UNIVERSITI TEKNOLOGI MARA

COMPARISON BETWEEN SEVERAL MACHINE LEARNING ALGORITHMS TO EVALUATE SOCIOECONOMIC INFLUENCES ON HYDROELECTRIC GENERATION IN MALAYSIA

'ALIAA AQILAH BINTI NOR AZMAN

MSc

May 2026

UNIVERSITI TEKNOLOGI MARA

COMPARISON BETWEEN SEVERAL MACHINE LEARNING ALGORITHMS TO EVALUATE SOCIOECONOMIC INFLUENCES ON HYDROELECTRIC GENERATION IN MALAYSIA

'ALIAA AQILAH BINTI NOR AZMAN

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Electrical Engineering)

Faculty of Electrical Engineering

May 2026

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 9th January 2026 to conduct the final examination of 'Aliaa Aqilah Binti Nor Azman on her Masters of Science thesis entitled "Comparison Between Several Machine Learning Algorithms to Evaluate Socioeconomic Influences on Hydroelectric Generation in Malaysia" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiner recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Abdul Hadi Abdul Razak, PhD
Associate Professor
Faculty of Electrical Engineering
Universiti Teknologi MARA
(Chairman)

Ahmad Ihsan Mohd Yassin
Associate Professor
Faculty of Electrical Engineering
Universiti Teknologi MARA
(Internal Examiner)

Norizam Sulaiman
Senior Lecturer
Universiti Malaysia Pahang Al-Sultan Abdullah
(External Examiner)

PROF DR HJH ZURAEDA IBRAHIM
Dean
Institute of Postgraduates Studies
Universiti Teknologi MARA

Date: 24 May 2026

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	'Aliaa Aqilah Binti Nor Azman
Student ID. No.	:	2023211408
Programme	:	Master of Science (Electrical Engineering) – CEE750
Faculty	:	Electrical Engineering
Thesis Title	:	Comparison Between Several Machine Learning Algorithms to Evaluate Socioeconomic Influences on Hydroelectric Generation in Malaysia

Signature of Student	:	……..
Date	:	May 2026

ABSTRACT

Accurate forecasting of hydroelectric power generation is important for supporting long-term energy planning and improving the management of renewable energy resources in Malaysia. While many previous hydropower forecasting studies emphasise hydrological variables such as rainfall and reservoir inflow, the potential influence of socioeconomic indicators on hydroelectric generation has received comparatively less attention. However, hydroelectric forecasting remains challenging due to complex and nonlinear relationships between influencing factors, which are often not fully captured by traditional methods. In addition, limited studies have compared machine learning models for hydroelectric forecasting in the Malaysian context, particularly when incorporating socioeconomic indicators. This study aims to analyse the relationship between selected socioeconomic indicators and hydroelectric power generation in Malaysia and to evaluate the performance of several machine learning models for forecasting hydropower output. Annual data covering the period from 1980 to 2021 were collected for Gross Domestic Product (GDP), energy consumption, population, inflation rate, and hydroelectric generation. Three Artificial Neural Network (ANN) configurations were first developed using different input combinations: Model A (GDP and energy consumption), Model B (GDP, energy consumption, and population), and Model C (GDP, energy consumption, and inflation). To assess the reliability of the input configurations, a five-fold cross-validation procedure was applied. The results showed that Model B produced the most favourable cross-validation performance, achieving an average Mean Squared Error (MSE) of 1.8955×10^5 with an R-value of 0.7731. Based on this result, the selected input combination was further evaluated using four machine learning models, namely Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest, and XGBoost. Among the evaluated models, ANN demonstrated the most consistent predictive performance, achieving a testing MSE of 1.1541×10^4 and an R-value of 0.9962. These results suggest that socioeconomic indicators such as GDP, energy consumption, and population may provide useful explanatory signals for hydroelectric generation forecasting. However, the findings should be interpreted cautiously due to the relatively limited dataset used in this study. Despite this limitation, the study demonstrates the potential of machine learning approaches for supporting data-driven energy forecasting and planning.

ACKNOWLEDGEMENT

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Ts. Dr. Mohd Azri Abdul Aziz, Ts. Dr. Noorfadzli Abdul Razak, and Assoc. Prof. Ts. Mahanijah Md Kamal, for their invaluable guidance, patience, and encouragement throughout my Master's journey. Their willingness to share knowledge, provide constructive feedback, and offer constant support has been instrumental in shaping the direction and quality of this research. Each of them has contributed not only to my academic progress but also to my personal growth as a learner and researcher.

To my beloved family, I am forever thankful for your unconditional love, prayers, and understanding. You have been my constant source of strength and motivation, and your belief in me, even during the moments when I doubted myself, has kept me moving forward. To my parents, thank you for teaching me the value of hard work and perseverance. To my siblings, thank you for your encouragement, humour, and for being there whenever I needed a reminder that I am never alone in this journey.

To my friends, I am deeply grateful for your support in both big and small ways, whether it was lending an ear when I needed to talk, sharing words of encouragement when I felt overwhelmed, or simply spending time together to give me moments of relief from the pressure of deadlines. The shared laughter, kindness, and understanding have been a much-needed balance to the challenges of this academic path.

Lastly, I would like to thank myself. For showing up on the difficult days, for continuing to push forward despite setbacks, and for choosing not to give up when the road felt uncertain. This achievement is a result not only of the guidance and love I have received from others but also of the quiet determination, patience, and resilience that I have nurtured within myself. This journey has been as much about completing a thesis as it has been about discovering my own strength.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbols

Gwh	Gigawatt-Hour, Unit of Energy
Mwh	Megawatt-Hour, Unit of Energy
μ	Mean of Data Set
R	Correlation Coefficient
σ	Standard Deviation of Data
x_{norm}	Set Normalized Value or x
x_{std}	Standardization Value of x

LIST OF ABBREVIATIONS

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMAX	Autoregressive Integrated Moving Average with Exogenous Variable
GDP	Growth Domestic Product
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MSE	Mean Squared Error
PSO	Particle Swarm Optimisation
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SVM	Support Vector Machine
SVR	Support Vector Regression
XGBOOST	Extreme Gradient Boost

CHAPTER 1
INTRODUCTION

1.1 Research Background

A global transition toward sustainable energy sources is currently reshaping national policies and economic frameworks. This shift is motivated by environmental imperatives such as climate change mitigation and the pursuit of long-term energy security and economic resilience. The energy sector, as highlighted by Malaysia's Economic Planning Unit, is a central pillar in this transition, contributing approximately 28% to GDP and employing 25% of the national workforce, while supporting economic development strategies (Hashemizadeh et al., 2024; Ministry of Economy (MOE), 2023).

Renewable energy has emerged as a cornerstone in global energy planning, providing an environmentally sustainable alternative to fossil fuels. As of 2020, renewable sources represented about 37% of global installed power capacity, with Germany reaching around 46% renewable electricity generation (Sustainable Energy Development Authority (SEDA), 2021). Malaysia, however, reported a comparatively lower renewable energy share of 23% in the same year. To address this gap, the Malaysia Renewable Energy Roadmap (MyRER) has set a national target of achieving 31% of installed capacity from renewable energy sources by 2025 and 40% by 2035, placing emphasis on operational efficiency, industrial growth, and environmental sustainability (SEDA, 2021). Furthermore, Malaysia has introduced mechanisms such as the Large-Scale Solar (LSS) program and Feed-in Tariff (FiT) initiatives to complement hydropower in reaching these targets, indicating a broader policy shift toward diversified renewable portfolios (MOE, 2023).

Among the available renewable sources, hydroelectric power remains one of the most reliable and widely adopted forms of energy, particularly in countries endowed with rich water resources (Benti et al., 2023). It offers stable and dispatchable electricity output, supports grid reliability, and aligns with green energy objectives. In Malaysia, hydropower currently accounts for around 11% of total electricity generation, with future plans aiming to increase its contribution to 35% by 2030 (Hazmin et al., 2024).

Large-scale developments such as the Bakun and Murum dams under the Sarawak Corridor of Renewable Energy (SCORE) exemplify how hydropower can support both industrial and regional development goals (Hazmin et al., 2024; SEDA, 2021). In addition to these mega-projects, smaller run-of-river and mini-hydro schemes have also been identified as potential contributors to the energy mix, particularly for rural electrification and off-grid solutions in remote areas. The Malaysia Renewable Energy Roadmap (MyRER) projects hydroelectric capacity growth as part of the strategy to reach the 40% renewable energy share target by 2035 (SEDA, 2021).

Despite its promise, hydroelectric power generation is influenced by multiple external factors such as climate variability and socioeconomic dynamics. Variables like GDP, population, energy consumption, and inflation can directly or indirectly impact generation patterns (Ren et al., 2023). For example, higher industrial activity driven by economic growth can increase electricity demand, while demographic changes and energy consumption patterns can influence planning and infrastructure needs. Fluctuations in rainfall, extreme weather events, and seasonal water availability, often linked to climate change, add another layer of complexity to predicting hydroelectric output. In this context, accurate forecasting of hydroelectric output becomes essential for informed energy management, especially as Malaysia seeks to balance energy diversification, rural electrification, and environmental considerations (Hazmin et al., 2024; Muhammad et al., 2024).

Artificial intelligence (AI) and machine learning (ML) techniques have been increasingly applied to better understand and forecast hydroelectric generation. These models offer advantages over traditional statistical methods in handling complex, nonlinear relationships among influencing variables. In particular, data-driven approaches such as ANN, SVM, Random Forests, and XGBoost have shown promising results in energy-related studies (Rajarajeswaran, 2025; Sauhats et al., 2016). Artificial Neural Networks (ANN) are commonly categorised under machine learning approaches because they learn patterns from data through training processes, although deeper architectures with multiple hidden layers are sometimes referred to as deep learning models. In this study, ANN is treated as a machine learning model due to its relatively simple network structure and its widespread application in energy forecasting studies. Their ability to learn patterns from historical and socioeconomic data makes them suitable for energy demand and supply forecasting, as demonstrated by recent studies in China, India, and Southeast Asia. Recent work suggests hybrid and deep learning-

based approaches can significantly enhance prediction accuracy. For instance, the Gaussian Process Energy Efficient Maintainable Environment (GPEEME) model demonstrated over 91% improvement in energy consumption forecasting (Sattar Hanoon et al., 2023; Shamout et al., 2022; Villeneuve et al., 2023). However, for the scope of this study, the standard ANN is selected as the primary framework. This choice ensures a balance between model stability and the ability to effectively capture non-linear trends within the socioeconomic indicators, while maintaining a level of interpretability necessary for energy planning. These technological advancements offer new opportunities to support Malaysia's energy planning, especially in optimising the integration of socioeconomic indicators into hydroelectric forecasting frameworks.

Several previous studies have investigated hydroelectric forecasting using different modelling approaches. For example, time-series forecasting techniques have been applied to hydroelectric generation systems to improve prediction performance under varying operating conditions (Barzola-Monteses et al., 2025). Many of these studies focus primarily on hydrological inputs such as reservoir level, inflow, or other water-related variables when predicting hydroelectric generation (Sapitang et al., 2020). In addition, machine learning techniques such as Artificial Neural Networks have also been explored in hydropower forecasting due to their ability to capture nonlinear relationships among influencing variables (Sattar Hanoon et al., 2023; Villeneuve et al., 2023). Although these approaches have shown encouraging results, existing studies have placed greater emphasis on environmental and hydrological factors, with a notable gap in the literature regarding socioeconomic indicators, which remain under-explored despite their potential influence on long-term hydroelectric generation patterns.

However, existing studies on hydroelectric forecasting in Malaysia have emphasised mainly hydrological or weather-related inputs, with less focus on socioeconomic indicators such as GDP, energy consumption, and inflation. This creates a knowledge gap in understanding how economic growth and demographic factors influence long-term hydropower generation. Addressing this gap by applying advanced ML models can provide actionable insights for policymakers and energy planners, particularly in aligning national energy strategies with sustainable development goals.

1.2 Problem Statement

The increasing demand for reliable and sustainable energy has made accurate forecasting of hydroelectric generation a critical component of energy planning. In Malaysia, hydropower accounts for a growing share of the renewable energy mix, with targets to increase its contribution to 35% of total electricity generation by 2030. However, the generation capacity of hydropower is highly dependent on external factors, including environmental and hydrological conditions such as rainfall, reservoir inflow, climate variability, and water availability (Lee et al., 2022; Lamidi et al., 2024), as well as socioeconomic dynamics such as GDP growth, population changes, and energy consumption patterns. These interdependencies make forecasting complex and challenging. Although useful for trend analysis, traditional statistical methods often fail to capture the nonlinear and dynamic relationships between these variables, resulting in limited forecasting accuracy (Barzola-Monteses et al., 2025).

