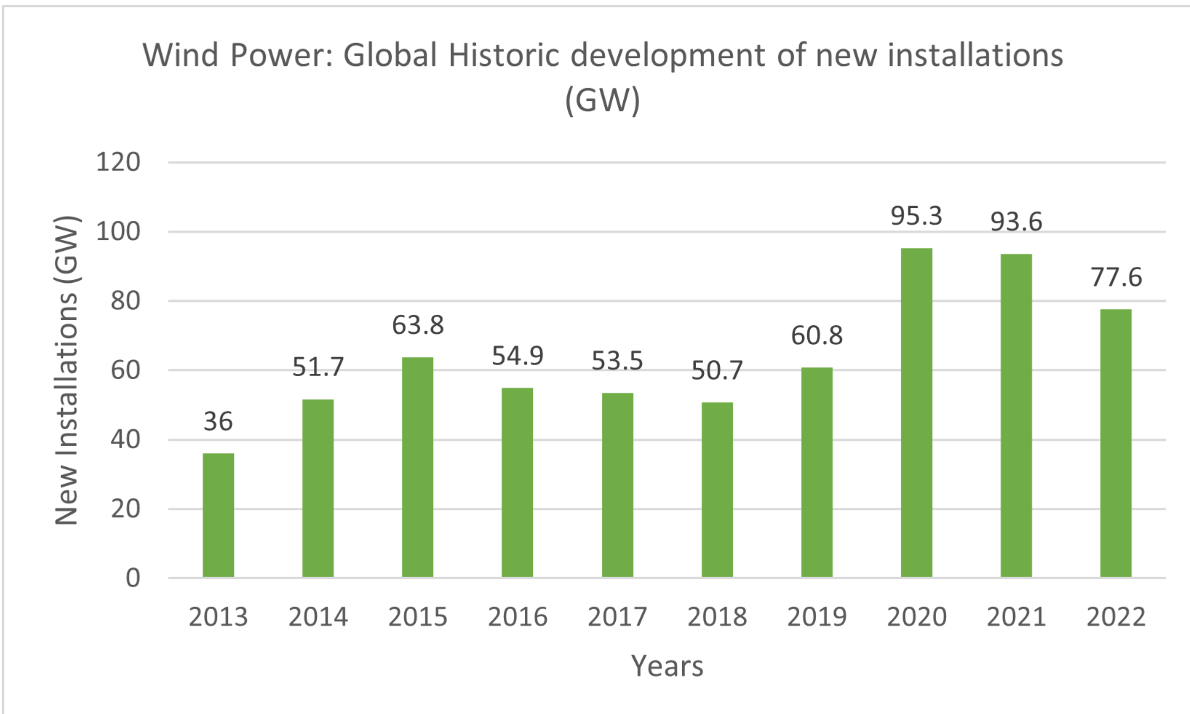


Figure 1.1: Global Wind Power New Installation

In Pakistan, there has been a noticeable and significant trend towards the adoption of wind energy as a critical component of the country's power generation mix. This trend reflects Pakistan's commitment to addressing energy challenges and its desire to harness the potential of renewable energy sources. According to data from the National Electric Power Regulatory Authority (NEPRA), the historic development of wind power capacity in Pakistan has shown substantial growth in the last ten years. Over the last decade, NEPRA's industry state report [10] highlights Pakistan's noteworthy emphasis on wind energy. The nation has made commendable strides in this sector, evident in the data revealing its growth from a modest 50 MW in 2013 to a significant 1,838 MW in 2022. This remarkable progression underscores Pakistan's commitment to advancing its wind energy landscape.

Notably, the addition of 590 MW in 2022 reflects the country's ongoing dedication to expanding its wind energy infrastructure. This consistent growth aligns with Pakistan's broader sustainability and environmental objectives. Looking ahead, it is anticipated that Pakistan will sustain its momentum, further enhancing its wind energy sector in the years to come. This trajectory not only contributes to the nation's energy diversification but also reinforces its position in the global shift towards cleaner and more sustainable energy sources.

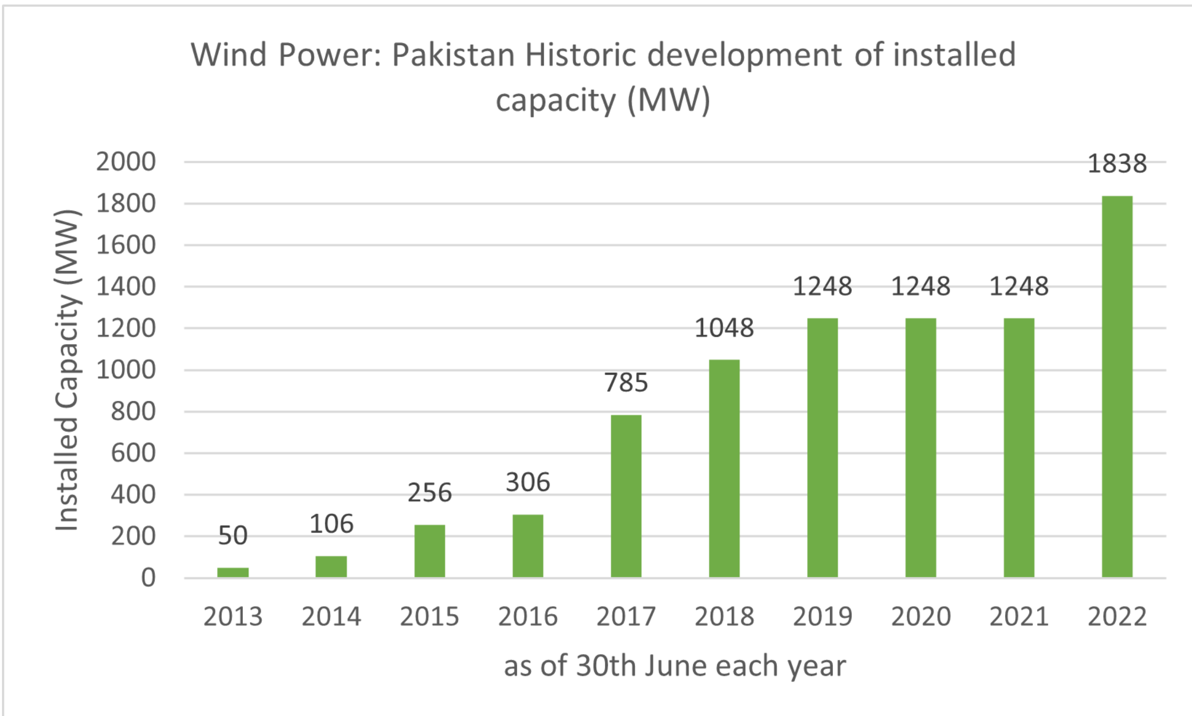


Figure 1.2: Installed Wind Capacity, Pakistan

1.7 Research Objective

Following are the objectives of this work.

1. To build AI driven Model for Wind Energy Forecasting.

2. Compare the model with state-of-the-art technique.

1.8 Thesis Structure

This work is organized into five main chapters. The first chapter provides an introduction to the topic, covering aspects such as motivation, contributions, goals, observations, national interests, and benefits. The chapter concludes by outlining the structure of the thesis. Moving on to the second chapter, a comprehensive literature review is presented. The third chapter delves into the methodology employed to reach the set milestones, detailing the data pre-processing techniques, feature selection and provides a detailed examination of the proposed models. The fourth chapter provides discussing the outcomes of proposed model and comparing their results through tables and graphs. Finally, the fifth and concluding chapter offers a conclusion and discussing potential future contributions in this field.

Chapter 2

LITERATURE REVIEW

2.1 Related Work

Numerous approaches to wind (WS/WP) forecasting exist, categorized in 2.1 Forecast models are also classified by timescales (ultra-short, short, medium, long-term) as shown in 2.1, aiding real-time control, planning, trading, and maintenance scheduling in power systems.

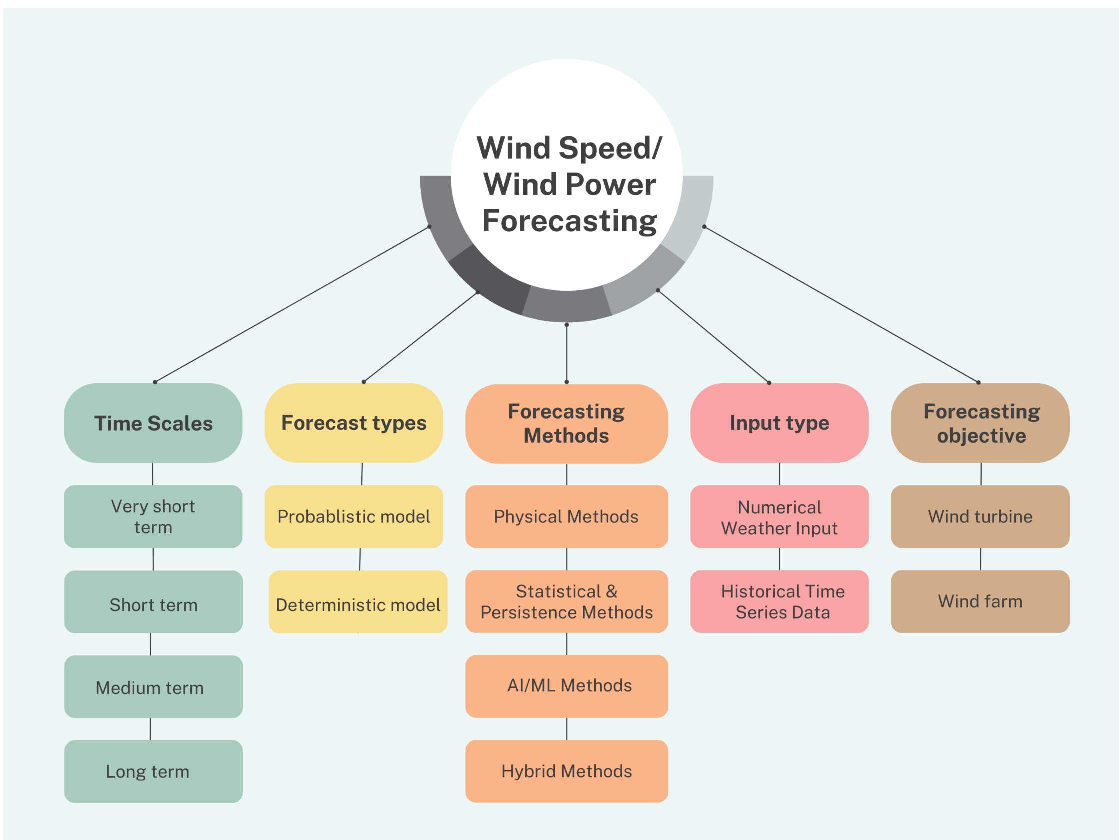


Figure 2.1: WS/WP Forecasting Classification

Forecasting models for WS/WP can also be classified into two types depending on their forecasting goals: wind turbine based forecasting and wind farm based forecasting. Wind turbine models estimate individual turbine output, while wind farm models predict the combined power output of multiple turbines, the latter being more complex [11]. Deterministic models offer point forecasts but may produce unsatisfactory results due to environmental complexity. Probabi listic models provide decision-makers with prediction intervals (PIs) [12], classifiable as parametric or non-parametric.

Parametric methods [13] assume forecast uncertainty follows a specific probability distribution, includes Gaussian Process (GP), delta method, and MVE (mean variance estimation). Quartile Regression (QR) and Lower Upper Bound Estimation (LUBE) [14] are the non-parametric methods that are distribution free and generate PIs with the help of techniques like quality measures for LUBE based models and asymmetric Laplace loss for QR models.

Forecasting models can be classified into four types based on their modeling theory [8]: physical models, traditional or statistical models, artificial intelligence (AI) / Machine learning (ML) based models, and hybrid models. In these models, the physical models are the models that forecast the metrological factors like air pressure, temperature and humidity for the WS forecasts [15], as they are good in this medium and can provide long term forecasts. They use Weather Research Forecasting (WRF) and Numerical Weather Prediction (NWP) techniques for these forecasts that enable them to forecast the wind power using the turbine curves. Other are the traditional models that use Autoregressive Integrated Moving Average (ARIMA), Frictional-ARIMA (f-ARIMA) Autoregressive Moving Average (ARMA) [16] and Hammerstein Autoregressive model. Their main focus is on the WS/WP time series data linear fluctuations to be able to perform well in short term to enable the short term forecasting.