Recent advancements in AI and ML have shown promising results in addressing the complexity of energy forecasting. Models such as ANN, SVM, Random Forest, and XGBoost can effectively handle nonlinearities and identify historical and socioeconomic data patterns that traditional models overlook. Machine learning techniques have increasingly been applied in hydropower and renewable energy forecasting because of their capability to model complex nonlinear relationships within large datasets (Sattar Hanoon et al., 2023; Villeneuve et al., 2023). However, despite the growing adoption of these techniques, there is limited research on the comparative performance of these ML models in the context of hydroelectric power generation in Malaysia. Most existing studies focus on other energy sources, such as solar or wind, leaving a gap in the literature on hydroelectric forecasting within the Malaysian context (Sapitang et al., 2020).

Furthermore, while socioeconomic factors such as GDP, population, and energy consumption are widely acknowledged as influential in shaping electricity demand, their integration into hydroelectric forecasting models remains underexplored. Although environmental variables are commonly used in hydropower forecasting because they directly influence water availability and generation capacity (Lee et al., 2022; Lamidi et al., 2024), this study focuses on socioeconomic indicators to examine their potential relationship with hydroelectric generation. Without a robust and accurate

forecasting approach that accounts for these factors, there is a risk of suboptimal planning and resource allocation, potentially hindering Malaysia's progress toward its renewable energy targets. This research seeks to address these gaps by evaluating and comparing the performance of ANN, SVM, Random Forest, and XGBoost models in predicting hydroelectric generation based on key socioeconomic indicators. While environmental factors like rainfall are important, it is often difficult to find consistent and detailed water data for all of Malaysia over a long period. In contrast, socioeconomic data like GDP and population are more reliable and easier to get from national records, making them better suited for the long-term forecasting goals of this study.

1.3 Research Objectives

The main objective of this study is to develop and evaluate machine learning models for forecasting hydroelectric generation in Malaysia using key socioeconomic indicators. The following specific objectives support this overall aim:

i) To analyse the influence of socioeconomic factors, including GDP, energy consumption, population, and inflation, on hydroelectric power generation
ii) To develop, validate and evaluate the performance of selected machine learning models (ANN, SVM, Random Forest, and XGBoost) in forecasting hydroelectric power generation and identify the most suitable model.

1.4 Research Questions

To achieve the objectives, this study addresses the following research questions:

i) How effectively can machine learning models (ANN, SVM, Random Forest, and XGBoost) forecast hydroelectric generation when trained on socioeconomic indicators such as GDP, energy consumption, and population?
ii) Which model demonstrates the highest predictive performance based on evaluation metrics such as MSE and R-values?

iii) What insights do the error distribution and regression plot analyses offer regarding model reliability and forecasting stability?

1.5 Significance of Study

This study contributes to understanding how socioeconomic indicators influence hydroelectric power generation and demonstrates the application of ML techniques in forecasting energy output. By comparing multiple models, including ANN, SVM, Random Forest, and XGBoost, the study highlights these approaches' relative strengths and limitations when applied to structured historical and socioeconomic datasets. This comparative framework provides insights into the suitability of data-driven models for energy forecasting tasks where accuracy, reliability, and adaptability are essential.

The findings are relevant for energy planners and policymakers as they provide an evidence-based perspective on the factors that significantly impact hydroelectric generation in Malaysia. This research supports better planning for infrastructure development and energy demand management by identifying which socioeconomic variables contribute most to predictive performance. Moreover, evaluating model performance using established metrics such as MSE and R-values ensures that the conclusions are grounded in objective measures rather than assumptions.

On a broader level, the study demonstrates the role of artificial intelligence and machine learning in addressing energy forecasting challenges. Although focused on Malaysia, the methodology and comparative approach can be adapted to other regions with similar energy profiles. This flexibility makes the research valuable for future studies seeking to optimise renewable energy forecasting while integrating economic and demographic factors.

1.6 Scope and Limitation of Work

This study focuses on forecasting hydroelectric power generation in Malaysia using socioeconomic indicators, specifically GDP, Energy Consumption, and Population, as input variables. These variables were selected because they have consistent long-term data records and are commonly associated with electricity demand and energy consumption patterns. The analysis is based on annual data from 1980 to

2021, obtained from verified national and international databases to ensure data quality and reliability. Four machine learning models are developed and evaluated: ANN, SVM, Random Forest, and XGBoost. These models are selected due to their frequent use in energy forecasting studies and their capability to model nonlinear relationships in datasets**.** A consistent data partitioning strategy is applied, with the dataset divided into training, validation, and testing subsets. Model performance is assessed using MSE and R-values, supported by additional evaluation tools such as error histograms, learning curves, and regression plots.

The models are implemented using MATLAB software. All simulations are conducted on a personal computer running the Windows operating system with an Intel® Core(TM) i7-8550U CPU @ 1.80GHz and 8.00 GB memory. These specifications describe the computational environment used during the model development and evaluation process.

The scope of this work is limited by the availability and resolution of historical data. Only socioeconomic variables with consistent long-term records are considered, while other potential predictors, such as climatic and hydrological factors, are excluded. As a result, the models emphasise the socioeconomic influence on hydroelectric generation but may not fully capture environmental variations, including rainfall fluctuations or extreme weather events. Despite these limitations, the methodology has been designed to provide a fair and practical comparison of model performance within the available data constraints. The outcomes will offer valuable insights for energy forecasting and policy planning.

1.7 Thesis Organisation

This thesis is organised into five chapters that collectively guide the research from its conceptual foundation to its conclusions and recommendations. Chapter 1 introduces the study by outlining the research background, problem statement, objectives, research questions, significance, scope, and limitations. It establishes the motivation behind the research and provides a clear framework for the work undertaken. Chapter 2 presents a comprehensive review of the relevant literature, focusing on hydroelectric generation forecasting, renewable energy policies, and the application of machine learning techniques in energy prediction. This chapter highlights key findings

from prior studies and identifies research gaps that this work aims to address. Chapter 3 explains the methodology adopted in the study, including data collection, preprocessing, and model development. It describes the machine learning models applied, such as ANN, SVM, Random Forest, and XGBoost, along with the evaluation metrics used to assess performance. Chapter 4 details the results and provides an in-depth discussion of the findings, including the analysis of performance metrics, error distributions, and regression plots, while also comparing the predictive capabilities of different models. Finally, Chapter 5 concludes the study by summarising the key results, emphasising the contributions of the research, and offering recommendations for future work and potential improvements.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Hydroelectric power continues to serve as a key component of the global renewable energy landscape due to its high efficiency, reliability, and contribution to energy security. As the demand for sustainable energy sources grows, forecasting hydroelectric generation has become essential for policymakers, utility planners, and researchers. However, various environmental, technical, and socioeconomic factors often complicate the forecasting process. These include variable weather conditions, limited data availability, and the dynamic nature of energy demand. As a result, there is a continued need for improved modelling approaches that can accommodate the complexity of hydropower systems.

This literature review provides an overview of the current research efforts to address these challenges. It first discusses the importance of hydroelectric forecasting and outlines key environmental and operational barriers. Following this, the review explores the impact of seasonal changes and climate-related uncertainties on hydropower generation. In shaping energy demand patterns, attention is given to socioeconomic indicators such as gross domestic product, population, and inflation. The review also highlights the growing use of machine learning methods in energy forecasting, including models such as ANN, SVM, Random Forest, and XGBoost. Finally, it examines the use of neural network-based models in more detail, focusing on performance limitations and overfitting issues

2.2 Hydroelectric Generation Forecasting: Importance and Challenges

Hydroelectric power remains a cornerstone of the global renewable energy portfolio, offering high reliability, efficiency, and flexibility in supporting national energy security goals. For instance, in countries like Indonesia, hydropower accounted for 57% of renewable electricity generation by the end of 2021, underscoring its dominance over intermittent sources like wind and solar (Hasanah et al., 2023). Its

operational efficiency, often exceeding 90%, and ability to respond rapidly to demand shifts make it a valuable asset in modern power systems (Alshammari, 2022). Moreover, hydroelectric generation supports grid stability by compensating for fluctuations in variable renewable sources and is regarded as a cleaner alternative to fossil-fuel-based power generation (Alshammari, 2022). Globally, hydropower contributes approximately 17% of total electricity production and remains the single largest renewable source, surpassing the combined output of solar, wind, and bioenergy (Gunasinghe et al., 2023). This importance is amplified in regions like Kyrgyzstan, where hydropower provides nearly 90% of total electricity generation, reflecting its critical role in national energy security (Gunasinghe et al., 2023).

Despite its advantages, various environmental, technical, and managerial factors challenge the accurate forecasting of hydroelectric generation. Unpredictable weather patterns, variable inflow rates, and operational constraints introduce significant uncertainty into forecast models (Ersan & Irmak, 2023). For instance, silt accumulation during monsoon seasons in the Indian Himalayan region has led to hydro plant shutdowns to avoid turbine damage, demonstrating how environmental factors can directly disrupt generation capacity (Kumar et al., 2024). Socioeconomic and regulatory influences, such as competing water uses for agriculture or drinking supply, further complicate the forecasting process (Alshammari, 2022; S.D. Grande et al., 2024). Reports from China and Ecuador show that poor hydropower forecasts during severe droughts have resulted in grid instability and rolling blackouts, underlining the risks of inadequate prediction models (Gunasinghe et al., 2023). Reliable forecasting is therefore essential not only for optimising energy output but also for infrastructure longevity, investment planning, and grid integration (Apaydin et al., 2020), (Rajarajeswaran, 2025). Emerging studies also highlight that improved forecasting approaches and data-driven methods that account for grid balancing and ramping requirements, are increasingly necessary to maintain power system reliability under volatile weather conditions (Bespalyy & Bespalaya, 2025).

2.2.1 Seasonal Variability

One of the primary challenges in hydroelectric forecasting is the significant seasonal variability in water availability. Hydropower output is heavily influenced by precipitation patterns, snowmelt, and temperature, which vary considerably across

seasons (Rasku et al., 2020; Matrenin et al., 2022). Ren et al. (2023) highlight the inherent inconsistency in hydropower generation throughout the year, marked by high-output wet seasons and low-output dry seasons. For example, in Turkey, winter precipitation contributes to peak generation, while drier summers result in power output reductions of up to 30% (Jāmiʿat al-Imārāt al-ʿArabīyah al-Muttaḥidah & Institute of Electrical and Electronics Engineers, n.d.; 2018 International Conference on Power System Technology [POWERCON], 2018). This seasonality is particularly pronounced in snow-fed river basins, where delayed snowmelt often causes mismatches between peak energy demand and water availability (Gunasinghe et al., 2023). In regions like Nepal, run-of-the-river hydropower plants experience extreme seasonal disparities, with monsoon periods contributing the majority of annual generation while winter months produce minimal energy output (Dahal et al., 2024). Similar patterns are evident in Kazakhstan, where small hydropower stations show a capacity factor of 42% but experience sharp seasonal generation drops, requiring improved forecasting for energy balancing (Bespalyy & Bespalaya, 2025).

This variability necessitates the inclusion of seasonal factors in forecasting models. Hybrid models that incorporate hydrological and climatic data, such as those applied in the Seyhan Basin, have shown promise in improving water use efficiency and reducing flood risks (Lee et al., 2022). However, these models often require extensive and high-quality datasets, which are not always available in regions with extreme or inconsistent weather patterns (Meng et al., 2023). Machine learning approaches have demonstrated potential in capturing seasonal hydrological patterns, particularly in situations where traditional forecasting models may struggle to represent complex relationships in the data (Alsamraee & Khanna, 2025). Recent approaches combine polynomial regression with precipitation-based predictors to model non-linear seasonal shifts, achieving high predictive accuracy across multiple timeframes (Dahal et al., 2024). Probabilistic seasonal flow forecasts (SFFs) are also gaining traction, particularly in East Asia and South Korea, where ensemble-based forecasting has improved drought preparedness by optimising reservoir operations for 3-month lead times (Gunasinghe et al., 2023). Furthermore, integrating reanalysis datasets, such as MERRA-2, with machine learning techniques enhances seasonal inflow predictions by capturing both historical and real-time precipitation anomalies (Dahal et al., 2024).