Many of the researchers have tried to find out some wind power forecasting models that are more accurate and reliable. Along with the two types of forecast (probabilistic and deterministic), literatures also discuss the four forecasting models as stated above (artificial intelligence based models, statistical models, physical models and hybrid model). The physical models use computational weather prediction to forecast wind speeds. They use meteorological data and computational methods to make these predictions, similar to conventional wind speed forecasting techniques. While the traditional or the statistic models use the statistical equations that analyze the relationships in the historical data and reveal enhanced performance in short term forecasting. It performs this function with the help of different forms like Autoregressive (AR), Persistence model (PM), Moving Average (MA), ARIMA, ARMA and SARIMA. Statistical models tend to be more accurate for short-term forecasts, diminishing in accuracy for longer forecast horizons.

The increased computation capability of processors and advancements in AI approaches have resulted in the extensive use of AI or ML-based models for wind power forecasting. The literature discusses different Supervised learning models like Extreme Learning Machine (ELM), Support Vector Machine (SVM), Kernel ELM (KELM) and Least Squares Support Vector Machine (LSSVM) models [17]. The usage of SVM models in wind energy and solar forecasting was reviewed in detail [18]. The research paper by [19] introduces a novel method that combines Extreme Learning

Machine (ELM) with the Adaboost algorithm, and further enhances its performance using particle swarm optimization. The paper [20] presents a self-adaptive kernel Extreme Learning Machine (KELM) that aims to enhance forecasting accuracy, minimize re-training expenses, and optimize training efficiency. [21] has utilized random forest [RF] for hour ahead wind power forecasting. [22] has employed Stacked denoising autoencoder with LSTM.

Along with the above discussed model, the paper also discusses different types of Artificial Neural Networks (ANNs). Both [23], [24] have used Back Propagation Neural Network (BPNN) in their research methodologies. In [25], modified multi-objective dragonfly algorithm helped to improve the loads of the BPNN. In [26], a hybrid decomposition technique was used for multi-step wind speed forecasting by constructing an upgraded BPNN. The study by [27] used three different fundamental predictors, namely Multi-layer Perceptron (MLP), Long Short-Term Memory (LSTM), and ARIMA, to forecast wind speed (WS). The combination weights of these predictors were adjusted using an intelligent optimization technique. To enhance the accuracy of forecasts, [28] employed the optimization of the Wavelet Neural Network through the utilization of the Multi-objective Sine Cosine algorithm. In addition, the authors of [29] have forecasted wind velocity by employing three artificial neural networks: an Elman neural network, a generalized regression neural network and cascade backpropagation neural network. The results were enhanced using the multi-objective particle swarm optimization technique and merged with an echo state network to provide the final wind power predictions. In order to make precise predictions about wind speed at a specific location, [30] introduced a multidimensional spatial-temporal graph neural network method. This method utilizes a Wind-Transformer as a node in a graph neural network to effectively gather information about local and surrounding wind speeds. Different forms of data like the wind direction, speed, air pressure and temperature is combined in it.

Moreover, deep neural networks are gaining substantial attention owing to their remarkable proficiency in addressing intricate nonlinear problems. The literature extensively discusses Long Short-Term Memory (LSTM) [31], Bidirectional LSTM (BiLSTM) [32], Gated Recurrent Unit (GRU) [33], Convolutional Neural Network (CNN) [34], Deep Belief Network (DBN) [35], Bidirectional GRU (BiGRU) [36], and Auto-encoder (AE) [37].

Recent research in wind power forecasting has demonstrated a significant shift towards employing hybrid models, integrating traditional numerical weather prediction techniques with modern machine learning algorithms. These models leverage the qualities of both methodologies, boosting the accuracy and dependability of wind power projections by properly reflecting the complex and dynamic nature of atmospheric conditions. As researchers continue to research novel combinations and optimizations, hybrid models emerge as a possible route for increasing the progress in wind power forecasting. Optimization methods such as Grey Combination Model and PSO-SVR [38] , Rao-1 algorithm [39] , and enhanced Jaya algorithm [40] have been introduced into forecasting frameworks to enhance overall predictive skills. Wavelet Transform (WT) and Autoregressive Moving Average (ARMA) have been integrated by Kaur et al. [41].

Recent research in wind power forecasting has demonstrated a significant shift towards employing hybrid models, integrating traditional numerical weather prediction techniques with modern machine learning algorithms. These models leverage the qualities of both methodologies, boosting the accuracy and dependability of wind power projections by properly reflecting the complex and dynamic nature of atmospheric conditions. As researchers continue to research novel combinations and optimizations, hybrid models emerge as a possible route for increasing the state-of-the-art in wind power forecasting. Optimization methods such as PSO-SVR and Grey Combination Model [38] , Rao-1 algorithm [39] , and enhanced Jaya algorithm [40] have been introduced into forecasting frameworks to enhance overall predictive skills. Wavelet Transform (WT) and Autoregressive Moving Average (ARMA) have been integrated by Kaur et al. [41]. It used a method called Wavelet Transform to analyze the wind speed data. First, the nonlinear time series was decomposed into smaller subsequences. Then it applied individual ARMA model to each subsequence. By combining the results of several ARMA applications, were able to get the final prediction, revealing that this hybrid model significantly enhanced the precision of the forecasts. In order to tackle the challenges of making multi-step forecasts for the near term, Liu et al. [42] improved wind power generation forecasting by combining K-shape, K-means, CNN, and Gate Recurrent Unit (GRU) into a complete hybrid model. The operational dataset was used to validate this model, which takes physical, geographical, and temporal information as input tensors. Xiong et al. [43] used Attention Mechanism to resolve the model issue not noticing the importance of the input data by assigning weights to physical attribute data. They minimized the likelihood of erroneous predictions caused by data mixing and preprocessed datasets by using CNN to extract relevant features for the short term and LSTM to capture trends in localized multidimensional characteristics.

Chapter 3

METHODOLOGY

We have introduced a new model for wind power forecasting based on artificial intelligence and data-driven. The model's flowchart is shown in 3.1. There are following two main steps, each tailored to tackle a different problem: (i) Clustering stage (ii) Features selection and Model training. The Clustering stage employs DBSCAN clustering Algorithm and proposed our enhance DBSCAN method to eliminate outliers from the dataset, it add value to refining predictions by eliminating outliers and enhancing accuracy. In Features Selection and Model Training stage, we have utilized Recursive Feature Elimination (RFE) coupled with XGBoost to elevate the model's precision in wind power predictions. This dynamic duo refines the feature set by iteratively eliminating redundant variables. The outcome is a lean and optimized set of features, ensuring the model's heightened capacity for generating accurate wind power predictions. The ensuing points highlight our forecasting model's contributions.

- Applied clustering analysis to eliminate outliers and enhance predicting accuracy.
- Proposed knee-point detection based algorithm for Automation of Eps value required for clustering.
- Engineered features that can potentially capture temporal patterns and structures.
- Applied Recursive feature elimination to eliminate redundant features from dataset before feeding it to XGBoost for prediction to improve accuracy of model.

This model represents a sophisticated integration of clustering analysis and recursive feature elimination (RFE), strategically employed to augment the forecasting capabilities of the XGBoost model. The utilization of clustering analysis ensures effective outlier removal and improved data accuracy, while RFE systematically refines the feature set, optimizing the model for enhanced predictive performance.

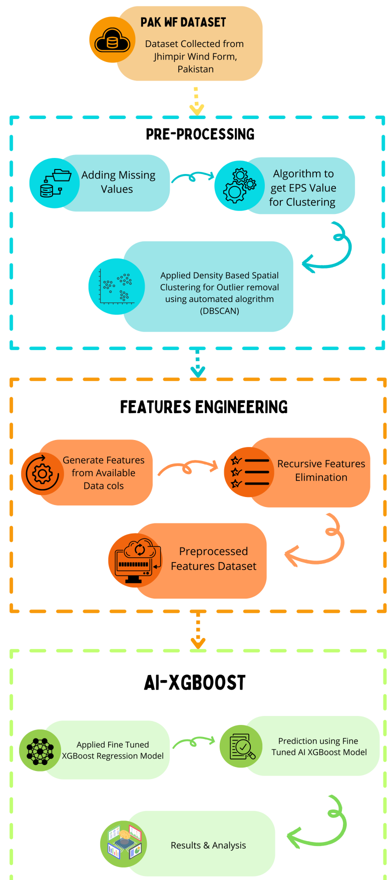


Figure 3.1: Proposed Methodology Structure

3.1 Outlier Removal Using Proposed Algorithm with DBSCAN

There are different clustering techniques that have been presented in the literature, including k-means [44], mean shift [45], DBSCAN [22], spectral clustering [46], mixes of Dirichlet model [47], clustering based on supervised learning [48], and clustering by local cores [49]. Adaptive to cluster forms, noise handling, and complicated pattern discovery—these are the hallmarks of DBSCAN, a strong clustering method. Due to its robustness, DBSCAN is able to adapt to varied densities in training samples and effectively filter noise, making it quite effective for wind power forecasting. Two parameters are required: the scanning radius reffered as Eps and the density threshold reffered as MinPts, that assesses whether a point to be considered as a core point.

The algorithm 1 shows how the conventional DBSCAN clustering algorithm works. Step one is to go through the dataset point by point, mark them as visited, and then use the EPS parameter to find their neighborhood. The point is classified as noise if there are less than MIN_PTS points in the neighborhood; otherwise, it starts a new cluster. In order to grow the cluster and include nearby points, the algorithm iteratively investigates the area around the core points. Every visited point and its neighbors help to form clusters, as outlined in the iterative process. A set of clusters (CS) found in the provided dataset by DBSCAN is the output.

Algorithm 1 AL:DBSCAN Clustering

```
Require: Dataset D, ε, MIN_PTS
Ensure: Clusters Set CS
procedure DBSCAN(D, ε, MIN_PTS)
    for Each traversed point dp ∈ X do
        Set dp as Visited
        N ← NN(dp, ε)
        if |N| < MIN_PTS then
            Set dp as Noise
        else
            CS ← {dp}
            for dp′ ∈ N do
                N ← N \ dp′
                if dp′ is not visited then
                    Set dp′ as visited
                    N′ ← NN(dp′, ε)
                    if |N′| ≥ MIN_PTS then
                        N ← N ∪ N′
                    end if
                end if
                if dp′ ∉ CS then
                    CS ← CS ∪ {dp′}
                end if
            end for
        end if
    end for
end procedure
```

The literature advises employing a K-distance plot or other optimization methods to identify parameters values. Using the formula from Euclidean geometry, we can find the distances between any two points in a k-distance plot. Equation 3.1 shows that the K-th value is obtained by sorting these distances in ascending order to create a distance array dist(x). Visualizing the k-distance plot is the traditional method for locating the knee point. In [22] they developed a novel approach for deep feature extraction. Their proposed method relies on visualizing the k-dist plot to identify the EPS value. After that, LSTM neural networks were fed the deep features retrieved from the raw historical data in order to make predictions.

$$K_{\text{dist}} = \{\text{dist}_{\text{MinPts}}(x) \mid x \in D\} \tag{3.1}$$

An comprehensive visualization with three separate clusters is produced by applying the DBSCAN clustering framework to a subset of the data. Figure 3.2 shows the clustering results when the Minpts value is 20 and the EPS value is 0.8. The diagram gives a detailed picture of the various types of cluster points. Solid circles encircled by solid line circles distinguish core points, which are important in cluster definition, and emphasize their importance in the clustering process. A solid triangle inside a dotted line circle represents a non-core point, indicating the outer limits inside the cluster. A triangle with a cross mark on top makes noise points, or points that don't fit the clustering patterns, easy to detect. This visual assistance makes it easy to identify data points that don't belong in the specified clusters, which provides useful context for understanding the dataset's subtleties.