2.2.2 Climate Change and Hydrological Uncertainty

Climate change represents a major long-term challenge to the reliability of hydropower forecasting. Shifts in temperature and precipitation patterns, altered snowpack accumulation, and changing river flows significantly influence water inflows to hydroelectric facilities (Hodoroski & Jiang, 2023; Hartono et al., n.d.). These disruptions introduce uncertainties affecting seasonal forecasts and long-term generation planning (Lamidi et al., 2024; Sarpong & Agyei, 2022). For example, changes in river discharge patterns have already led to both surpluses and deficits in hydropower production across various regions (IEEE Power & Energy Society & Institute of Electrical and Electronics Engineers, n.d.). More frequent extreme weather events, such as prolonged droughts and flash floods, make it challenging to maintain consistent inflow predictions, increasing the risk of energy shortages or sudden operational curtailments (Gunasinghe et al., 2023). Climate trends in the Himalayan region show that run-of-the-river plants do not achieve proportional energy gains even with higher rainfall during monsoon seasons due to infrastructure and operational constraints (Dahal et al., 2024).

To address these complexities, researchers have increasingly explored data-driven forecasting approaches capable of modelling nonlinear relationships in hydrological systems. These approaches aim to improve predictive accuracy by capturing complex interactions between climatic variables and hydrological processes. However, forecasting performance remains constrained by climate-induced variability, particularly when historical inflow patterns no longer reflect future hydrological behaviour. Studies in South Asia have indicated that hydropower capacity factors could decline by up to 6.9% by 2100 due to rising temperatures and altered monsoon cycles, highlighting the increasing uncertainty surrounding long-term hydropower production (Gunasinghe et al., 2023).

Scenario-based planning, such as the pumped hydro storage strategy used in Jordan (Almashayikh et al., 2022), also demonstrates how climate resilience can be embedded into long-term hydropower forecasting. Despite these advancements, one of the most significant challenges is the mismatch between increasing hydrological volatility and existing operational frameworks designed for more predictable flow regimes. Without effective integration of climate projections into reservoir planning and grid dispatch, regions heavily reliant on hydro may face worsening energy shortages or

operational disruptions during extreme climate events (Gunasinghe et al., 2023). Nevertheless, these advanced approaches demand significant computational resources and access to high-quality meteorological datasets, which may limit their broad applicability.

.

2.2.3 Operational and Management Challenges

In addition to environmental variability, forecasting is also influenced by operational and management-related complexities (Ahmad et al., 2021: Villeneuve et al., 2023). Efficient hydropower operation depends on various factors, including reservoir characteristics, streamflow conditions, and real-time water usage demands. Reservoir operators often need to balance hydropower generation with competing priorities such as irrigation, flood control, and municipal water supply (Jāmi‘at al-Imārāt al-‘Arabīyah al-Muttaḥidah & Institute of Electrical and Electronics Engineers, n.d.). Variability in upstream inflows and limited storage capacities further complicate predictive accuracy

Accurate forecasts play a vital role in operational decision-making. Studies show that streamflow forecasting can improve the efficiency and cost-effectiveness of water management strategies, provided that the forecast system aligns with the physical and regulatory design of the reservoir (Lee et al., 2022). Jixian et al. (2018) and Nanayakkara et al. (2009) emphasise how infrastructure reliability is impacted by fluctuations in temperature and water flow, which may require frequent equipment maintenance. Furthermore, Feng et al. (2021) highlight the challenge of modelling the inherently non-linear and random behaviour of water level data, particularly when datasets are limited, increasing the risk of overfitting or underfitting forecasting models.

There is therefore a growing need for forecasting approaches that can capture complex hydrological patterns and support more reliable operational planning. Developing models that can improve predictive accuracy under varying environmental and operational conditions remains an important area of research in hydropower forecasting.

2.3 Socioeconomic Indicators and Their Role in Forecasting

Due to their direct and indirect influence on energy demand, forecasting hydroelectric energy generation increasingly incorporates socioeconomic indicators such as GDP, population, and inflation. These indicators help capture broader trends in industrial activity, infrastructure development, and demographic shifts that are critical to the formulation of accurate energy models. As energy plays a vital role in economic development, its demand is often reflective of a nation's growth trajectory, making socioeconomic factors key inputs in long-term energy planning and forecasting (Shahgholian et al., 2023).

Research has shown that electricity consumption patterns, particularly in residential and commercial sectors, are strongly shaped by variables like population density, household occupancy, and economic conditions. For example, predictive models that include occupancy and weather variables alongside energy use data have reported improved forecasting accuracy, with R^2 values exceeding 0.85 (Dubey et al., 2024). This suggests that integrating socioeconomic factors not only contextualises energy use patterns but also improves the precision of forecasting systems.

Furthermore, these indicators are essential in evaluating the broader feasibility of hydroelectric systems. Hydropower plants often serve multiple roles beyond electricity generation, such as irrigation, water storage, and flood control functions that are closely linked to regional development priorities and population needs (Shahgholian et al., 2023). However, the implementation of such systems is frequently constrained by socioeconomic challenges, including fiscal limitations, limited skilled labour, and community opposition to large-scale infrastructure projects (Alshammari, 2022). As such, forecasting models that omit socioeconomic inputs may fall short in accurately capturing the complexity of energy dynamics in developing and transitional economies.

2.3.1 Economic Growth and Energy Demand Nexus

The link between economic growth and energy demand has been widely examined, particularly in the context of electricity consumption. Empirical evidence indicates that rising GDP often correlates with increased energy use, especially in developing economies where economic expansion drives industrialisation and infrastructure growth (Jahanger et al., 2022). In Malaysia, patterns of hydropower

consumption have reflected this trend, aligning with the Environmental Kuznets Curve hypothesis, which posits that environmental impacts initially worsen with growth but improve at later stages of development (Jahanger et al., 2022).

Comparative studies in Europe support the existence of a general positive relationship between economic activity and energy demand, though regional disparities are evident due to different economic structures, energy policies, and consumption behaviours (Datsyuk et al., 2024). Despite this established relationship, much of the existing literature tends to focus on total energy use rather than disaggregating renewable sources like hydropower. This gap limits understanding of how specific economic trends influence hydroelectric power generation and highlights the need for more targeted research in this domain.

In addition to GDP, financial indicators such as inflation also impact the economics of energy systems. Inflation affects both production costs and electricity pricing structures, with higher energy tariffs potentially reducing demand in cost-sensitive sectors (Jahanger et al., 2022). Economic evaluations of hydropower systems are increasingly taking such variables into account, particularly when assessing long-term viability. Villeneuve et al. (2023), in their review of AI-based hydropower forecasting, stress the relevance of incorporating external economic factors, including operational costs and financial pressures, into model frameworks.

2.3.2 The Role of Population in Energy Economics

Population growth plays a significant role in shaping energy economics, directly influencing both electricity demand and broader consumption patterns. As populations expand, particularly in developing countries, the demand for energy-intensive services such as housing, transportation, and industrial activity increases, placing greater pressure on energy infrastructure. These demographic dynamics necessitate the development of more adaptive and context-sensitive forecasting models.

In the Malaysian context, Jahanger et al. (2022) investigated the influence of socioeconomic factors, including population growth, on hydropower consumption. Their study emphasised the broader implications of demographic trends for energy planning and aligned with recent approaches in which ANNs integrate population data alongside economic indicators to improve forecasting accuracy. These models suggest that population trends can play an important role in shaping energy demand trajectories.

Similarly, in Iran, an ANN model incorporating GDP and population growth was effectively applied to forecast long-term energy consumption, reinforcing the importance of including demographic variables in predictive modelling (Villeneuve et al., 2023). These findings suggest that traditional models based solely on energy consumption metrics may overlook essential factors that drive demand, particularly in regions undergoing rapid demographic transformation.

Incorporating population into energy forecasting models improves their ability to capture evolving consumption trends over time. For instance, in smart city contexts, predictive frameworks that account for population shifts have shown promise in aligning energy supply with fluctuating demand, although integrating heterogeneous data sources remains a technical challenge (Dubey et al., 2024). In rapidly developing countries, the inclusion of population in forecasting has led to more robust and reliable predictions of energy needs (Fard & Ardehali, 2023a).

Beyond forecasting, population size and density also influence the design and operation of hydroelectric systems. In high-growth regions, hydropower infrastructure often serves multifunctional roles, including domestic water supply, irrigation, and flood control, highlighting the need for coordinated planning that reflects demographic realities (Shahgholian et al., 2023). However, challenges such as community displacement, fiscal limitations, and labour shortages may hinder the development of such systems, particularly in underserved or rural areas (Shahgholian et al., 2023).

Overall, the integration of population data into energy forecasting not only enhances model accuracy but also provides essential insights for policy and planning. As global demographic trends continue to evolve, incorporating these variables will remain vital for ensuring the feasibility, sustainability, and social responsiveness of hydroelectric energy systems.

Despite the growing body of literature examining the relationship between socioeconomic indicators and energy consumption, relatively fewer studies have specifically investigated how such indicators influence hydroelectric power generation forecasting. Many previous studies have primarily focused on electricity demand forecasting or other renewable energy sources such as solar and wind power (Shahgholian et al., 2023). Consequently, there remains a need to further explore how socioeconomic variables can be integrated into forecasting models to improve the prediction of hydroelectric generation. Addressing this gap may provide additional

insights into how economic and demographic dynamics influence hydropower production and forecasting performance.

2.4 Machine Learning in Energy Forecasting

Machine learning (ML) has emerged as a powerful approach for energy forecasting due to its capacity to model complex, nonlinear, and time-dependent relationships that are often difficult for traditional statistical models to capture. Unlike conventional methods, ML algorithms learn patterns directly from historical data, enabling them to adapt to changing conditions and seasonal variations. In hydroelectric power generation, this ability is crucial because factors such as rainfall, reservoir inflows, temperature, and historical energy output often exhibit high variability and complex interdependencies (Regularized Probabilistic Forecasting of Electricity Wholesale Price_and Demand, n.d.; Shahgholian et al., 2023). By leveraging these data-driven techniques, forecasting accuracy improves significantly, with studies reporting substantial reductions in error metrics like RMSE and MAE when compared to baseline time-series models (Alsamraee & Khanna, 2025).

Artificial Neural Networks (ANN) have played a central role in energy forecasting due to their universal function approximation capabilities and ability to capture nonlinear relationships among multiple input variables. They are particularly effective for time-series data, where historical hydrological patterns and environmental indicators can be mapped to future energy outputs. Studies have shown that ANN models reduce forecast errors by up to 20–30% compared to traditional regression models, particularly when input data is pre-processed through normalisation and noise removal (Alsamraee & Khanna, 2025; Priya et al., 2025). Data decomposition techniques, such as empirical mode decomposition (EMD), have been successfully paired with ANN to separate high-frequency noise from meaningful trends, allowing for improved model convergence and more accurate predictions (Parvathareddy et al., 2025). These improvements highlight the value of combining ANN with systematic feature engineering for hydropower forecasting tasks.

Support Vector Machines (SVMs) are also widely used in energy forecasting, especially when dealing with smaller or medium-sized datasets. By applying kernel functions, SVM can map input data into higher-dimensional spaces to uncover

nonlinear patterns that might otherwise remain hidden (Rajarajeswaran, 2025b). Applications in solar and renewable energy forecasting have shown that SVM, when combined with key meteorological variables such as temperature, humidity, and cloud cover, achieves consistently lower Mean Absolute Percentage Error (MAPE) compared to baseline statistical models (Rajnish et al., 2023). Results from broader energy demand studies reveal that SVM models can reduce forecasting errors by up to 15% compared to persistence models, with MAE values often falling below 3% when preprocessing steps such as feature selection and scaling are applied (Menon et al., 2023). These characteristics make SVM a suitable candidate for modelling complex relationships within energy forecasting datasets.

Ensemble learning has proven to be a highly effective approach in energy modelling due to its ability to combine multiple models to improve prediction accuracy. Techniques such as Random Forest and XGBoost are particularly noted for their robustness and ability to handle complex, nonlinear relationships in data. Random Forest, by combining multiple decision trees, helps reduce variance and overfitting, thus enhancing prediction reliability. It has been successfully applied in energy forecasting tasks, including predicting energy consumption for buildings and hydropower generation based on climate data. In one study, Random Forest demonstrated lower normalised root mean square error (NRMSE) compared to other models, showcasing its effectiveness in handling multivariate regression tasks with features like cooling degree days and historical consumption (*Grande et al.,*2024; Li et al., 2020). These results suggest that Random Forest models are capable of capturing nonlinear relationships between multiple energy-related variables.

XGBoost, an optimised gradient boosting technique, offers strong predictive performance by iteratively correcting the errors of previous models. Known for its high flexibility and accuracy, XGBoost has been widely used in short-term load forecasting and renewable energy prediction. In many cases, XGBoost outperforms other methods, particularly when dealing with large, complex datasets that involve intricate feature interactions. For example, a detailed comparative analysis found that XGBoost achieved superior results with an R^2 value of 0.94, a MAE of 0.8680, and a root mean square error (RMSE) of 1.2078 in long-term energy forecasting tasks (Alsamraee & Khanna, 2025). Similarly, studies focusing on energy utility sectors have reported R^2 values as high as 84% for XGBoost and 75.7% for Random Forest, when forecasting net revenue with small datasets containing engineered lag features (Gupta, n.d.). These

results demonstrate that tree-based models can effectively handle non-linear interactions, though they may still encounter challenges in generalising to unseen data, particularly with limited samples. XGBoost and Random Forest have also been applied in modelling CO_2 emissions from energy consumption and transportation sectors, where XGBoost achieved RMSE values as low as 0.036 in complex urban datasets (Abbasian Hamedani & Talebi, 2025). Random Forest demonstrated superior performance compared to support vector regression (SVR) and other baseline models, achieving MAPE values around 0.072, which underscores its capacity for accurate forecasting in energy-related emissions studies. These results underscore its strong generalisation capabilities, even in scenarios with partial missing data, where the RMSE increase was minimal at just 0.24 MWh with 10% data gaps (Alsamraee & Khanna, 2025).

Choosing the right settings (hyperparameters) is a vital step in making machine learning models work. When working with smaller datasets, researchers often choose standard or widely used settings instead of searching for perfect values through trial and error. This is done to prevent overfitting, where the model becomes too focused on small errors in the data rather than the overall trend. Studies in energy forecasting (Benti et al., 2023), (Barzola-Monteses et al., 2025), (Sattar Hanoon et al., 2023b) show that using these established settings such as 10 hidden neurons for ANN or specific learning rates for trees helps the model perform better on new, unseen data. This approach ensures the model focuses on real energy-economy trends rather than random noise in the historical records.

In conclusion, machine learning techniques such as Artificial Neural Networks, Support Vector Machines, Random Forest, and XGBoost have been widely applied in energy forecasting due to their ability to capture nonlinear relationships and manage complex datasets (Alsamraee & Khanna, 2025), (Rajnish et al., 2023). Each model offers different learning mechanisms, ranging from neural network structures and kernel-based learning to ensemble tree algorithms. Previous studies have demonstrated that these approaches can improve forecasting accuracy when modelling complex energy systems and multivariate datasets. Therefore, these four machine learning models are considered suitable for evaluating hydroelectric generation forecasting performance when socioeconomic indicators are used as predictive inputs.

2.4.1 Limitations of Traditional Forecasting Models

Traditional time series models such as Autoregressive Integrated Moving Average (ARIMA), exponential smoothing, and multiple linear regression have long been used in hydropower forecasting due to their simplicity and interpretability (Barzola-Monteses et al., 2019; Kumari et al., 2022; Shoaga et al., 2022). ARIMA models are particularly effective in modelling stationarity and seasonality. For instance, studies in Ecuador and Rwanda used Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) and Seasonal Autoregressive Integrated Moving Average (SARIMA) variants to model monthly hydropower production, integrating rainfall and seasonal inputs for improved accuracy (Barzola-Monteses et al., 2019; Shoaga et al., 2022). However, while these methods are efficient in linear trend detection and short-term forecasting, they often fail to account for the complex interactions that arise in nonstationary or multi-variable systems commonly found in renewable energy forecasting (Alsamraee & Khanna, 2025). Their reliance on fixed statistical assumptions can limit adaptability, especially when input variables, such as rainfall or inflow rates, exhibit strong nonlinearities or sudden fluctuations (Y. Wang, 2024).

However, these statistical models rely on linear assumptions and often fall short when modelling nonstationary or highly nonlinear hydrological systems. Regression models may also struggle to accommodate shifting relationships between input variables over time, such as changes in rainfall patterns or energy demand (S. Wang & Ma, 2023). They are less suitable for capturing extreme or rare events, such as peak inflows or unexpected operational disruptions, which are critical for risk management in hydropower operations (Z. Zhang et al., 2023). These limitations have become more evident with the growing influence of climate variability and dynamic energy demand patterns, both of which require models that can learn and adapt in real time.

In addition, traditional statistical models are often less capable of predicting extreme values such as peak inflows or sudden output drops, which are crucial for operational and risk planning in hydroelectric systems (Paul et al., 2021). As hydropower systems become increasingly influenced by climatic variability and changing energy demand patterns, forecasting models must be able to capture nonlinear relationships and interactions among multiple variables. Consequently, machine learning techniques have increasingly been explored as alternative approaches because

of their ability to model complex and nonlinear systems without relying on strict statistical assumptions (Y. Wang, 2024; Z. Zhang et al., 2023). These characteristics make machine learning models particularly suitable for energy forecasting applications involving large and dynamic datasets.

2.5 Artificial Neural Network (ANN) in Energy and Hydropower Forecasting

Artificial Neural Networks (ANN) have gained growing interest in energy forecasting studies due to their ability to capture complex and nonlinear patterns in data. These models are beneficial when traditional statistical approaches fall short, mainly when the input variables include multiple interrelated factors such as energy consumption, economic growth, and environmental conditions. Although widely applied in various energy domains such as wind, solar, and electricity demand forecasting, ANNs have also shown potential in hydropower-related studies.

2.5.1 ANN Application in Energy Forecasting

Artificial Neural Networks (ANNs) have become increasingly popular in energy forecasting due to their ability to model complex nonlinear relationships between input variables. In the context of wind power, optimised ANN models have been applied for short-term and ultra-short-term forecasting, with results showing significant accuracy improvements when hyperparameters are finely tuned (Zha & Jiang, 2022a). Another study successfully employed ANNs to forecast weather parameters essential for sizing stand-alone renewable power systems, underscoring the flexibility of ANN models across different renewable energy contexts (N. K. Dubey et al., 2021).

Hydropower forecasting also benefits significantly from ANN integration. At the Nam Theun 2 Hydropower Plant, ANN-based models have also been applied to predict energy generation at hydropower plants, demonstrating their usefulness in modelling complex energy systems (Kongpaseuth & Kaewarsa, 2023). Similar efforts have utilised particle swarm optimisation with ANN to enhance prediction accuracy in forecasting available energy at hydropower plants (Kaewarsa & Kongpaseuth, 2024). Recent studies have reported high accuracy in ANN-based hydropower forecasting, such as in major Nigerian hydropower plants, where daily inflow and generation forecasts achieved correlation coefficients above 0.96 using optimised ANN

architectures (Nwobi-Okoye et al., 2024). Likewise, multi-layer perceptron (MLP) models have been implemented in Italy to forecast hydropower production 100 days ahead, using weather-driven lag features and walk-forward validation for improved reliability (Grande et al., 2024).

Beyond hydropower, ANNs have also been applied to broader energy forecasting domains, including electricity consumption and greenhouse gas emissions. For instance, in a study involving six Asian countries, ANN models that incorporated socioeconomic variables such as GDP, population, and primary energy consumption successfully captured the dynamic relationship between energy use and emissions (Abbasian Hamedani & Talebi, 2025). The integration of optimisation techniques such as particle swarm optimisation (PSO) and grey wolf optimiser (GWO) into ANN frameworks has been shown to further enhance performance by fine-tuning model parameters and reducing forecasting errors. This hybrid approach is particularly beneficial when dealing with large and complex datasets where traditional training algorithms may struggle to converge effectively.

In long-term electricity demand forecasting, ANN models have proven their ability to handle multivariate inputs and nonlinear patterns that are difficult for statistical methods to capture. A study focusing on India's electricity demand demonstrated that ANN models performed better than traditional approaches like ARIMA and multiple linear regression, especially when economic and demographic indicators were used as predictors (Salem et al., 2025). While support vector regression (SVR) slightly outperformed ANN in certain metrics, the ANN framework was still recognised for its flexibility and adaptability to complex forecasting tasks. These findings highlight the growing role of ANN models in providing data-driven insights for energy planning and policy development.

2.5.2 ANN in Socioeconomic-Based Energy Forecasting

ANNs have also been successfully employed in predicting energy consumption based on socioeconomic indicators. For instance, ANN models have been developed to forecast electricity demand using temperature and economic variables, outperforming conventional methods due to their flexibility in capturing multivariate nonlinear relationships (Q. Wang et al., 2020). When applied to energy consumption in buildings, ANNs effectively captured intricate demand patterns that traditional models struggled

to represent (Luo & Li, 2023). In hydropower modelling, ANNs have shown promising results in forecasting both reservoir inflows and energy output, although results heavily depend on input configuration and data quality (Paul et al., 2021). Furthermore, the strength of ANNs lies in their ability to approximate continuous functions, making them well-suited for dynamic systems like electricity markets or hydropower generation, especially under fluctuating seasonal or economic conditions (Kumari et al., 2022). ANN models have also shown notable success in wind speed forecasting, providing more accurate results than ARIMA models, particularly when handling data with high variability (IEEE Staff, 2011).

In addition, long-term studies have confirmed the effectiveness of ANN models when socioeconomic indicators such as GDP, population, and energy trade (exports and imports) are used as inputs for national-scale forecasting. For example, an ANN model optimised using PSO and improved PSO (IPSO) was employed to forecast Iran's electrical energy consumption from 2023 to 2050. The results indicated that incorporating socioeconomic variables led to improved forecasting accuracy, with an MAPE as low as 1.94% (Fard & Ardehali, 2023). The study also compared multiple ANN configurations and found that deeper architectures, when properly optimised, yielded more reliable long-term forecasts. For instance, an IPSO-ANN model was applied to forecast Iran's electricity consumption up to 2050, demonstrating the capability of ANN models to capture long-term energy demand trends influenced by socioeconomic development (Fard & Ardehali, 2023). This suggests that, when properly trained and tuned, ANN models are capable of capturing long-range energy consumption dynamics influenced by national socioeconomic development trends. However, as highlighted in the study, the performance of these models remains highly sensitive to network architecture and the selection of appropriate input features (Fard & Ardehali, 2023).

2.5.3 Limitations, Overfitting, and Learning Curve Considerations in ANN Models

Despite their advantages in predictive accuracy, ANN models present several limitations that must be acknowledged, particularly in the context of energy forecasting. One of the most significant concerns is overfitting, where a model performs well on training data but fails to generalise to new, unseen datasets. This problem becomes more

pronounced when models are trained on noisy or limited data, leading to poor forecasting reliability in real-world applications (*Regularized Probabilistic Forecasting of Electricity Wholesale Price and Demand*, n.d.). Furthermore, ANN models often demand extensive computational resources and large datasets to achieve stable and accurate performance, which may not be feasible in all forecasting environments (Ersan & Irmak, 2023).

Another key challenge is the lack of interpretability. Unlike traditional regression models, ANN structures operate as a "black box," making it difficult to trace how input features influence the final output. This poses limitations in policy-making and other areas where model transparency is critical for decision-making and reporting (Sauhats et al., 2016). Additionally, tuning the network's hyperparameters, such as learning rates, the number of hidden layers, and neuron configurations, often requires trial-and-error experimentation, demanding considerable technical expertise and computational time.

Overfitting in ANN models has been widely discussed in recent literature, with various strategies proposed to enhance generalisation. For instance, in the field of smart hydropower management, machine learning methods were applied with careful parameter tuning to maintain high accuracy while avoiding overfitting (Sahin & Ozbay Karakus, 2024). Likewise, in short-term load forecasting, it was shown that performance could be significantly improved and overfitting reduced by optimising the ANN's architecture and learning parameters (*IEEE, 2020*). These findings underscore the importance of model design in ensuring consistent predictive power.