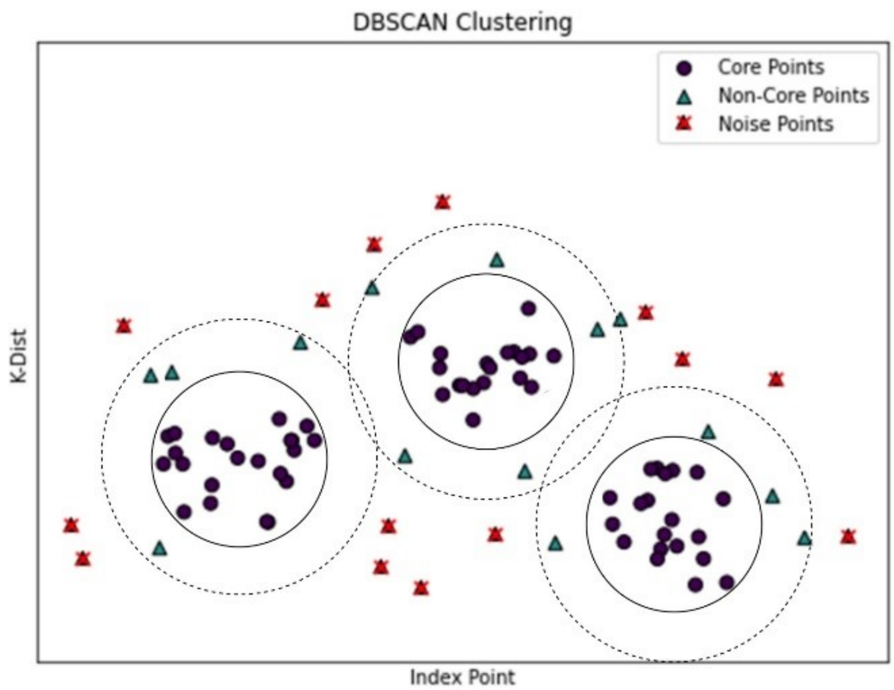


Figure 3.2: DBSCAN Process

Also, the in-depth view of one cluster in Figure 3.3 helps to clarify things even more. A solid red circle stands out at the core block's center, drawing attention to its a vital role in the cluster. Owing to this visualization and the provided Minpts and EPS data, we can see a larger context of the clustering dynamics, which simplifies and clarifies the complex process.

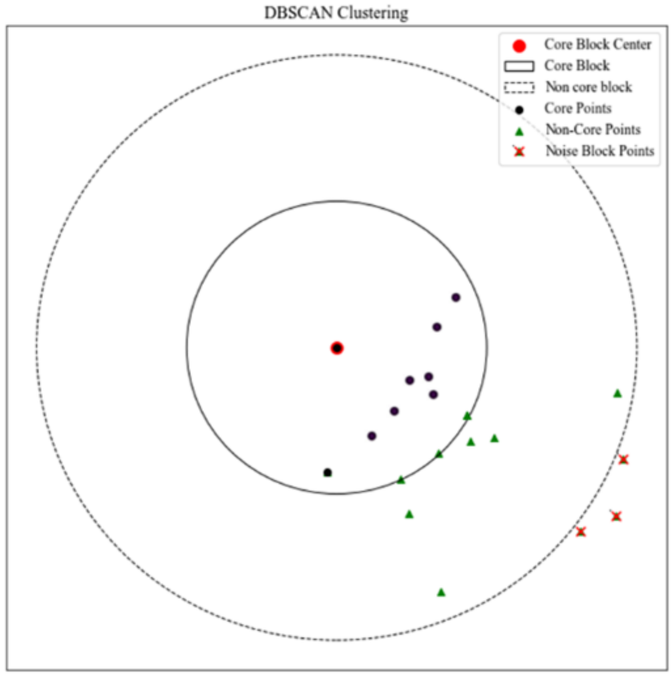


Figure 3.3: DBSCAN Cluster View

Our approach stands out in its distinct advantage over traditional methods of visualizing k-distance plots, as discussed in the literature [22], [50]. In contrast to the conventional approach, our proposed methodology integrates enhanced DBSCAN, introducing a more efficient and automated process. Notably, our approach eliminates the need for manually interpreting k-distance plots.

In our methodology, the determination of the knee point in the k-distance plot is revolutionized through the implementation of an automated process for calculating Eps values as shown in Figure 3.4. This innovative step is achieved through our algorithm, which operates in a data-driven manner. Rather than relying on the visual analysis of k-distance plots, our algorithm mathematically computes the Eps value for DBSCAN. Specifically, it identifies the maximum curvature along the k-distance plot, as expressed in Equation 3.2, 3.3, and 3.4. This data-driven calculation not only streamlines the process but also enhances the accuracy of Eps determination, contributing to the overall robustness and efficiency of our approach in comparison to conventional visualization techniques.

Slope

$$m_i = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \tag{3.2}$$

Second Derivative

$$D_i^2 = \frac{m_{i+1} - m_i}{m_{i+2} + m_{i+1}} \tag{3.3}$$

Knee Point

$$K \quad D_K^2 = \min(D^2) \tag{3.4}$$

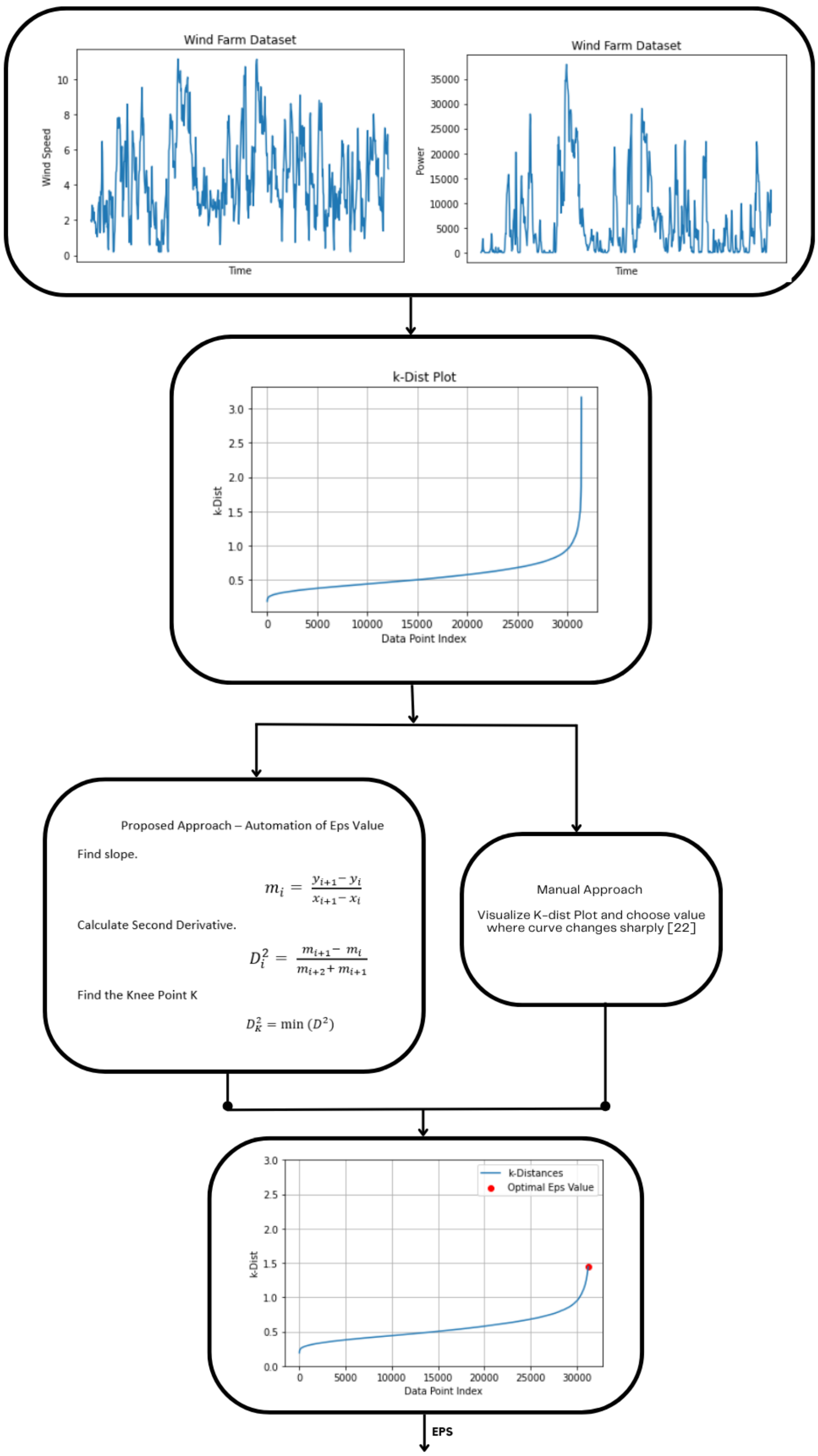


Figure 3.4: Proposed Enhanced DBSCAN vs Traditional clustering approach

To identify outlier points without affecting with the cluster determination process, a suitable MinPts value is required. The accuracy of the cluster identification process can be affected by a smaller MinPts number, which could be used for outlier detection, or the reverse, depending on the situation. Hence, we have chosen a marginally higher MinPts value of 10. After we found the two parameters needed for DBSCAN, we ran the algorithm on our dataset. It removed samples from outliers and helped us choose more representative samples to use as training for our forecasting model. Not only does this cut down on effort, but it also makes parameter selection less susceptible to human bias. While k-distance charts have long been the go-to for manually selecting Eps values, doing so can introduce inaccuracies or slow down the algorithm if not done correctly.

3.2 Application of Recursive Feature Elimination with XGBoost

Previous research conducted by researchers has used a variety of machine learning methods for forecasting, with algorithms chosen for their strengths in certain areas and compatibility with existing datasets. Factors such as data characteristics, input variables, and particular components are used to select these models. As an example, [51] suggested a Support Vector Machine classifier for electric price predictions that uses differential evolution, based on the features of the input data and the presence of high frequency components. Similarly, researchers propose robust linear regression models for data when there is a high degree of linear correlation between the input features and the predicted variables. In his proposed method for wind energy forecasting, [52] used multiple linear regression in his study. Iteratively building decision trees is how XGBoost optimizes model accuracy; the algorithm is prominently accurate in predictions. By modifying the weights of misclassified samples, XGBoost uses gradient boosting to improve predictions as shown in Figure 3.5, in contrast to conventional trees that are susceptible to overfitting. Its superiority over other forecasting models is demonstrated by its capacity to manage noisy data. When it comes to training and improving the models' robustness, feature selection is just as important as model selection. The model's robustness can be affected by redundant features. That is why these superfluous features are targeted for removal using the recursive feature elimination method.

A loss function and a regularization term make up XGBoost's objective function for regression, which is rooted in the gradient boosting framework. Along with the second-order derivative, the loss function measures the model's data fit, while the regularization term limits the model's complexity by penalizing an excess of leaf nodes. The penalty increases with the number of leaf nodes, effectively limiting the complexity of the model. The objective function can be generally stated mathematically as in Equation 3.5.

$$\hat{y}_i^m = \sum_{j=1}^{N} f_j(x_i) = \hat{y}_i^{(m-1)} + f_t(x_m) \tag{3.5}$$

Given that $\hat{y}_i^m$ represents the outcome of the forecast for the m-th iteration of sample i. f_j is the outcome of the prediction made by the j-th tree on sample i, and j is the set of CART trees. $f_t(x_m)$ is the outcome of the m-th tree, while $\hat{y}_i^{(m-1)}$ is the outcome of the first $m - 1$ trees. The loss function and regularization term defined as

$$O = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{j=1}^{T} \Omega(f_i) \tag{3.6}$$

Here, $l(y_i, \hat{y}_i)$ is the loss function, which measures the difference between the true label y_i and the predicted value $\hat{y}_i$, and N represents the total number of training instances. $\Omega(f_i)$ is the regularization term for the i-th tree, T represents the number of trees in the combination. The Taylor series of the loss function is often approximated to the second order by removing the constant term. When there are m CART trees, the objective function takes the following form:

$$O = \sum_{i=1}^{N} \left[d_i f_m(x_i) + \frac{1}{2} h_i f_m^2(x_i) + \Omega(f_m) \right] \tag{3.7}$$

$$d_i = \frac{\partial}{\partial y_i^{(t-1)}} l(y_i, \hat{y}_i^{(m-1)}) \tag{3.8}$$

$$h_i = \frac{\partial^2}{\partial (y_i^{(t-1)})^2} l(y_i, \hat{y}_i^{(m-1)}) \tag{3.9}$$

Where d_i and h_i are the first-order and second-order derivatives, respectively, and the regularization term in the above equation will be as follows:

$$\Omega(f) = \gamma N + \frac{1}{2} \lambda \sum_{j=1}^{n} W_j^2 \tag{3.10}$$

Where W_j is the weight fraction for the j-th child node of the tree f, n is the total number of nodes in tree f, γ and λ are custom parameters. γ represents the penalty term for the L1 norm, while γN controls tree complexity by adjusting the number of leaf nodes and their coefficients, thereby regulating model complexity. Additionally, λ is the penalty term for the L2 norm, where $\frac{1}{2}\lambda \sum_{j=1}^{n} W_j^2$ helps control the weight distribution among leaf nodes to prevent overfitting.