To further address overfitting, learning curves serve as a valuable diagnostic tool. They help visualise how well the model is learning by comparing performance on training and validation sets across epochs. For example, studies on wind power prediction demonstrated that learning rate adjustments and network structure refinements based on learning curve trends generalisation (Zha & Jiang, 2022). Similarly, in hydropower forecasting, optimisation techniques like particle swarm optimisation were integrated with ANN models to guide the training process and ensure reliable outcomes throughout the learning phase (Kaewarsa & Kongpaseuth, 2024).

Moreover, comparative studies have shown that while ANN models often outperform traditional regression approaches in accuracy, they are more vulnerable to overfitting if not properly calibrated (Predicting World Energy Consumption Comparison of ANN and Regression Analysis, n.d.).

2.6 Summary of Literature Review

This chapter reviewed the key literature related to hydroelectric generation forecasting and the factors influencing energy production. The discussion first examined the importance of hydropower forecasting and the environmental and operational challenges that affect hydroelectric systems, including seasonal variability, climate change, and operational constraints.

The literature also highlighted the influence of socioeconomic indicators such as GDP and population on energy demand patterns. Previous studies have shown that incorporating these variables can provide a more comprehensive understanding of long-term energy system behaviour. However, relatively limited research has focused specifically on integrating socioeconomic indicators into hydroelectric generation forecasting models.

In addition, this chapter reviewed the application of machine learning techniques in energy forecasting. Methods such as Artificial Neural Networks, Support Vector Machines, Random Forest, and XGBoost have demonstrated strong capabilities in modelling nonlinear relationships and complex datasets. The limitations of traditional statistical forecasting models were also discussed, highlighting the need for more flexible data-driven approaches.

Overall, the literature indicates that machine learning models provide promising tools for improving hydroelectric forecasting accuracy. Based on these findings, the next chapter presents the research methodology used to develop and evaluate the machine learning models applied in this study.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This study adopts a structured methodological framework to investigate the relationship between socioeconomic factors and hydroelectric power generation in Malaysia. The methodology is organised into several key stages: data collection, preprocessing, model development, training, and performance evaluation. Initially, the modelling process focused exclusively on ANN due to its strong capability in capturing nonlinear relationships in time-series data. Three different input configurations are trained using an ANN. The goal of this phase was to identify the most effective input combination for predicting hydroelectric generation. After selecting the most promising input configuration based on ANN results, the study proceeded to apply three additional machine learning models SVM, Random Forest, and XGBoost) using the same chosen input variables. This allowed for a fair and meaningful comparison of how different modelling techniques perform under identical data conditions, providing insight into their relative strengths and practical applicability in the energy forecasting context. The methodological framework is illustrated in the flowchart (Figure 3-1), which visually outlines the sequence of tasks and decisions taken throughout the study.

The methodology emphasises transparency and comparability throughout all stages of model development. All models were trained using the same dataset and input combinations, and preprocessing methods were applied in line with each model's algorithmic requirements. The training and evaluation phases employed consistent data splits and performance metrics to prevent bias in the outcome assessment. Particular attention was given to input variable preparation, model configuration, and the use of both numerical and visual tools for evaluating predictions. By maintaining a cautious and structured approach, the methodology aims to generate results that are interpretable, and relevant for understanding the potential role of data-driven models in forecasting hydroelectric generation within a broader energy-economy context.

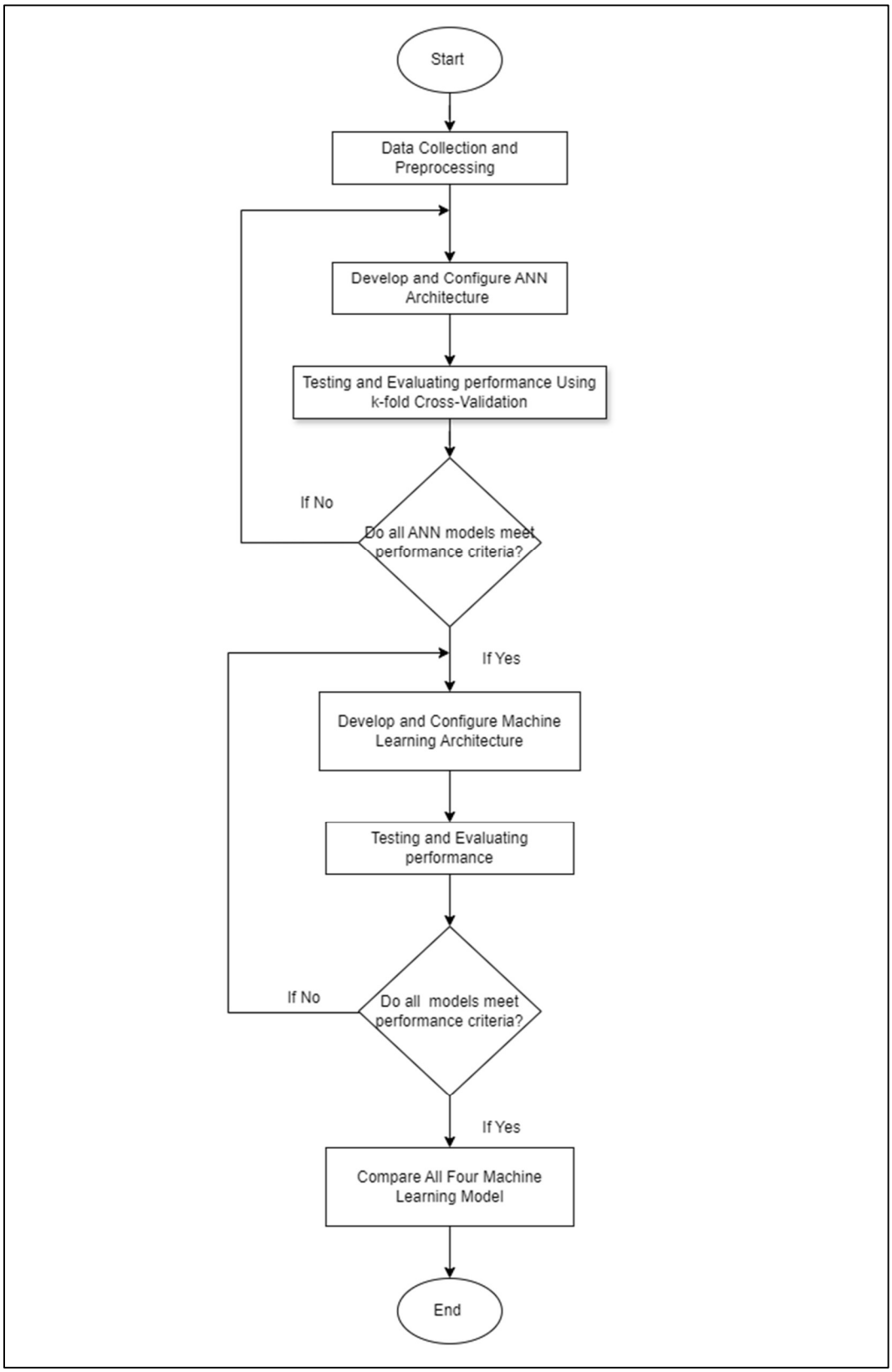

Figure 3.1 Flowchart

3.2 Data Collection and Preprocessing

The dataset used in this study consists of annual observations from 1980 to 2021, collected from two well-recognized online sources, namely the Malaysia Energy Information Hub and Macrotrends. The Malaysia Energy Information Hub, which is administered by the Energy Commission of Malaysia (Suruhanjaya Tenaga), provides official energy statistics that are regularly updated and validated through national reporting protocols. These data are widely referenced for energy planning, system modelling, and policy evaluation. Macrotrends compiles long-term economic indicators from reputable international organisations such as the World Bank and the International Monetary Fund. Its extensive time-series coverage makes it an appropriate source for macroeconomic variables that span several decades. By combining these two data repositories, the study ensured that both the energy-related and economic indicators were reliable and consistent over the observed period.

It should be noted that the hydroelectric generation data obtained from the Malaysia Energy Information Hub for this study cover the period from 1980 to 2021, which represents the most recent official data available at the time of this research. As a result, the dataset is limited to the available historical records. To ensure reliable model training and evaluation despite the limited number of observations, k-fold cross-validation was applied during the modelling process. This approach allows the available data to be utilised more effectively during model development while maintaining a fair evaluation procedure. A more detailed explanation of the cross-validation process is provided in Section 3.3.

The input variables selected for this study were GDP, Energy Consumption, Population, and Inflation Rate. These variables were chosen because they are commonly recognised as key drivers of energy demand and renewable energy generation in the literature. GDP is a fundamental measure of economic performance, reflecting the level of industrial and commercial activity that often correlates with electricity demand. Energy Consumption captures historical usage patterns, which are critical for understanding and forecasting future load requirements. Population is an essential demographic factor since changes in population size and distribution can directly affect infrastructure needs and energy demand. Inflation Rate was included to account for macroeconomic instability and price fluctuations that may indirectly influence

investment behaviour, energy policies, and operational costs within the energy sector. Together, these indicators provide a comprehensive overview of the economic and demographic factors that shape hydroelectric generation.

The output variable for all models developed in this study is Hydroelectric Generation, which is expressed in gigawatt-hour (GWh). This measure represents the total annual electricity produced by hydroelectric sources in Malaysia and directly reflects the country's renewable energy contribution to the national power grid. Before conducting any modelling, the dataset was examined thoroughly to ensure quality and consistency. Both visual inspection and statistical analyses were performed to detect missing values, anomalies, and abnormal trends. Missing values were addressed through appropriate imputation methods. For continuous variables such as GDP and Energy Consumption, linear interpolation was employed because these indicators typically follow smooth, predictable trends over time. For relatively stable variables such as Population, mean substitution was used to fill in missing entries while avoiding artificial fluctuations. These data cleaning procedures were performed with care to minimise the risk of introducing distortions that could compromise the training and generalisation performance of the models.

Due to the differences in scale and units across the input variables, normalisation was carried out to prepare the data for the machine learning models. For the ANN model, Min–Max normalisation was applied to rescale all variables into a [0, 1] range. This scaling method improves gradient-based optimisation by ensuring faster convergence and more stable weight updates during training (Kim et al., 2025). The Min–Max normalisation is shown in Equation (3.1). For the SVM, Random Forest, and XGBoost models, Z-score normalisation was applied instead of Min–Max scaling. This method is particularly effective for algorithms that are sensitive to the magnitude of input features, as it standardises each variable to have zero mean and unit variance (Kim et al., 2025). The Z-score transformation was computed as shown in Equation (3.2).

To avoid data leakage, the same statistical parameters derived from the training subset were applied to normalise the validation and testing subsets. This ensures that the models are evaluated under conditions similar to those encountered during training and prevents any inadvertent exposure of future information to the learning process.

Preliminary experiments with the ANN models revealed that including the Inflation Rate as an input variable did not improve predictive performance. In some cases, its inclusion slightly increased variability in model outputs, likely due to its

indirect and less consistent relationship with hydroelectric generation compared to the other variables. Consequently, the Inflation Rate was excluded from the final input configuration used for the comparative evaluation of the SVM, Random Forest, and XGBoost models. The refined set of inputs, consisting of GDP, Energy Consumption, and Population, was then split into training (70%), validation (15%), and testing (15%) subsets. The splitting process preserved both the statistical distribution and the chronological order of the data to ensure that each subset accurately represented the temporal characteristics of the full dataset. This approach allowed for a fair and consistent comparison of model performance while maintaining the integrity of time-series forecasting.

$$x_{norm} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{3.1}$$

$$z = \frac{(x - \mu)}{\sigma} \tag{3.2}$$

3.3 Cross-Validation

To evaluate the generalisation performance of the ANN and guide the selection of the most relevant input features, a five-fold cross-validation approach was implemented using MATLAB. Cross-validation is a widely adopted resampling technique in machine learning that provides an unbiased estimate of model performance, particularly when the dataset is limited or moderate in size. It reduces the risk of overfitting by ensuring that every data point is used for both training and testing across multiple iterations, thereby providing a more realistic assessment of how the model will perform on unseen data. K-fold cross-validation achieves a balance between computational efficiency and performance variability, mitigating the influence of any single train-test partition. In this study, a five-fold configuration (k = 5) was selected to maintain sufficient data for training in each fold while retaining the ability to test the model on a representative set of unseen samples during every cycle.