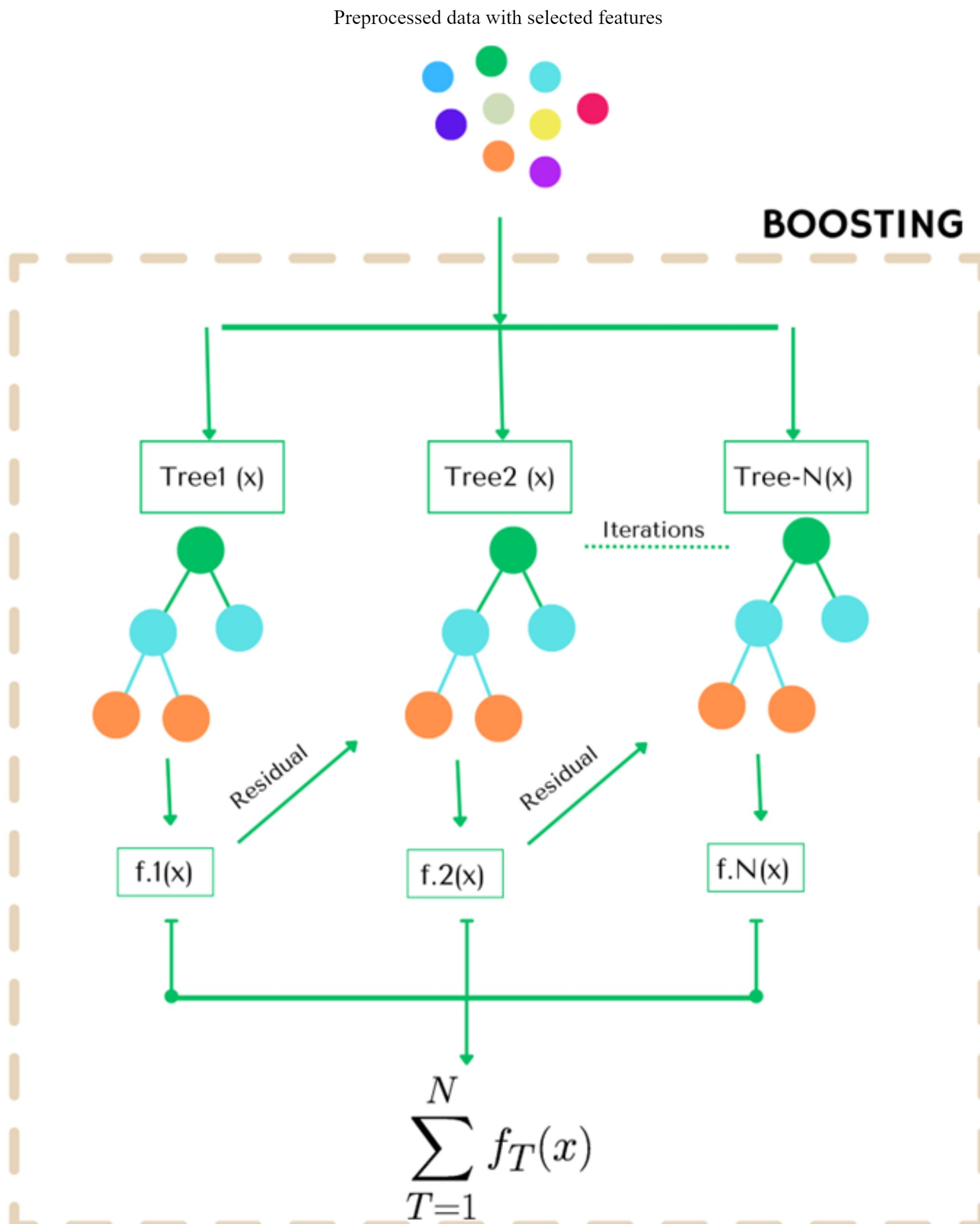


Figure 3.5: XGBoost Structure

3.3 PURPOSED DBSCAN-RFE-XGBOOST

An important factor influencing the accuracy of wind speed forecasts is the wind farm dataset. Wind farm data, however, contains abnormal samples. Due to noise introduction and lower learning efficiency, wind speed forecasting algorithms can be severely hindered by outliers and incorrect entries in wind farm statistics. In response, we offer a refined DBSCAN method for pre-processing data, which can detect outliers and remove them from the training set.

While density-based clustering is an advantage of DBSCAN, the effectiveness of the method is highly dependent on the value of the epsilon (eps) parameter. The determination of eps in traditional approaches frequently involves visually inspecting k-distance maps, which can lead to discrepancies and human error. In response to this difficulty, we present an enhanced DBSCAN method (as shown in algorithm 2 that uses an effective algorithm based on knee point detection to automate eps selection. Optimal clustering for wind power forecasts is guaranteed by this automation, which removes subjectivity. Models for predicting wind speeds frequently include a number of meteorological variables, such as wind speed, wind direction, temperature, humidity, and pressure. Having said that, the predictive power of every feature is not proportional to its own. Using recursive feature selection, a method that systematically assesses feature subsets and selects the top 10 features from our dataset that show the best predictive potential, we are able to discover the most important features. This focused method increases prediction accuracy while decreasing model complexity and increasing computational efficiency.

In algorithm 2 The first phase of is to process the dataset, which is represented as F', by scaling its features with a scaling function, SF. The next step is to make a NN model and then use the `Model.Fit()` operation to fit it to the scaled dataset. Each point in the processed dataset $D(F')$ is then used to calculate the k-distance using the NN model. Then, we use the Knee Locator algorithm to find the knee point on the k-distance curve, and we set this knee point as the epsilon (EPS) parameter. At last, the dataset $\overline{D}$ that has been processed after outliers have been removed is acquired by using the standard DBSCAN algorithm in conjunction with the calculated EPS and $MinPts$. This mixed method uses a knee-point detection algorithm to improve DBSCAN clustering's optimal neighborhood radius determination.

Algorithm 2 Proposed Approach for Outlier Removal

Require: Dataset D, $MinPts = 10$
Ensure: Processed Dataset D'

Scaling $F' = SF(D)$
Compile Model $= NN(MinPts)$
Apply Model.Fit()
Compute k-dist using $D(F')$
$Dist = Model.kneighbors(F',$ return_distances=True)
K-dist $= Dist[:, -1]$
Set `curve` $=' convex'$, `direction` $=' increasing'$ for Knee Locator
Get
- $kneedle = knee.elbow$
- $Kneedle = KneeLocator$(range(len(K-dist)),
- $K - dist,$ curve=curve, direction=direction)$.elbow$

Set
- EPS= kneedle

$D' = DBSCAN(EPS, MIN\_PTS, D(F'))$

Algorithm 3 Recursive Feature Elimination (RFE)

Require:

- Dataset X of shape (m, n), where m is the number of samples, and n is the number of features.
- Target variable y of shape $(m, 1)$.
- XGBoost Model XGB.
- k, the desired number of features

Ensure: Subset of features selected for the model: x_{selected}

Initialize $x_{\text{selected}} = \emptyset$

Initialize model XGB

Segregate temporal feature

- X["year"] = X["DnT"].dt.year
- X["month"] = X["DnT"].dt.month
- X["day"] = X["DnT"].dt.day
- X["hour"] = X["DnT"].dt.hour

while len$(x_{\text{selected}}) < k$ **do**

  XGB.Train(X)

  Feature Importance $FI = XGB$.FI(X)

  Identify $ftop = \text{argmax}(FI)$

  Mark the selected feature: x_{selected}.append$(ftop)$

  Remove the feature from dataset

end while

Return the selected features subset x_{selected}

Algorithm 3 outlines a feature selection model designed to identify a subset of features that significantly contribute to the predictive performance of a model. Given a dataset X with dimensions (m, n), representing m samples and n features, along with the corresponding target variable y, the model employs XGBoost (XGB) as the underlying predictive model. The user specifies the desired number of features to be selected, denoted as k. The algorithm initializes an empty set x_{selected} and the XGBoost model. After that temporal features year, month, day and hour are segregated from feature Date and Time.It iteratively trains the XGBoost model on the dataset X, computes the feature importance (FI), identifies the feature with the highest importance (f_{top}), adds it to the selected features, and removes it from the dataset. This process continues until the desired number of features, k, is reached. The final output is the subset of selected features, x_{selected}, representing the most influential features for the given XGBoost model and dataset. This approach ensures that the selected features collectively contribute to the model's predictive accuracy. We build a prediction model with the best features found and clusters created by enhanced DBSCAN using XGBoost, a scalable and very efficient gradient boosting algorithm. A promising option for wind speed prediction, XGBoost can deal with nonlinear interactions and avoid overfitting. Our improved DBSCAN-RFE-XGBoost framework as shown in figure 3.7 provides a reliable method for wind power forecasting by combining XGBoost's predictive powers with automated eps selection and feature optimization. We demonstrate that our proposed model works well in situations where data quality is poor by including automatic estimation for missing data, XGBoost's built-in protections (such as regularization and random sampling), and accurate forecasts.

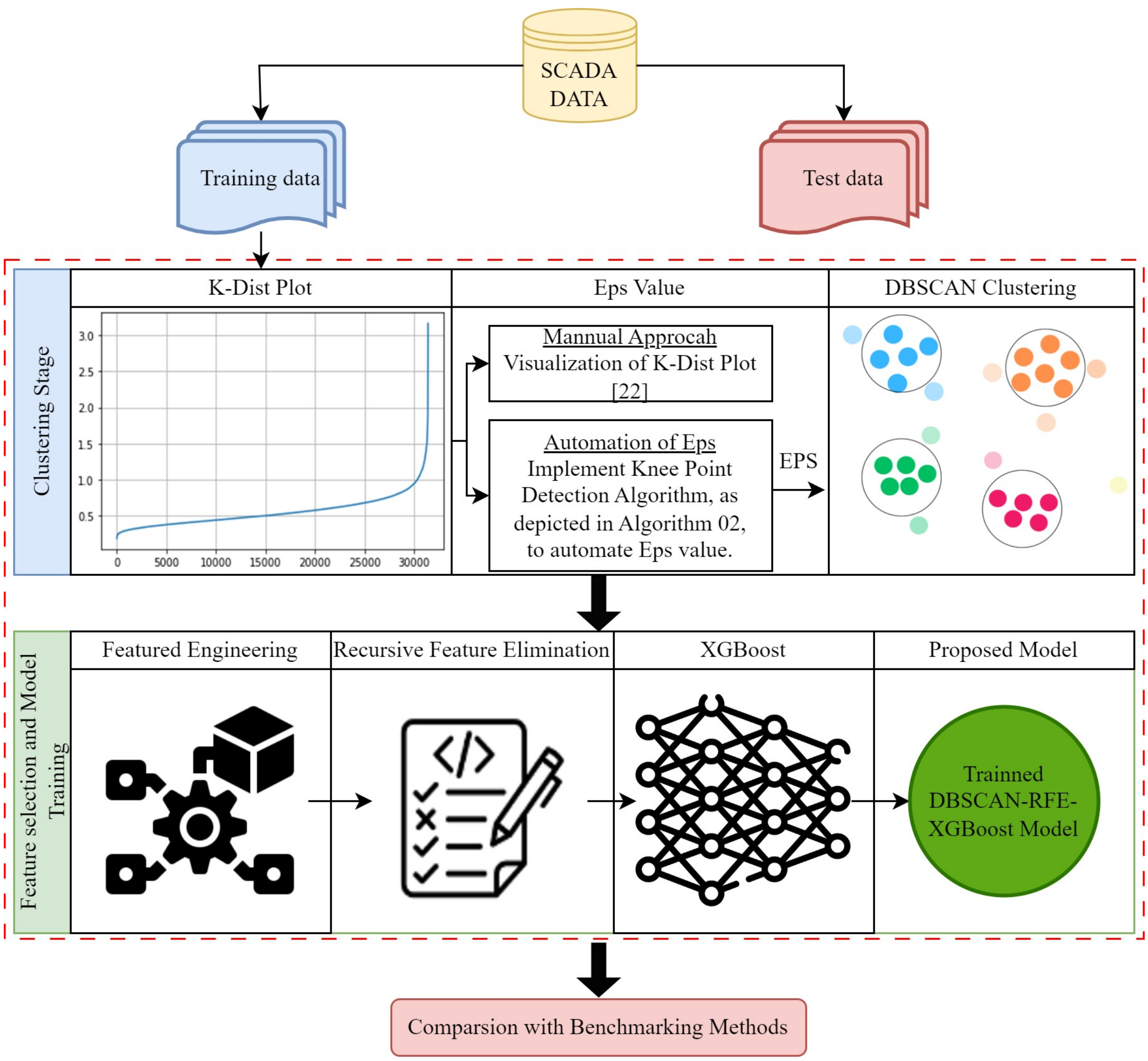


Figure 3.6: Proposed Enhanced DBSCAN-RFE-XGBoost

Our enhanced DBSCAN-RFE-XGBoost framework consists of two phases, as shown in Figure 3.7 During the first stage of data preprocessing, we have implemented a clustering algorithm based on density-based spatial clustering (DBSCAN). This algorithm automatically removes outliers from the dataset and calculates the eps value needed for DBSCAN clustering using a K-dist plot and the Knee Point Detection Algorithm. The second stage involves applying recursive feature elimination on the preprocessed dataset to select the most suitable features. These features will then be fed into the XGBoost algorithm, which can forecast wind power and offer comparisons with benchmarking methods.