Before executing cross-validation, the dataset was arranged in a column-based format, with each row representing a specific year and each column representing a variable, including both input features and the output target. The dataset was divided into five equal subsets, known as folds. In each iteration, four folds were used for training the ANN, while the remaining fold was used exclusively for testing. This process was repeated five times, with each fold serving as the test set exactly once. The performance metrics from all iterations were then averaged to produce a stable and reliable estimate of the model's predictive accuracy. This approach ensured that the evaluation was not dependent on a particular data split and that the selection of input variables was guided by data-driven insights rather than arbitrary assumptions. Such systematic partitioning is consistent with standard practices in applied energy forecasting studies, as discussed by Zhang and Liu (2023), where emphasis is placed on model generalisation rather than overfitting to training data.

In each fold, the ANN model employed a feedforward architecture with a single hidden layer comprising ten neurons. This configuration was chosen because it balances simplicity with the ability to approximate complex nonlinear relationships. The universal approximation theorem states that a neural network with a single hidden layer and a sufficient number of neurons can approximate any continuous function to a desired level of accuracy, making this structure an effective baseline for evaluating variable relevance. Ten neurons were selected as a practical choice, providing adequate capacity for learning patterns without unnecessarily increasing complexity. Training was carried out using the backpropagation algorithm with MSE as the loss function. A separate validation subset was intentionally excluded within each fold since the goal of this cross-validation stage was not hyperparameter optimisation but rather the comparative evaluation of different input combinations. Keeping the model structure consistent across folds also ensured a fair and unbiased comparison of performance metrics.

Model evaluation during cross-validation relied on two key metrics: MSE and the R-value. MSE quantifies the average squared difference between predicted and observed values, thereby penalising larger deviations more heavily and providing a clear measure of prediction accuracy. As highlighted by Salem et al. (2025), MSE is particularly useful in regression-oriented energy forecasting because it directly reflects numerical precision. The R-value, on the other hand, measures the strength and direction of the linear relationship between predicted and actual values. A higher R-

value indicates that the model captures the trend of the data more effectively, which is especially important in time-series datasets like annual hydroelectric generation. By combining these two metrics, the evaluation framework provided both a measure of error magnitude (through MSE) and a measure of trend alignment (through R-value), offering a comprehensive perspective on performance.

Beyond its role in numerical performance evaluation, the five-fold cross-validation procedure was pivotal for refining the final set of input features. By ensuring that each data point contributed to both training and testing across multiple iterations, the method minimised biases arising from any single partition and produced a more robust estimate of generalisation error. This process confirmed which input configurations consistently yielded superior predictive results, guiding the selection of variables for subsequent modelling with ANN, SVM, Random Forest, and XGBoost. As noted by Teodorescu and Obreja Brașoveanu (2025), rigorous cross-validation practices such as this are essential in predictive modelling, particularly when variable selection can have a direct impact on forecasting accuracy and practical applicability.

3.4 Evaluating ANN Models for Energy-Economy Forecasting

The methodological framework outlines the steps taken to develop and evaluate ANN models in examining the relationship between socioeconomic indicators and hydroelectric power generation. This section builds on the earlier data preparation efforts and introduces the modelling phase, which focuses on understanding how economic and demographic variables contribute to predictive accuracy.

The ANN modelling approach was selected due to its capacity to handle nonlinear relationships and to learn patterns from moderately sized datasets. To ensure reliable evaluation, all models were developed using consistent preprocessing techniques, a uniform network structure, and identical performance metrics. This allowed for a systematic assessment of how varying combinations of input features influenced prediction outcomes.

Rather than immediately assuming which variables would be most influential, the study employed different configurations to reflect practical economic scenarios. These variations, along with the architectural details, are presented in the next section. By isolating different socioeconomic indicators in each configuration, the framework

allowed the research to assess the relevance of each factor in forecasting hydroelectric generation, while maintaining fairness and comparability across models.

3.4.1 Model Architecture and Configuration

Following the methodological overview, this section describes the architectural design and configuration of the three ANN models developed to analyse how selected socioeconomic indicators influence hydroelectric power generation. The models were implemented using MATLAB's Neural Net Fitting tool, a supervised learning environment designed for function approximation and regression tasks. This tool provides a streamlined interface for building feedforward neural networks, offering integrated features for data division, training, validation, and performance visualisation. Its algorithmic robustness and ability to model nonlinear relationships make it particularly well-suited for energy forecasting applications, where the input-output interactions are often complex and lack clear analytical formulations.

The choice of ANN as the primary modelling framework is supported by its proven ability to capture nonlinear dependencies between predictors and target outputs. The universal approximation theorem states that a feedforward neural network with a single hidden layer and a sufficient number of neurons can approximate any continuous function to an arbitrary level of accuracy when provided with adequate data and appropriate training. This theoretical foundation has driven widespread adoption of ANN techniques in energy-related studies, where forecasting tasks require models that can adapt to complex trends and variable interactions.

A single ANN architecture was employed with three different input configurations to examine how varying combinations of input variables influence forecasting performance:

- **Model 1:** Gross Domestic Product (GDP) and Energy Consumption.
- **Model 2:** GDP, Energy Consumption, and Population.
- **Model 3:** GDP, Energy Consumption, and Inflation Rate.

The selection of these input combinations was guided by empirical and policy relevance rather than by automated feature selection techniques. This approach reflects practical scenarios, such as evaluating the effect of population growth or inflationary trends on renewable energy generation. Similar studies have suggested that selecting variables based on domain knowledge, rather than algorithmic filtering alone, enhances

the interpretability of results and strengthens their value for policy and planning (Hooshyar et al., 2024).

To ensure a fair and unbiased comparison, all three ANN models shared the same network architecture. Each model consisted of an input layer with two or three neurons corresponding to the selected predictors, a single hidden layer containing ten neurons, and an output layer with one neuron representing hydroelectric generation, as illustrated in Figure 3.2. This configuration was chosen to balance modelling capacity with computational efficiency. The use of ten neurons in the hidden layer was determined through preliminary trials and is consistent with common practices in energy forecasting, where moderately sized hidden layers have been shown to effectively capture nonlinear dynamics without introducing excessive risk of overfitting (Scorzato, 2024).

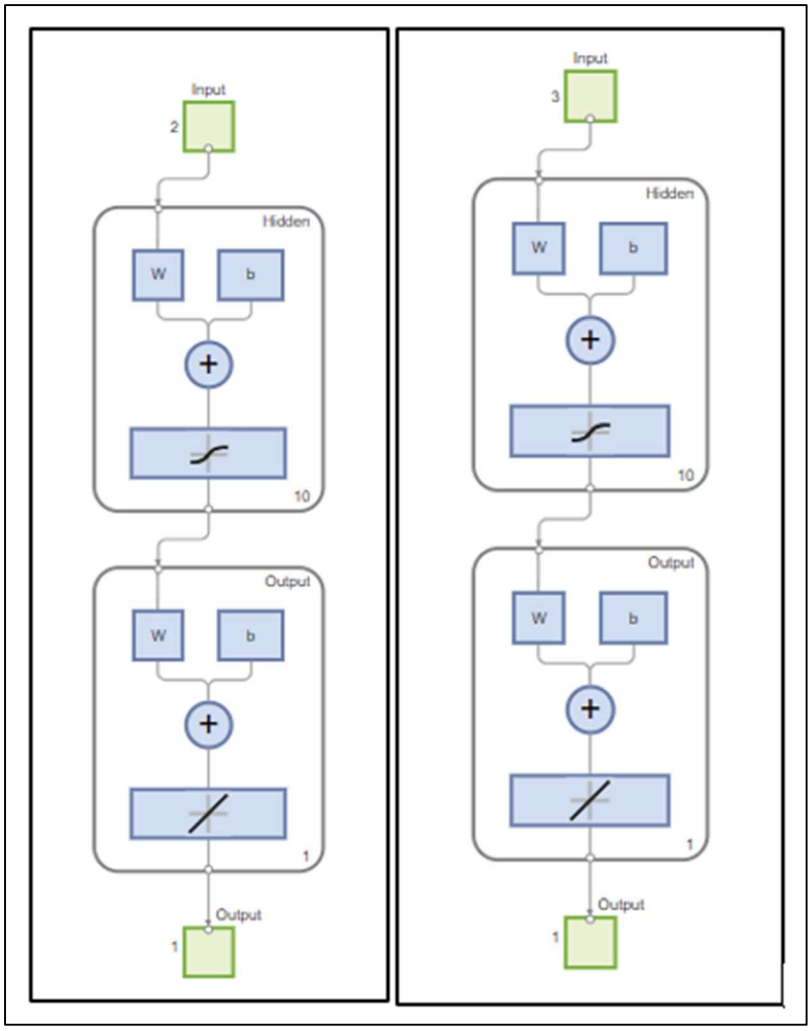

Figure 3.2 Architecture of ANN

Model training was carried out using the Levenberg–Marquardt algorithm across all three ANN configurations. The specific training parameters and convergence indicators for each model are summarised in Table 3.1.

Table 3.1
Training Parameters for ANN Configurations

Model	Training Algorithm	Stopping Iteration	Target	Gradient	Mu	Validation Check
1	Levenberg-Marquardt	13	1000	2.93x10^7	0.001	6
2	Levenberg-Marquardt	13	1000	2.4x10^7	0.001	6
3	Levenberg-Marquardt	10	1000	9.15x10^6	0.001	6

This algorithm is widely recognised for its speed and accuracy in nonlinear regression tasks, combining the advantages of both gradient descent and the Gauss–Newton method to efficiently minimise the mean squared error (MSE). The Levenberg–Marquardt algorithm is widely adopted in ANN-based forecasting studies because of its fast convergence and high optimisation efficiency when training multilayer perceptron networks on small to medium-sized datasets. The Levenberg–Marquardt algorithm was selected as the training method for the ANN model because it is widely used for nonlinear regression problems and has been applied in various energy forecasting studies. This algorithm provides efficient weight optimisation and stable convergence behaviour when training feedforward neural networks on datasets of moderate size. The stopping criterion in all models was set to a target of 1000 epochs, although early stopping occurred well before reaching this limit due to convergence based on validation performance.

The final gradient values also differed, with Models 1 and 2 achieving 2.93×10^{7} and 2.40×10^{7}, respectively, while Model 3 reached a lower gradient of 9.15×10^{6}. The final gradient values differed across models, reflecting variations in the optimisation process during training. The parameter Mu, maintained at 0.001 for all three models, governed the adaptive learning rate in the Levenberg–Marquardt algorithm, balancing

between the steepest descent and Gauss–Newton updates. A constant Mu value reflects stable convergence behaviour without the need for significant adjustments during training.

Additionally, all models exhibited six validation checks before stopping, meaning no further improvement in validation performance was observed for six consecutive iterations. This safeguard ensured that training ceased before overfitting occurred, preserving generalisation ability. The target value, representing the mean squared error goal, was set at 1000 across all models but was not fully reached due to early stopping based on validation performance. By maintaining a uniform experimental setup, any observed differences in model performance can be attributed solely to the variation in input variables rather than inconsistencies in training conditions. This approach reinforces the credibility of the comparative analysis and is aligned with established best practices in applied machine learning research. Systematically comparing training behaviour and parameter convergence across all three configurations, insights can also be gained into the relative difficulty of learning from each input combination.

By systematically varying the input combinations while keeping the architecture and training conditions constant, this modelling framework provides a robust foundation for understanding how economic activity, energy demand, demographic factors, and macroeconomic conditions influence hydroelectric power forecasting. The results from these models form the basis for subsequent evaluations of prediction accuracy.

3.4.2 Performance Evaluation

To assess the predictive capability of the ANN models, two primary evaluation metrics were used: Mean Squared Error (MSE) and the correlation coefficient (R-value). MSE quantifies the average squared difference between the predicted and actual hydroelectric power generation values, with lower values indicating higher prediction accuracy. It is calculated using Equation (3.3), which y_1 represents the actual values, $\hat{y}_1$ represents the predicted values, and n is the number of test samples. The R-value, as defined in Equation (3.4), measures the strength of the linear relationship between actual and predicted values. A value closer to 1 suggests a stronger correlation,

indicating that the model is better at capturing the underlying patterns in the data. In this formula, ŷ is the mean of the actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_1 - \hat{y}_1)^2 \tag{3.3}$$

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_1 - \hat{y}_1)^2}{\sum_{i=1}^{n}(y_1 - \hat{y})^2}} \tag{3.4}$$

In addition to numerical metrics, graphical analysis tools were used to further evaluate the performance of the ANN models developed with different input combinations. Error histograms were generated using the built-in tools of the Neural Net Fitting App to visualise the distribution of prediction errors across the training, validation, and test sets. These histograms display the frequency of errors across defined intervals along the x-axis, where each bar represents how many predictions fall within a particular error range. A vertical line at zero indicates the point of perfect prediction accuracy.