3.4 Benchmarking Methods

After that, we compare the outcomes of our enhanced DBSCAN-RFE-XGBoost model to those of two well-known benchmark methods: SVMs and Long Short-Term Memory networks (LSTMs).

3.4.1 Long Short-Term Memory networks (LSTMs)

Long short-term memories (LSTMs) are widely utilized in wind power forecasting because to their exceptional ability to grasp temporal relationships. In addition to wind power forecasting, it finds extensive usage in load forecasting, price forecasting, and solar energy forecasting [53], [54], and [55], respectively. For the purpose of day-ahead wind power forecasting, Ko et al. [56] employ LSTMs.

The vanishing gradient problem is one obstacle that conventional RNNs have when trying to learn and remember data over long time periods; LSTMs were developed to circumvent this difficulty. LSTM's have a complex structure in which we have three gates with a memory cell. These three gates are the input gate, the output gate and the forget gate. At each time step t, the relevant information for new input (x_t) is determined by the input gate (i_t), the information to be discarded from the previous cell state (C_{t-1}) is decided by the forget gate (f_t), and the information to be outputted is controlled by the output gate (o_t). These gates and the input (g_t), which represents the new data that could be added to the cell state, are used to update the cell state (C_t).

Input Gate Activation (i_t): It decides which parts of the new input (x_t) and the previous hidden state (h_{t-1}) should be added to the cell state.

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{3.11}$$

Forget Gate Activation (f_t): It determines which parts of the previous cell state (C_{t-1}) should be forgotten or retained.

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{3.12}$$

Cell State Update (C_t): The cell state is updated by combining the information from the input gate (i_t) and the new information (g_t).

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \tag{3.13}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t \tag{3.14}$$

Output Gate Activation (o_t): It regulates the information that will be outputted as the hidden state (h_t).

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{3.15}$$

Hidden State Update (h_t): The final hidden state is computed by applying the output gate to the cell state.

$$h_t = o_t \odot \tanh(C_t) \tag{3.16}$$

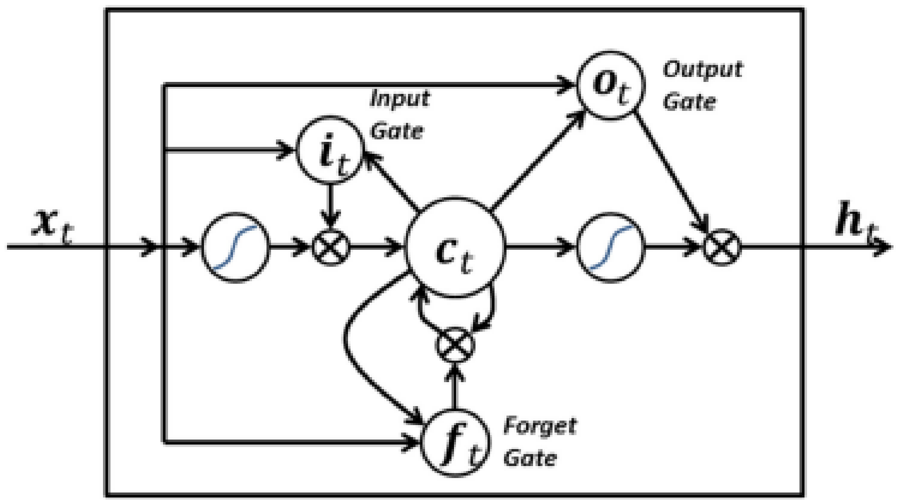


Figure 3.7: LSTM Network Cell

3.4.2 Support Vector Machines (SVMs)

A supervised machine learning technique, Support Vector Machine (SVM) excels at handling complex correlations and patterns, making it a useful tool for predicting. Many applications have made use of support vector machines (SVM), including wind speed forecasting [57], demand forecasting [58], stock price prediction [59], GDP forecasting [60], and several others. Achieving the maximum margin between classes while finding the best hyperplane to divide data points into them is its main goal. How far is the hyperplane from the closest data point in either class? That's the margin. A key component of support vector machines (SVMs) is their ability to detect support vectors, or data points that are critical for hyperplane positioning. Important in determining the margin are these support vectors, which are located around the decision border.

The equation of the hyperplane in a linear SVM can be expressed as

$$f(x) = w \cdot x + b \tag{3.17}$$

where $f(x)$ is the decision function that predicts the class of a given input x. w represents the weight vector, which is perpendicular to the hyperplane and determines its orientation. b is the bias term, allowing for the translation of the hyperplane. SVM finds the optimal values for w and b that maximize the margin while ensuring that all data points are correctly classified.

To gain insight into the workings of Support Vector Machines (SVM), Figure 3.8, the visualization portrays that an n dimensional space represent each data item, where 'n' signifies the number of features. Each data item, denoted as a feature within the dataset, is associated with a particular value that belongs to its specific coordinate in this multidimensional space. This representation allows us to visually comprehend the distribution and relationships among data points across the various features, forming a comprehensive understanding of the dataset's structure in the context of SVM.

In scenarios where achieving linear separation poses challenges, Support Vector Machines (SVM) offer a powerful solution by leveraging kernel functions. These functions help convert the input information into a higher-dimensional space. This allows us to find a hyperplane that is capable of separating the data. By employing kernel functions, SVM enhances its capacity to handle complex relationships and patterns that may not be linearly separable in the original feature space. This approach allows SVM to capture intricate decision boundaries, making it particularly advantageous in situations where a linear classification is not viable. The ability to map data into higher dimensions through kernel functions broadens the applicability of SVM, enabling it to address a wide range of real-world problems characterized by intricate and non-linear relationships among variables.

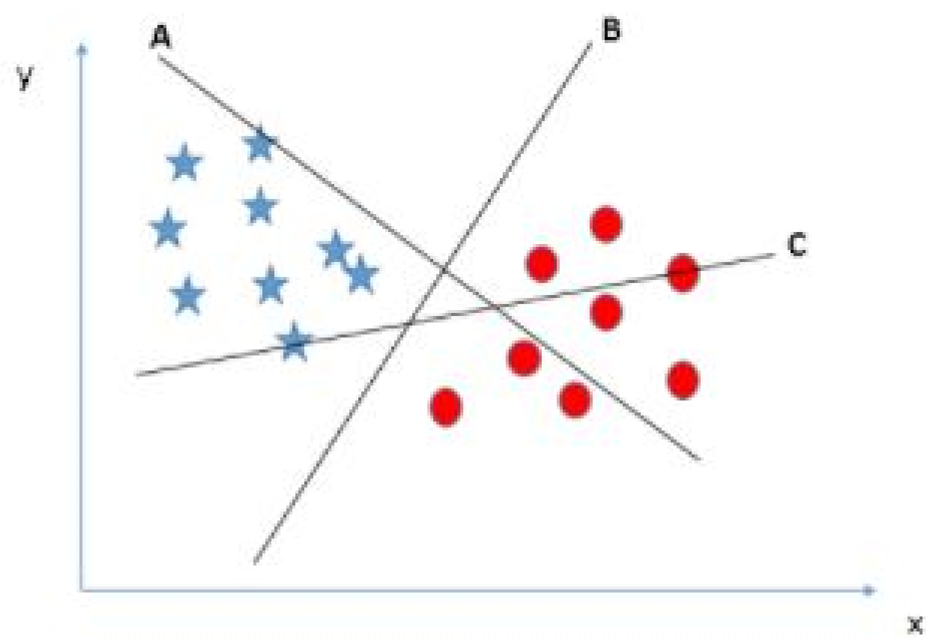


Figure 3.8: Visualization of Support vector machine

Chapter 4

RESULTS DISCUSSION

4.1 Dataset Description

We used a dataset obtained from the Three Gorges Wind power plant in Jhimpir, Pakistan, to evaluate the proposed model's performance using real-time data.Accurate date and time records spanning from January 1, 2015, to July 31, 2018, along with 31392 samples, make up the wind farm dataset, which contains environmental and meteorological parameters suitable for Wind Power Forecasting. The dataset also includes wind speed data from three separate sensors, wind direction, wind pressure, temperature, and humidity. An additional feature, denoted as "Date and Time" (DnT), has been thoughtfully integrated into our dataset, partitioned into year, month, day, and hour components enabling a more detailed and granular analysis of the data.

Notably, the wind speed measurements are recorded in "meter/second," exhibiting a dynamic range spanning from 0 m/s to 18.02 m/s. The wind direction data, collected at a height of 80 meters above ground level, is expressed in degrees and ranges from 0 to 356. An additional feature, denoted as "Date and Time" (DnT), has been thoughtfully integrated into our dataset, partitioned into year, month, day, and hour components. These temporally resolved variables will play a pivotal role within our AI-Xgboost model, facilitating a more intricate and granular analysis of the data. This integration empowers our model to draw on temporal trends and patterns, further enhancing its capabilities in wind power prediction and optimization.

The box plot, depicted visually in Figure 4.1, shows the monthly average wind speed. Different patterns become apparent when looking at the figure, especially in the IQR. The IQR values are much higher in May, June, July, and August, suggesting a wider range of wind speeds throughout the summer. The interquartile range (IQR) for July, for example, covers a wide range of wind speeds, from 8.75 m/sec to 12 m/sec.

On the other hand, November, December, January, and February had lower IQR values. Winter weather is typically more predictable and consistent, with wind speeds staying within a narrow range, thus this finding fits in with that. With the help of the box plot, we can see how the average wind speed changes with the seasons and how it trends overall. Figures 4.2, 4.3, 4.4, and 4.5 depict datasets representing wind power, wind speed, and corresponding weather parameters (temperature and humidity). These visualizations offer concise insights into the samples associated with each dataset. They provide a clear overview of the variations and trends in wind-related variables, facilitating a quick understanding of the data characteristics.