To interpret the histograms, emphasis is placed on the shape and concentration of the bars relative to the zero-error line. When the majority of bars are tall and closely grouped near the centre, it indicates that the model's predictions were generally close to the actual values. By observing the section of the histogram where most of these bars appear, one can estimate the typical range of prediction errors without relying solely on numeric summaries. A narrower, more centred distribution implies better consistency and generalisation, while wider or uneven spreads may suggest less stable performance or the presence of outliers. This visual interpretation supported the comparative analysis of input configurations by offering insight into the stability and accuracy of each model’s predictions.

Regression plots were also produced, displaying predicted values against actual values for each test sample. These plots provided a visual assessment of the model’s fit, where closer alignment to the 45-degree line indicated better performance.

This allowed for quick identification of under- or overestimations across different magnitudes of hydroelectric power output.

Lastly, gradient curves were examined during training to observe how model weights evolved over time. These curves, which show the magnitude of gradient updates across iterations, helped confirm whether the network was converging properly. A steadily declining gradient indicated that the learning process was stabilising, whereas erratic changes could signal difficulties in optimisation or learning rate settings. Together, these visual diagnostics complemented the MSE and R-value results by offering additional perspectives on model behaviour and training dynamics.

3.5 Comparative Framework of Machine Learning Models for Forecasting Hydroelectric Generation

Before comparing the four machine learning models, the architecture and configuration of each algorithm were first established to ensure that the subsequent comparison was conducted under consistent experimental conditions. The ANN models provided a baseline for evaluating the relationship between socioeconomic indicators and hydroelectric generation. However, relying on a single modelling approach can limit the depth of analysis. To enhance the evaluation, three additional machine learning algorithms were included: SVM, Random Forest, and XGBoost. These models were selected due to their established effectiveness in regression problems and their ability to handle structured datasets.

Each of the additional models offers distinct strengths based on different learning principles. SVM is known for its ability to model complex, nonlinear patterns using kernel functions, while Random Forest applies ensemble learning through bootstrap aggregation to improve generalisation and reduce variance. XGBoost builds upon gradient boosting by incorporating regularisation and sequential tree refinement to enhance accuracy while mitigating overfitting. Including these three models alongside ANN provides a broader perspective on predictive performance and highlights how different algorithmic paradigms respond to identical socioeconomic input conditions.

To ensure a fair comparison, all models were trained on the same dataset and input variables, which included GDP, Energy Consumption, and Population. Data preprocessing and normalisation procedures were applied as described in Section 3.2,

ensuring consistency in feature scaling across all models. The dataset was divided into training (70%), validation (15%), and testing (15%) subsets, maintaining uniform partitioning to avoid bias caused by variations in data handling.

The evaluation of all models was based on MSE and the R-value, which together provide a balanced measure of predictive accuracy and trend alignment. The aim of this comparative analysis was not solely to identify a single best-performing model but to explore the strengths and weaknesses of each approach. By adopting this framework, the study ensures that the insights derived from the analysis are robust and reflect the diverse modelling strategies applied to the same dataset.

3.5.1 Machine Learning Model Architecture and Configuration

In addition to the ANN architecture, the configuration of the other machine learning models (SVM, Random Forest, and XGBoost) is described in this subsection to ensure that the comparative analysis is conducted under clearly defined modelling conditions. Table 3.2 summarises the key parameter settings used for each machine learning model. These include both manually specified values and the default parameters applied during implementation in MATLAB.

The parameter values used for the machine learning models were selected based on configurations commonly adopted in previous energy forecasting and regression studies (Barzola-Monteses et al., 2025; Benti et al., 2023; Sattar Hanoon et al., 2023). These settings provide a practical baseline for evaluating model behaviour while minimising excessive hyperparameter tuning, which could otherwise bias the comparison across algorithms. Consistent with the theoretical approach discussed in Chapter 2, these settings provide a practical baseline for evaluating model behaviour while minimising excessive hyperparameter tuning, which could otherwise bias the comparison across algorithms.

Table 3.2
Machine Learning Model Configuration

Model	MATLAB Tool / Function	Learning Method	Key Parameters (Values)
ANN	Neural Net Fitting App	Levenberg–Marquardt	Hidden layer: 10 neurons Performance goal: 0 Max iterations: 1000

			Validation checks: 6 Initial Mu: 0.001
SVM	fitrsvm	Support Vector Regression (RBF Kernel)	Box constraint: 1 Kernel scale: auto Epsilon: 0.1
Random Forest	fitrensemble	Bagging (Ensemble Trees)	Learning cycles: 100 Learner type: Tree Max splits: 10 Min leaf size: 1
XGBoost	fitrensemble	LSBoost (Boosted Trees)	Learning cycles: 100 Learning rate: 0.1 Learner type: Tree

Table 3.3 summarises the training progress and convergence behaviour of the ANN model during optimisation. The table presents key indicators such as iteration count, performance improvement, gradient magnitude, damping factor (Mu), and validation checks, which collectively describe the stability and progression of the learning process.

Table 3.3
ANN Training Parameters and Convergence Summary

Metric	**Initial Value**	**Stopped Value**	**Target Value**
Iteration	0	13	1000
Elapsed Time	–	00:00:03	–
Performance (MSE)	1.46e+07	4.03e+03	0
Gradient	2.4e+07	1.02e+04	1e-07
Mu	0.001	100	1e+10
Validation Checks	0	6	6

The ANN model was implemented using MATLAB's Neural Net Fitting App, which provides a graphical interface for configuring and training feedforward neural networks for regression problems. A relatively simple architecture consisting of a single hidden layer with ten neurons was selected to balance modelling capacity with computational efficiency. A moderate number of hidden neurons is commonly used in energy forecasting studies because it allows the network to capture nonlinear relationships while reducing the risk of overfitting when the dataset is limited in size. The training was carried out using the Levenberg–Marquardt algorithm, which is widely applied in nonlinear regression tasks due to its fast convergence and stable optimisation behaviour when training feedforward neural networks.

For the SVM model, the fitrsvm function in MATLAB was used with a Radial Basis Function (RBF) kernel. The RBF kernel was selected because it is commonly applied in regression problems where nonlinear relationships may exist between input variables and the target output. The model employed MATLAB's default parameter settings, including a box constraint of 1, epsilon value of 0.1, and automatic kernel scaling. These settings provide a balanced baseline configuration that allows the model to learn nonlinear patterns without introducing excessive complexity.

The Random Forest model was developed using the fitrensemble function with the Bagging method. Ensemble tree models such as Random Forest are often configured using a moderate number of learning cycles to stabilise predictions while maintaining computational efficiency. In this study, 100 learning cycles were used together with the standard tree learner provided in MATLAB.

Similarly, the XGBoost model was approximated using the Least-Squares Boosting (LSBoost) method available in the same function. Boosting models iteratively refine prediction errors, and a moderate learning rate of 0.1 with 100 boosting cycles is frequently adopted as a practical starting configuration in regression studies. This setup allows the model to gradually improve predictions while reducing the likelihood of excessive model variance.

3.5.2 Training and Validation Process

To maintain consistency across all forecasting models, a structured and standardised training and validation process was implemented. The dataset was divided into three subsets: 70% for training, 15% for validation, and 15% for testing. This three-

way partitioning approach is commonly adopted in machine learning studies to reduce overfitting and ensure that model evaluation is based on data not seen during training. For the SVM, Random Forest, and XGBoost models, this partitioning was carried out programmatically in MATLAB using the cvpartition function with fixed random seeds to ensure reproducibility. A comparable data split was configured for the ANN model using the built-in data division utility in MATLAB's Neural Net Fitting App. This ensured that all models were trained and tested on the same data partitions, allowing for fair performance comparisons.

During training, all models were supplied with the same input features: GDP, Energy Consumption, and Population, which were selected based on their demonstrated relevance in the previous part and their measurable influence on hydroelectric power generation. By keeping the input features consistent, the study ensured that observed performance differences were due to the models' learning capacities rather than differences in input data.

Model validation was conducted in parallel with training to monitor generalisation and guide early stopping decisions. For the ANN model, early stopping was automatically triggered when the validation error failed to improve over six consecutive iterations, as specified by the default setting in the Levenberg–Marquardt training algorithm. This regularisation technique helps prevent overfitting by retaining the network configuration that yields the lowest validation error. Details of the ANN's training and convergence behaviour are provided in Table 3.3 in Section 3.5.1.

For the SVM, Random Forest, and XGBoost models, similar performance monitoring procedures were implemented using custom MATLAB scripts. In each case, the validation set was used to periodically assess model error during training cycles. In the case of XGBoost, early stopping rounds were also incorporated to terminate boosting iterations once no significant improvement was detected in the validation loss, improving computational efficiency and avoiding overfitting.

To facilitate qualitative performance assessment, visual diagnostic tools were generated for each model. For the ANN, the Neural Net Fitting App automatically produced regression plots, learning curves, and error histograms. Equivalent visualisations were created for SVM, Random Forest, and XGBoost using MATLAB, including predicted-versus-actual scatter plots, residual error distributions, and MSE progression curves across training iterations. These plots provided visual insights into convergence behaviour, error structure, and model fit.

In terms of quantitative evaluation, each model's performance was assessed on both the validation and test sets using two key metrics: Mean Squared Error (MSE) and the R-value. MSE reflects the average squared difference between predicted and actual values, offering a direct measure of prediction error magnitude. The R-value, on the other hand, quantifies the strength of the linear correlation between predicted and actual outcomes, capturing the degree to which each model preserved trend behaviour.

By applying a uniform training and validation framework, the study minimised confounding variables related to data handling or experimental setup. This allowed for a clearer interpretation of how each model, ANN, SVM, Random Forest, and XGBoost, responded to the same set of socioeconomic predictors in forecasting hydroelectric power generation.

3.5.3 Performance Evaluation and Comparison

The performance of the ANN, SVM, Random Forest, and XGBoost models was evaluated using the same core performance metrics introduced earlier: Mean Squared Error (MSE) and the correlation coefficient (R-value). These two metrics were applied across the training, validation, and testing datasets to assess not only how well each model learned from the training data but also how effectively it generalised to unseen data. MSE provided a measure of the average squared difference between actual and predicted values, where lower values indicated higher accuracy. Meanwhile, the R-value offered an indication of the strength of the linear relationship between predicted and actual outputs, with values closer to 1 suggesting better alignment. Using these metrics uniformly across all models helped ensure that the evaluation remained consistent and that the results were directly comparable.

In addition to standard performance indicators, visual diagnostic tools were employed to interpret the predictive behaviour and reliability of each machine learning model. Regression plots were generated to illustrate the relationship between actual and predicted hydroelectric generation values. In these plots, a tighter clustering of data points along the 45-degree line was interpreted as a sign of stronger predictive alignment and minimal bias, whereas systematic deviations from this line suggested potential underfitting or overfitting.

Error histograms were also used to examine the distribution of residuals across all datasets. These plots display how frequently different error magnitudes occurred,

where the x-axis shows the error range and the height of each bar reflects the number of predictions falling into that range. The zero-error line serves as a reference for ideal predictions. Interpretation focused on identifying whether errors were concentrated near zero or spread widely across the graph. A narrow and centred error distribution indicated that the model made relatively small and consistent prediction errors. In contrast, wider spreads, skewed shapes, or multiple peaks in the distribution signalled instability, sensitivity to specific input conditions, or limitations in the model's ability to generalise to unseen data. This graphical insight complemented the numerical metrics and helped highlight behavioural differences between the models.

While these visualisations do not serve as definitive proof of model superiority, they complement the MSE and R-value metrics by offering a more qualitative perspective on model behaviour. Together, the numerical and graphical evaluations provide a more comprehensive view of model performance, allowing for the detection of subtle strengths or limitations that might not be evident from scalar metrics alone. Importantly, the purpose of this comparative analysis was not solely to identify the best-performing algorithm, but rather to gain insights into how each model responded to the same set of socioeconomic inputs. By observing these behaviours under a consistent experimental framework, the study was better positioned to draw cautious conclusions about the suitability of different machine learning methods for forecasting hydroelectric power generation in data-driven energy planning applications.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Introduction

In this study, a structured experimental procedure was employed to examine the predictive relationship between hydroelectric generation and key socioeconomic indicators, namely GDP, Energy Consumption, Population, and Inflation. The analysis began by developing ANN models under three distinct input configurations to evaluate the contribution of different predictor combinations. These models were assessed using five-fold cross-validation, with performance measured by MSE and R-values. Based on this evaluation, the most effective input set was identified and subsequently used to benchmark the ANN model against three additional machine learning algorithms: SVM, Random Forest, and XGBoost. To ensure a comprehensive evaluation, performance was also examined using learning curves, error histograms, and regression plots across all phases of training, validation, and testing.