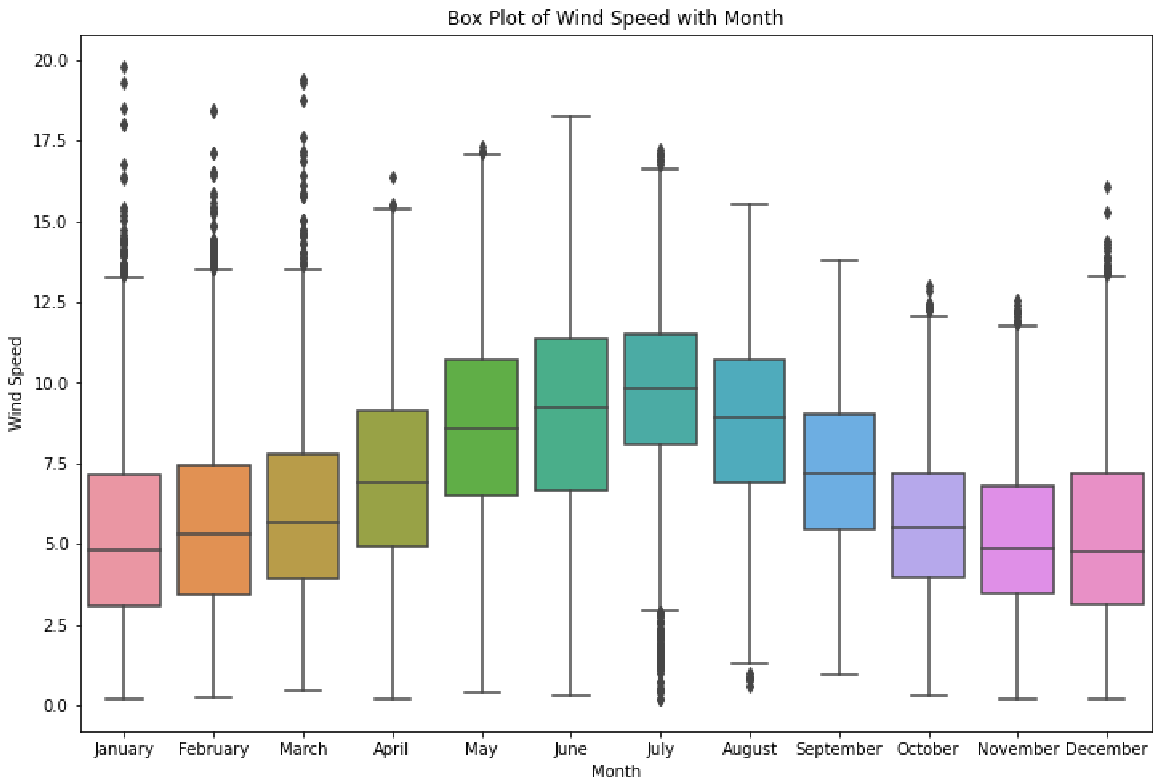


Figure 4.1: Dataset

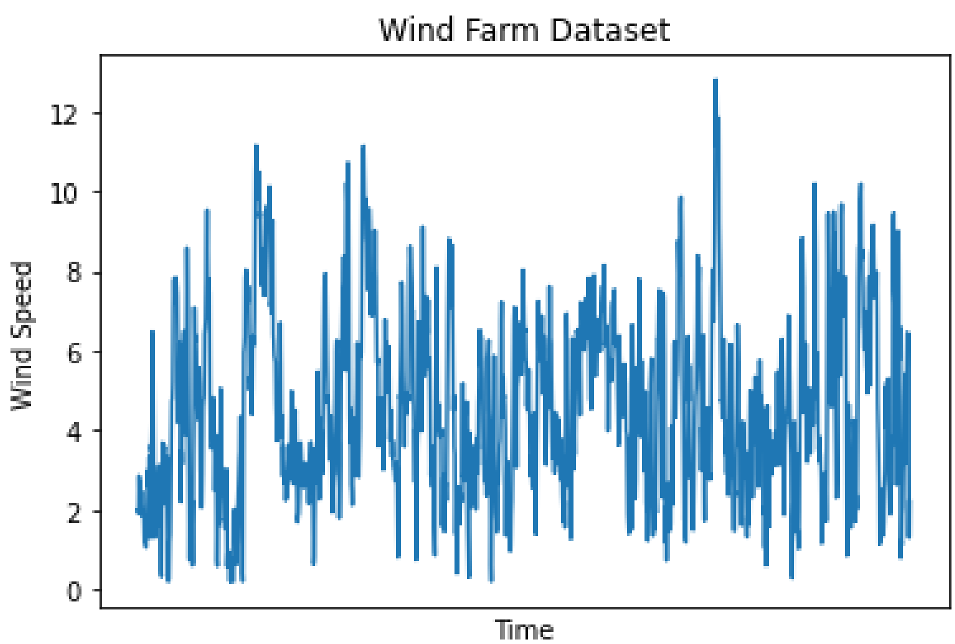


Figure 4.2: Wind Speed

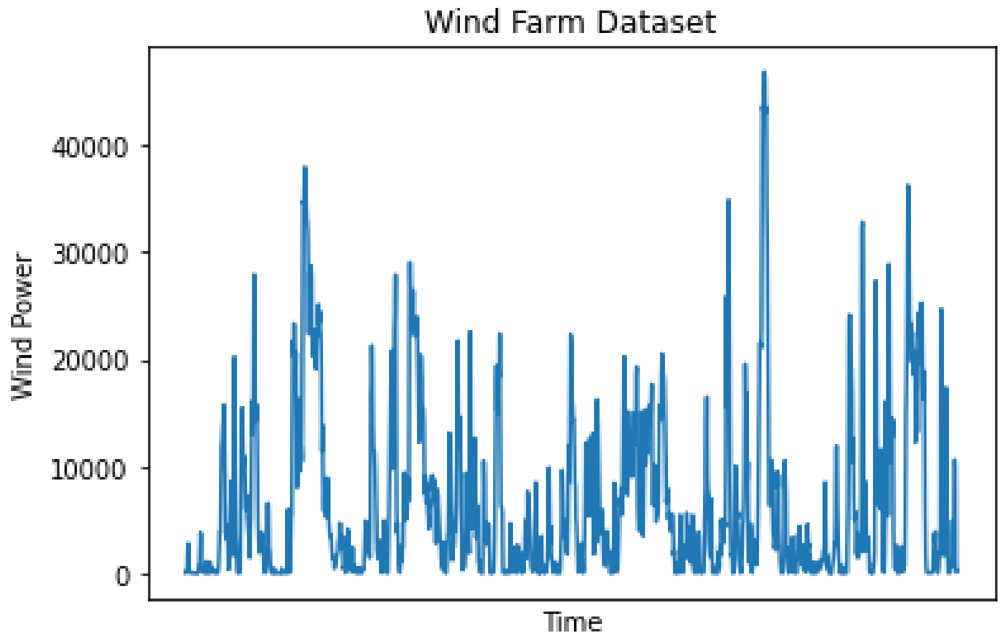


Figure 4.3: Wind power

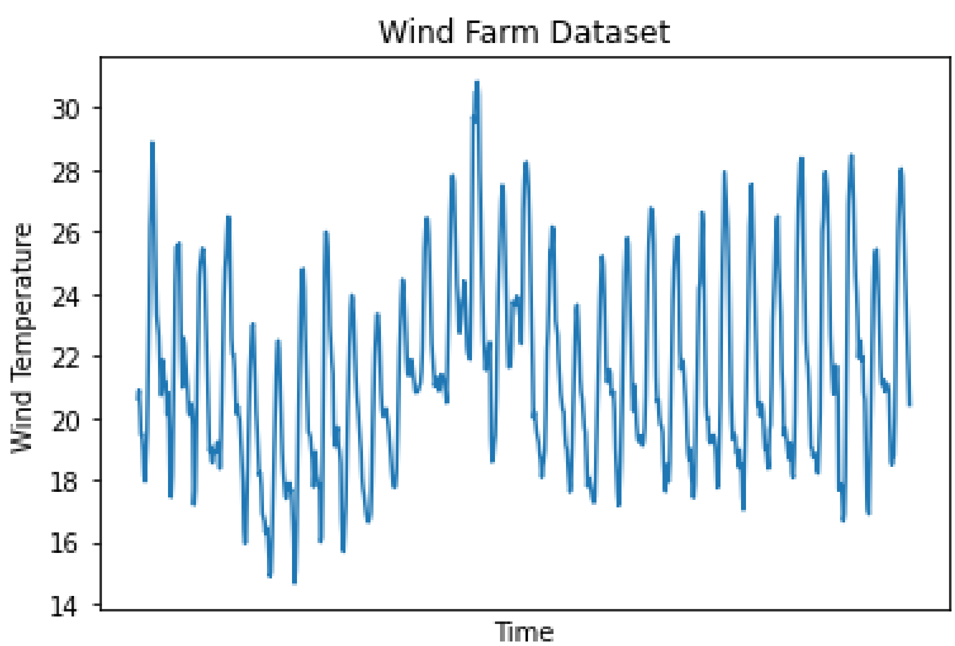


Figure 4.4: Wind Temprature

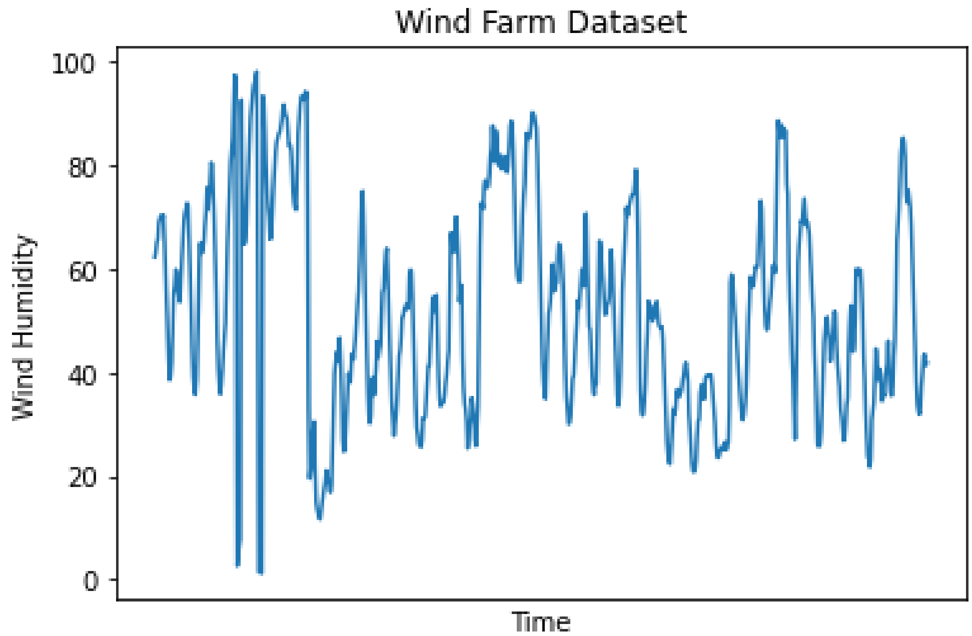


Figure 4.5: Wind Humidity

4.2 Performance Assessment

When assessing the accuracy of the model's predictions, we have used four distinct metrics: MSE, RMSE, MAE, and R2. Those Metrics can be defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.1}$$

Where:

- n is the number of data points.
- y_i is the actual value for data point i.
- $\hat{y}_i$ is the predicted value for data point i.

The first of these criteria is the Mean Squared Error (MSE), which measures the average squared difference between the predicted and actual values. A better fit is indicated by lower MSE values, which provide a better understanding of the overall model fit.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.2}$$

Where:

- n is the number of data points.

- y_i is the actual value for data point i.
- $\hat{y}_i$ is the predicted value for data point i.

Another variant of Mean Squared Error, Root Mean Squared Error (RMSE) is computed by dividing the average squared error by its square root. It gives a numerical number to the mean discrepancy between expected and observed data. An essential statistic for assessing prediction ability, lower RMSE values show improved model accuracy.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4.3}$$

Where:

- n is the number of data points.
- y_i is the actual value for data point i.
- $\hat{y}_i$ is the predicted value for data point i.

The mean absolute deviation (MAE) is a measure of the variance between the actual and expected values. It provides a direct estimate of the precision of the model, is resilient, and is less affected by outliers. Better predictive performance is shown by smaller MAE values.

$$R^2 = 1 - \frac{SS_{\text{RES}}}{SS_{\text{TOT}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{4.4}$$

Where:

- SS_{RES} is the sum of squared regression.
- SS_{TOT} is the sum of squared total.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.
- $\bar{y}_i$ is the mean of actual values.

The coefficient of determination, or R^2, measures how much of the variation in the dependent variable can be explained by the independent variables in the model. A greater correlation and a better fit of the predictive model are indicated by a higher R^2.

4.3 Experiment Results and Analysis

After cleaning the dataset of any outliers using the proposed enhanced DBSCAN technique, we divided it into training and testing sets. We have made sure that our subsequent modelling approach is more accurate and robust by removing these noisy data points. We have automated the process of determining Eps values for DBSCAN using a newly presented method. Using Recursive Feature Elimination, we were able to narrow down the features to ten that were most crucial for improving our prediction models. After that, the XGBoost model—a robust gradient boosting algorithm—is fed these chosen characteristics. Using a grid search method, the model's hyperparameters are fine-tuned to optimise its performance. In order to get the most accurate predictions from the model, grid search determines which parameters work best.

4.4 Enhanced DBSCAN Model Comparison

- **All Features in Enhanced DBSCAN:** The model using DBSCAN produces an MSE of 0.001737 in this comparison, as shown in Table 4.1, but the model without DBSCAN has a little higher MSE of 0.002294. With a lower MSE, the model that is based on DBSCAN shows better prediction accuracy. In the same way, the RMSE of the DBSCAN-incorporated model (0.041677) is lower than that of the non-DBSCAN-incorporated model (0.043602), highlighting the benefit of DBSCAN incorporation. When it comes to MAE, the tendency is the same: the model that incorporates DBSCAN has a lower error rate (0.018828) than the one that does not (0.020667). In addition, the model that incorporates DBSCAN has a higher R2 value (96.9555%) compared to the model without DBSCAN (95.9957%), indicating a more suitable fit to the data.

- **Selected Features (Power, Wind Speed, and Wind Direction) with DBSCAN:** By narrowing our attention to a subset of the features, we can observe from Table 4.1 that the model utilising DBSCAN performs superbly with an MSE of 0.001737, in contrast to the model devoid of DBSCAN, which exhibits a larger MSE of 0.001928. Using this reduced set of features does not affect the DBSCAN approach's superior performance in terms of root-mean-square error (RMSE) (0.041677 vs. 0.043915), mean absolute error (MAE) (0.018828 vs. 0.019462), and root-square percentage (R2) (96.9555 vs. 96.6859). That the DBSCAN-based model stands up well with varying feature sets can be observed.

Table 4.1: Results w.r.t Feature Selection

Metric	WITH DBSCAN		WITHOUT DBSCAN	
	All Features	**Wind Speed, Wind Direction, Wind Power**	**All Features**	**Wind Speed, Wind Direction, Wind Power**
MSE	**0.001737**	0.001901	0.002294	0.001928
RMSE	**0.041677**	0.043602	0.043602	0.043915
MAE	**0.018828**	0.018947	0.020667	0.019462
R^2	**96.9555%**	96.6211%	95.9957%	96.6859%

4.5 Model Comparison with LSTM and SVM

In Table 4.2 and Fgures 4.7, 4.8, we compare the proposed model with DBSCAN against two other models (LSTM and SVM).

- **MSE:** The effectiveness of the proposed model is evident as it surpasses the predictive accuracy of both LSTM (with an accuracy of 0.004645) and SVM (with an accuracy of 0.008374). Notably, the DBSCAN model achieves this superiority by minimizing the Mean Squared Error (MSE) to the lowest recorded value of 0.001737.

- **RMSE:** The RMSE of the proposed model, at 0.041677, is lower than that of the LSTM (at 0.068160) and SVM (at 0.091513) models, demonstrating that the DBSCAN-based model yields more accurate predictions.

- **MAE:** With a minimum MAE of 0.018828 compared to LSTM's 0.038600 and SVM's 0.064810, the proposed model surpasses both LSTM and SVM, having reduced prediction errors.

- **R2:** The R-squared value of 96.9555% is the highest for the proposed model, signifying its ability to account for a larger portion of the variance in the target variable. This shows that theproposed model has better explanatory power in capturing and interpreting the data variability compared to both LSTM (91.8571%) and SVM (85.6094%).

Table 4.2: Comparison of Proposed Model with LSTM & SVM

Metrics	Proposed Model	LSTM	SVM
MSE	**0.001737**	0.004645	0.008374
RMSE	**0.041677**	0.068160	0.091513
MAE	**0.018828**	0.038600	0.064810
R^2	**96.9555%**	91.8571%	85.6094%

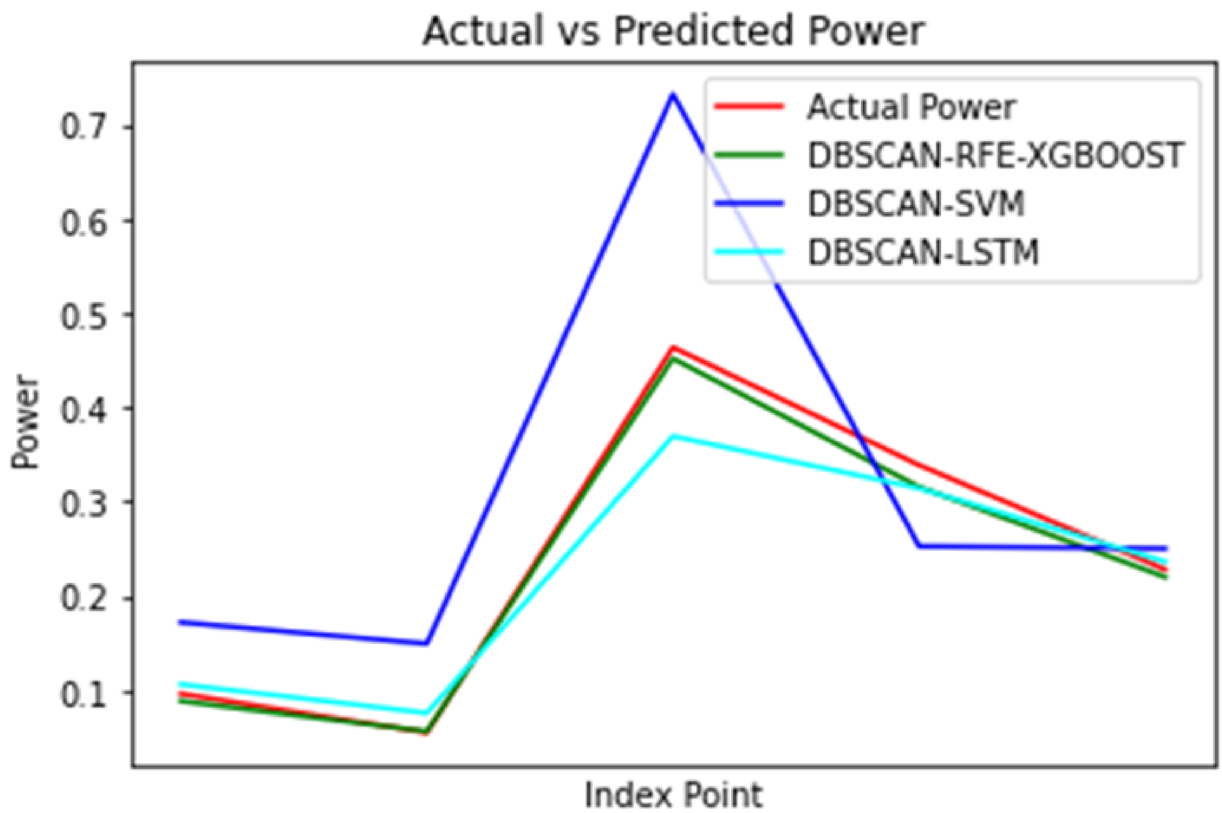


Figure 4.6: Actual Power vs Predicted Power (Comparison of Proposed with benchmark methods)

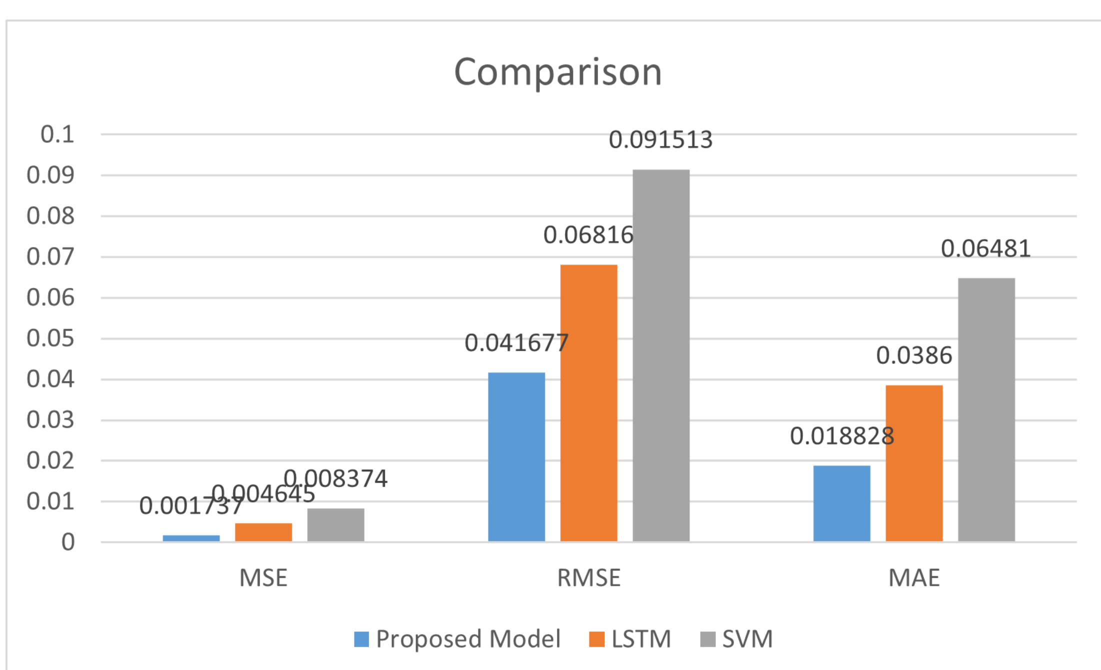


Figure 4.7: MSE,RMSE,MAE Comparison of Proposed with benchmark methods

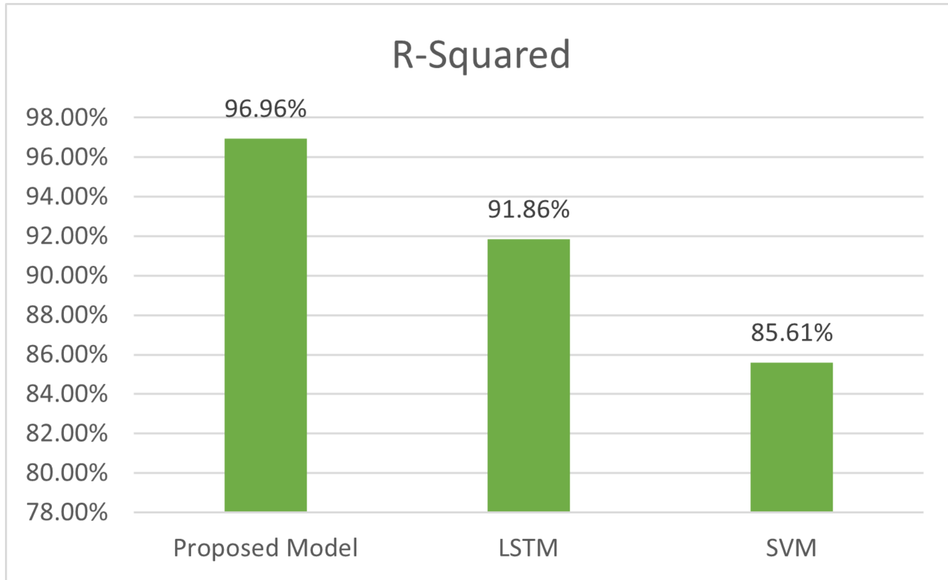


Figure 4.8: R-Sqaured Comparison of Proposed with benchmark methods

Chapter 5

FUTURE WORK AND CONCLUSION

For wind energy to be reliably and efficiently integrated into the power grid, wind power forecasting is essential. Wind energy output is inherently unpredictable and erratic, which makes it difficult for grid operators to maintain a steady power supply. In order to keep consumers' electricity demands met consistently and reliably, operators need accurate wind power forecasts that allow them to anticipate and proactively control these changes. The DBSCAN-RFE-XGBoost model is our recommendation for reliable wind power prediction. We have created a clustering technique called DBSCAN that takes into account the SCADA dataset's inherent uncertainty. This approach uses a K-dist plot and the Knee Point Detection approach to automatically compute the epsilon value needed for DBSCAN clustering. This eliminates the need for tedious, hand-tuning of parameters while yet guaranteeing dataset flexibility. An outlier detection and removal approach is presented for use in clustering data. We have utilized the preprocessed dataset first to extract tempoal features and then applied recursive feature elimination (RFE) to choose most suitable features. When using the XGBoost algorithm to forecast wind power, RFE is useful for determining which features are most relevant to use. We used the SCADA dataset collected from Pakistan's Three Gorges Wind Farm to evaluate our suggested model's performance using real-time data. By reducing Root Mean Squared Error (RMSE) by 38.89% and improving R-squared (R2) by about 5.10%, our model outperformed other benchmark methods, proving the efficacy of our methodology.

Adding advanced feature extraction methods will greatly improve our model's performance in the next stages of our research. This requires diving farther into the input data to find and extract important elements that could lead to better forecasts. Our goal is to discover links and patterns in the data by using advanced algorithms and techniques for feature extraction. Additionally, we may look at other approaches for prediction.

REFERENCES

[1] T.-Z. Ang, M. Salem, M. Kamarol, H. S. Das, M. A. Nazari, and N. Prabaharan, “A comprehensive study of renewable energy sources: Classifications, challenges and suggestions,” *Energy Strategy Reviews*, vol. 43, p. 100939, 2022.

[2] E. Institute, “Statistical review of world energy,” 2023.

[3] Y. Wang, R. Zou, F. Liu, L. Zhang, and Q. Liu, “A review of wind speed and wind power forecasting with deep neural networks,” *Applied Energy*, vol. 304, p. 117766, 2021.

[4] Q. Wu, H. Zheng, X. Guo, and G. Liu, “Promoting wind energy for sustainable development by precise wind speed prediction based on graph neural networks,” *Renewable Energy*, vol. 199, pp. 977–992, 2022.

[5] M. L. Sørensen, P. Nystrup, M. B. Bjerregård, J. K. Møller, P. Bacher, and H. Madsen, “Recent developments in multivariate wind and solar power forecasting,” *Wiley Interdisciplinary Reviews: Energy and Environment*, vol. 12, no. 2, p. e465, 2023.

[6] Q. He, H. Zheng, X. Ma, L. Wang, H. Kong, and Z. Zhu, “Artificial intelligence application in a renewable energy-driven desalination system: A critical review,” *Energy and AI*, vol. 7, p. 100123, 2022.

[7] M. H. Lipu, M. S. Miah, M. Hannan, A. Hussain, M. R. Sarker, A. Ayob, M. H. M. Saad, and M. S. Mahmud, “Artificial intelligence based hybrid forecasting approaches for wind power generation: Progress, challenges and prospects,” *IEEE Access*, vol. 9, pp. 102 460–102 489, 2021.

[8] S. Hanifi, X. Liu, Z. Lin, and S. Lotfian, “A critical review of wind power forecasting methods—past, present and future,” *Energies*, vol. 13, no. 15, p. 3764, 2020.

[9] M. Hutchinson, “Global wind report,” 2023.

[10] NEPRA, "State of industry report," 2023.

[11] J. Zhang, J. Yan, D. Infield, Y. Liu, and F.-s. Lien, "Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and gaussian mixture model," *Applied Energy*, vol. 241, pp. 229–244, 2019.

[12] X. Yuan, C. Chen, M. Jiang, and Y. Yuan, "Prediction interval of wind power using parameter optimized beta distribution based lstm model," *Applied Soft Computing*, vol. 82, p. 105550, 2019.

[13] A. Khosravi and S. Nahavandi, "Combined nonparametric prediction intervals for wind power generation," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 4, pp. 849–856, 2013.

[14] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Lower upper bound estimation method for construction of neural network-based prediction intervals," *IEEE transactions on neural networks*, vol. 22, no. 3, pp. 337–346, 2010.

[15] J. Hu, J. Heng, J. Wen, and W. Zhao, "Deterministic and probabilistic wind speed forecasting with de-noising-reconstruction strategy and quantile regression based algorithm," *Renewable Energy*, vol. 162, pp. 1208–1226, 2020.

[16] Q. Han, F. Meng, T. Hu, and F. Chu, "Non-parametric hybrid models for wind speed forecasting," *Energy Conversion and Management*, vol. 148, pp. 554–568, 2017.

[17] B. Gu, H. Hu, J. Zhao, H. Zhang, and X. Liu, "Short-term wind power forecasting and uncertainty analysis based on fcm–woa–elm–gmm," *Energy Reports*, vol. 9, pp. 807–819, 2023.

[18] A. Zendehboudi, M. A. Baseer, and R. Saidur, "Application of support vector machine models for forecasting solar and wind energy resources: A review," *Journal of cleaner production*, vol. 199, pp. 272–285, 2018.

[19] G. An, Z. Jiang, X. Cao, Y. Liang, Y. Zhao, Z. Li, W. Dong, and H. Sun, "Short-term wind power prediction based on particle swarm optimization-extreme learning machine

model combined with adaboost algorithm," *IEEE access*, vol. 9, pp. 94 040–94 052, 2021.

[20] L. Xiao, W. Shao, F. Jin, and Z. Wu, "A self-adaptive kernel extreme learning machine for short-term wind speed forecasting," *Applied Soft Computing*, vol. 99, p. 106917, 2021.

[21] A. Lahouar and J. B. H. Slama, "Hour-ahead wind power forecast based on random forests," *Renewable energy*, vol. 109, pp. 529–541, 2017.

[22] Y.-X. Wu, Q.-B. Wu, and J.-Q. Zhu, "Data-driven wind speed forecasting using deep feature extraction and lstm," *IET Renewable Power Generation*, vol. 13, no. 12, pp. 2062–2069, 2019.

[23] P. Jiang, Z. Liu, X. Niu, and L. Zhang, "A combined forecasting system based on statistical method, artificial neural networks, and deep learning methods for short-term wind speed forecasting," *Energy*, vol. 217, p. 119361, 2021.

[24] Z. Qu, W. Mao, K. Zhang, W. Zhang, and Z. Li, "Multi-step wind speed forecasting based on a hybrid decomposition technique and an improved back-propagation neural network," *Renewable energy*, vol. 133, pp. 919–929, 2019.

[25] Q. Zhou, Q. Lv, and G. Zhang, "A combined forecasting system based on modified multi-objective optimization for short-term wind speed and wind power forecasting," *Applied Sciences*, vol. 11, no. 20, p. 9383, 2021.

[26] Z. Qu, W. Mao, K. Zhang, W. Zhang, and Z. Li, "Multi-step wind speed forecasting based on a hybrid decomposition technique and an improved back-propagation neural network," *Renewable energy*, vol. 133, pp. 919–929, 2019.

[27] X. He, Y. Nie, H. Guo, and J. Wang, "Research on a novel combination system on the basis of deep learning and swarm intelligence optimization algorithm for wind speed forecasting," *IEEE Access*, vol. 8, pp. 51 482–51 499, 2020.

[28] J. Wang, W. Yang, P. Du, and T. Niu, "A novel hybrid forecasting system of wind speed based on a newly developed multi-objective sine cosine algorithm," *Energy Conversion and Management*, vol. 163, pp. 134–150, 2018.

[29] Z. He, Y. Chen, Z. Shang, C. Li, L. Li, and M. Xu, "A novel wind speed forecasting model based on moving window and multi-objective particle swarm optimization algorithm," *Applied Mathematical Modelling*, vol. 76, pp. 717–740, 2019.

[30] Q. Wu, H. Zheng, X. Guo, and G. Liu, "Promoting wind energy for sustainable development by precise wind speed prediction based on graph neural networks," *Renewable Energy*, vol. 199, pp. 977–992, 2022.

[31] J. Wang, H. Zhu, Y. Zhang, F. Cheng, and C. Zhou, "A novel prediction model for wind power based on improved long short-term memory neural network," *Energy*, vol. 265, p. 126283, 2023.

[32] L. P. Joseph, R. C. Deo, R. Prasad, S. Salcedo-Sanz, N. Raj, and J. Soar, "Near real-time wind speed forecast model with bidirectional lstm networks," *Renewable Energy*, vol. 204, pp. 39–58, 2023.

[33] Z. Jiang, J. Che, M. He, and F. Yuan, "A cgru multi-step wind speed forecasting model based on multi-label specific xgboost feature selection and secondary decomposition," *Renewable Energy*, vol. 203, pp. 802–827, 2023.

[34] X. Peng, Y. Li, L. Dong, K. Cheng, H. Wang, Q. Xu, B. Wang, C. Liu, J. Che, F. Yang *et al.*, "Short-term wind power prediction based on wavelet feature arrangement and convolutional neural networks deep learning," *IEEE Transactions on Industry Applications*, vol. 57, no. 6, pp. 6375–6384, 2021.

[35] H. Jiajun, Y. Chuanjin, L. Yongle, and X. Huoyue, "Ultra-short term wind prediction with wavelet transform, deep belief network and ensemble learning," *Energy Conversion and Management*, vol. 205, p. 112418, 2020.

[36] F. Liu, Q. Tao, D. Yang, and D. Sidorov, "Bidirectional gated recurrent unit-based lower upper bound estimation method for wind power interval prediction," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 3, pp. 461–469, 2021.

[37] Y. Yin, Z. Shang, F. Li, and H. Shao, "Short-term wind speed prediction based on improved auto encoder," in *Journal of Physics: Conference Series*, vol. 2418, no. 1. IOP Publishing, 2023, p. 012098.

[38] Y. Zhang, H. Sun, and Y. Guo, "Wind power prediction based on pso-svr and grey combination model," *IEEE Access*, vol. 7, pp. 136 254–136 267, 2019.

[39] Z. Wang, L. Wang, C. Huang, and X. Luo, "A hybrid ensemble learning model for short-term solar irradiance forecasting using historical observations and sky images," *IEEE Transactions on Industry Applications*, vol. 59, no. 2, pp. 2041–2049, 2022.

[40] Z. Tang, G. Zhao, G. Wang, and T. Ouyang, "Hybrid ensemble framework for short-term wind speed forecasting," *IEEE Access*, vol. 8, pp. 45 271–45 291, 2020.

[41] D. Kaur, T. T. Lie, N. K. Nair, and B. Vallès, "Wind speed forecasting using hybrid wavelet transform-arma techniques," *Aims Energy*, vol. 3, no. 1, pp. 13–24, 2015.

[42] X. Liu, L. Yang, and Z. Zhang, "Short-term multi-step ahead wind power predictions based on a novel deep convolutional recurrent network method," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 3, pp. 1820–1833, 2021.

[43] B. Xiong, L. Lou, X. Meng, X. Wang, H. Ma, and Z. Wang, "Short-term wind power forecasting based on attention mechanism and deep learning," *Electric Power Systems Research*, vol. 206, p. 107776, 2022.

[44] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023.

[45] R. Lu, W. Zhang, Y. Wang, Q. Li, X. Zhong, H. Yang, and D. Wang, "Auction-based cluster federated learning in mobile edge computing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 4, pp. 1145–1158, 2023.

[46] J. Xie, W. Kong, S. Xia, G. Wang, and X. Gao, "An efficient spectral clustering algorithm based on granular-ball," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[47] F. Ascolani, A. Lijoi, G. Rebaudo, and G. Zanella, "Clustering consistency with dirichlet process mixtures," *Biometrika*, vol. 110, no. 2, pp. 551–558, 2023.

[48] M. Han, H. Wu, Z. Chen, M. Li, and X. Zhang, "A survey of multi-label classification based on supervised and semi-supervised learning," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 3, pp. 697–724, 2023.

[49] Q.-F. Yang, W.-Y. Gao, G. Han, Z.-Y. Li, M. Tian, S.-H. Zhu, and Y.-h. Deng, "Hcdc: A novel hierarchical clustering algorithm based on density-distance cores for data sets with varying density," *Information Systems*, vol. 114, p. 102159, 2023.

[50] C. Paik, Y. Chung, and Y. J. Kim, "Power curve modeling of wind turbines through clustering-based outlier elimination," *Applied System Innovation*, vol. 6, no. 2, p. 41, 2023.

[51] M. Zulfiqar, M. Kamran, M. Rasheed, T. Alquthami, and A. Milyani, "Hyperparameter optimization of support vector machine using adaptive differential evolution for electricity load forecasting," *Energy Reports*, vol. 8, pp. 13 333–13 352, 2022.

[52] Z. Wang, L. Wang, M. Revanesh, C. Huang, and X. Luo, "Short-term wind speed and power forecasting for smart city power grid with a hybrid machine learning framework," *IEEE Internet of Things Journal*, 2023.

[53] S. H. Rafi, S. R. Deeba, E. Hossain *et al.*, "A short-term load forecasting method using integrated cnn and lstm network," *IEEE Access*, vol. 9, pp. 32 436–32 448, 2021.

[54] F. Zhou, Z. Huang, and C. Zhang, "Carbon price forecasting based on ceemdan and lstm," *Applied Energy*, vol. 311, p. 118601, 2022.

[55] C.-H. Liu, J.-C. Gu, and M.-T. Yang, "A simplified lstm neural networks for one dayahead solar power forecasting," *Ieee Access*, vol. 9, pp. 17 174–17 195, 2021.

[56] M.-S. Ko, K. Lee, J.-K. Kim, C. W. Hong, Z. Y. Dong, and K. Hur, "Deep concatenated residual network with bidirectional lstm for one-hour-ahead wind power forecasting," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 2, pp. 1321–1335, 2020.

[57] M. Liu, Z. Cao, J. Zhang, L. Wang, C. Huang, and X. Luo, "Short-term wind speed forecasting based on the jaya-svm model," *International Journal of Electrical Power & Energy Systems*, vol. 121, p. 106056, 2020.

[58] P. Jiang, R. Li, N. Liu, and Y. Gao, "A novel composite electricity demand forecasting framework by data processing and optimized support vector machine," *Applied Energy*, vol. 260, p. 114243, 2020.

[59] C. Xiao, W. Xia, and J. Jiang, "Stock price forecast based on combined model of ari-ma-ls-svm," *Neural Computing and Applications*, vol. 32, pp. 5379–5388, 2020.

[60] G. N. Kouziokas, "A new w-svm kernel combining pso-neural network transformed vector and bayesian optimized svm in gdp forecasting," *Engineering Applications of Artificial Intelligence*, vol. 92, p. 103650, 2020.