The Results and Discussion section presents detailed findings from each stage of the experiment: data preprocessing and normalisation, ANN model development and variable testing, cross-validation analysis, and comparative evaluation across models. Emphasis is placed on identifying the most suitable input variables, understanding how each model generalises, and highlighting the strengths and limitations of different algorithmic approaches. These evaluations contribute to a data-driven understanding of how socioeconomic trends can support more accurate forecasting in the energy sector, particularly in the context of hydroelectric power generation.

This chapter is organised into several subsections. Section 4.2 describes the data preprocessing procedures, including data preparation and normalisation prior to model development. Section 4.3 presents the cross-validation analysis used to evaluate the performance of different input configurations for the ANN models. Section 4.4 discusses the performance of the selected ANN models in forecasting hydroelectric generation. Section 4.5 provides a comparative evaluation of multiple machine learning models, namely ANN, SVM, Random Forest, and XGBoost, to assess their relative

predictive capabilities. Finally, Section 4.6 summarises the key findings of the chapter and highlights the main insights derived from the analysis.

4.2 Data Preprocessing

In preparation for model training, the dataset underwent normalisation tailored to the specific needs of each machine learning algorithm. Table 4.1 summarises the normalisation details for each variable.

Table 4.1
Summary of Min-Max and Z-Score Normalisation

	Normalisation		**Z-Score Normalisation**	
	Min	**Max**	**Mean**	**Std**
GDP	53,308.00	1,548,898.00	518,284.5238	469,399.4716
Energy Consumption	747.00	13,647.00	5,716.9048	4,259.8587
Population	13,879.00	32,584.00	23,387.5714	6,020.8630
Inflation	-0.01	0.10	-	-
Hydroelectric Generation	120.00	2,676.00	-	-

For the ANN model, Min-Max normalisation was selected. This method is especially effective in neural network training because it minimises the risk of gradient saturation during backpropagation, which can hinder learning when input values vary widely in magnitude. Key input variables such as GDP, Energy Consumption, Population, and Inflation, along with the output variable, Hydroelectric Generation, were normalised using this approach. For example, GDP values ranged from 53,308.00 to 1,548,898.00, while Energy Consumption ranged from 747.00 to 13,647.00. Inflation values remained within a narrow band between -0.01 and 0.10. The output variable, Hydroelectric Generation, was also normalised to support smoother convergence during the ANN training process. This preprocessing ensured that all variables contributed equally to the model's learning, preventing domination by any single high-magnitude feature.

In contrast, the SVM, Random Forest, and XGBoost models were normalised using Z-score. This technique transforms the variables to have a mean of zero and a standard deviation of one, making it suitable for algorithms that rely on distance-based metrics or internal feature splitting, such as SVMs and tree-based models. In this case, GDP, Energy Consumption, and Population were standardised, while Inflation was omitted from the z-score normalisation process, as it was not part of the selected input configuration for these models. Hydroelectric Generation, serving as the output variable, was also left unstandardized to retain its original scale. The standard deviations observed for GDP, Energy Consumption, and Population were 469,399.47, 4,259.86, and 6,020.86, respectively, as mentioned in Table 4.1, reflecting moderate to high variability. By aligning the input features through normalisation, the models were better equipped to process the socioeconomic data effectively without being biased by scale differences. These preprocessing decisions helped ensure each model received well-prepared data appropriate to its architecture, promoting stable learning and improving overall reliability.

In the context of model interpretability and performance, the choice of normalisation technique is crucial. Neural networks are particularly sensitive to input magnitudes because their optimisation process depends on gradients computed during backpropagation. Without normalisation, features with larger values can dominate learning, leading to poor convergence and bias in weight updates. Conversely, models like SVMs rely on kernel functions that compute distances between data points, making consistent scaling across features vital. For tree-based models such as Random Forest and XGBoost, normalisation is less critical for accuracy but still beneficial for performance stability, especially when combining categorical and continuous inputs. If normalisation were omitted or misapplied, it could result in skewed decision boundaries, misrepresented feature importance, and increased training times. Therefore, the tailored normalisation approach used in this study not only enhanced model learning but also contributed to fair comparative analysis across algorithm types.

4.3 Cross-Validation

Before advancing to full-scale training, a five-fold cross-validation was conducted using the Neural Net Fitting App in MATLAB to evaluate how different

combinations of input variables performed in forecasting hydroelectric generation. This procedure involved partitioning the dataset into five equal parts, training the model on four subsets while testing it on the remaining one. This process was repeated until each subset was the test set once. Through this approach, every observation in the dataset contributed to both training and validation at different stages, allowing the model performance to be assessed more reliably despite the relatively limited size of the dataset.

The cross-validation focused on three input scenarios: Model A, GDP and Energy Consumption, Model B, GDP, Energy Consumption, and Population, and Model C, GDP, Energy Consumption, and Inflation. This preliminary phase was not aimed at final model selection but instead served as an exploratory step to identify which variable combinations consistently yielded better prediction accuracy and generalisation. The dataset was carefully structured to preserve the chronological alignment of input features across folds, ensuring reliable performance assessment without temporal distortion. Using cross-validation in this stage also helped reduce the risk that the model performance would depend too heavily on a single data partition.

The results of the cross-validation, presented in Table 4.2, revealed noticeable performance differences among the three configurations. Model B, which included Population in addition to GDP and Energy Consumption, achieved the lowest MSE and the highest R-value, indicating improved fit and predictive strength.

Table 4.2
Cross-Validation Evaluation of Input Variable Sets

Input Combination	**MSE (5-Fold Average)**	**R-Value (5-Fold Average)**
(a) GDP + Energy Consumption	2.8040×10^{6}	0.5979
(b) GDP + Energy Consumption + Population	1.8955×10^{5}	0.7731
(c) GDP + Energy Consumption + Inflation	5.8547×10^{5}	0.4279

Specifically, it recorded an average MSE of 1.8955×10^{5} and an R-value of 0.7731, outperforming the other configurations. The R-value of approximately 0.77

suggests that the model was able to capture a moderate relationship between the selected input variables and hydroelectric generation during cross-validation. Model A, the baseline with only GDP and Energy Consumption, produced a higher MSE and a weaker correlation, while Model C, where Inflation replaced Population, also showed lower correlation values and greater variability in prediction accuracy. These observations suggest that Population contributes useful information when modelling hydroelectric generation together with economic and energy indicators. In contrast, the inclusion of Inflation did not improve the predictive relationship during this preliminary evaluation.

By applying a structured and repeatable validation method, the study was able to make more informed decisions regarding input selection. This step helped ensure that the variable combination used for the subsequent ANN training was supported by consistent performance during cross-validation rather than relying on a single training outcome.

4.4 Performance of ANN Models for Energy-Economy Forecasting

Following the cross-validation phase, which provided data-driven insights into the suitability of various input variable combinations, the study advanced to the core modelling stage. This phase centred on the construction and evaluation of Artificial Intelligence (AI)-based models, specifically ANN, to examine the predictive relationships between socioeconomic indicators and hydroelectric power generation in Malaysia. The objective was not only to forecast energy output but also to gain a deeper understanding of the roles that different economic and demographic factors play in shaping long-term energy production patterns.

To achieve this, the modelling approach was systematically divided into three distinct scenarios, each representing a different combination of input variables. These configurations were chosen to reflect plausible real-world relationships that might exist between macroeconomic trends and renewable energy output, rather than relying solely on statistical feature ranking methods. This approach ensured that the analysis remained grounded in contextual relevance while allowing for interpretability of each model's behaviour.

Model A, employed a baseline configuration using GDP and Energy Consumption as the input features. This model aimed to capture the core dynamics of economic activity and energy demand, both of which are typically regarded as primary drivers of electricity generation.

Model B, built upon the first configuration by incorporating Population as a third input variable. This addition was intended to account for demographic growth and its potential impact on energy requirements. By including Population, the model was expected to reflect broader socioeconomic pressures on energy systems, particularly as urbanisation and population expansion may correlate with increased infrastructure and energy use.

Model C, introduced a macroeconomic dimension by replacing Population with Inflation Rate. This configuration sought to explore whether price stability and broader economic volatility might influence hydroelectric generation, possibly through investment trends, operating costs, or changes in energy pricing. Inflation was considered a less direct but potentially insightful variable for understanding the external financial pressures faced by the energy sector.

Each of these modelling scenarios enabled a structured and comparative evaluation of the ANN's sensitivity to different types of inputs. By holding the model architecture and training procedure constant while varying only the input features, the study ensured that observed differences in performance could be attributed primarily to the predictor configurations themselves. This design allowed for a focused assessment of how specific socioeconomic variables contribute to the predictive capability and generalisation strength of AI-based forecasting models.

In summary, this phase of the study established a controlled experimental framework for investigating the energy–economy nexus through machine learning. It highlighted the importance of careful input selection and offered insights into which indicators hold the greatest potential for improving the accuracy and robustness of hydroelectric power forecasting in a real-world context.

4.4.1 Model Performance

Table 4.3 summarises the performance of the three model configurations across training, validation, and testing phases. The table presents the Mean Squared Error (MSE) and correlation coefficient (R-value) for each configuration, providing a basis

for comparing predictive accuracy and generalisation capability. These results highlight how variations in input variables influence model performance, allowing for the identification of the most reliable configuration.

Table 4.3
Model Performance

		Training	Validation	Test
Model A	**MSE**	7.4874e+03	1.7485e+04	1.1441e+04
	R	0.9918	0.9678	0.9923
Model B	**MSE**	2.5077e+03	1.4350e+04	6.4307e+03
	R	0.9970	0.9532	0.9969
Model C	**MSE**	3.1628e+03	4.6403e+03	1.8151e+04
	R	0.9962	0.9810	0.9945

The model in Model A, as shown in Table 4.3, using GDP and Energy Consumption as inputs, demonstrated reasonable predictive capabilities during validation and testing. With a validation MSE of 1.7485×10^4 and an R-value of 0.9678, the model showed a strong correlation between predicted and actual outcomes. Testing results were similarly strong, with an MSE of 1.1441×10^4 and an R-value of 0.9923. However, the observed gap between validation and testing suggests mild overfitting, indicating that while the model effectively captured relationships in the data, its ability to generalise to unseen data could be further improved. These results underscore the importance of balancing model complexity with generalisation.

Model B in Table 4.3, which incorporates Population alongside GDP and Energy Consumption, significantly improved the model's performance. The validation MSE was reduced to 1.4350×10^4, with an R-value of 0.9532, demonstrating better generalisation compared to Model A. The test results further confirmed this improvement, with an MSE of 6.4307×10^3 and an R-value of 0.9969, the highest among all configurations. Adding Population contributed to a more comprehensive understanding of the relationships within the data, reducing overfitting while ensuring reliable predictive accuracy. These findings suggest that incorporating relevant socioeconomic factors, such as Population, can enhance the predictive reliability of models for hydroelectric generation forecasting.

Model C, as shown in Table 4.3, replacing Population with Inflation Rates introduced variability, leading to a decline in overall performance. The validation MSE increased to 4.6403×10^{3}, with an R-value of 0.9810, reflecting weaker generalisation compared to Model B. The test results followed a similar trend, with an MSE of 1.8151×10^{4} and an R-value of 0.9945. While the model still performed reasonably well on unseen data, the higher errors suggest that Inflation Rates might introduce noise due to their inherent variability or weaker correlation with hydroelectric generation. These results highlight the importance of carefully evaluating the relevance and stability of input features when designing models.

A closer inspection of the generalisation gaps, the difference between training and test performance, provides additional insight into model behaviour. In Model A, the MSE increased from 7.4874×10^{3} in training to 1.1441×10^{4} in testing, reflecting a generalisation gap of approximately 53%. This moderate gap suggests that while the model learned underlying patterns, it also partially captured noise or structures that did not generalise well. In contrast, Model B, demonstrated a larger generalisation gap, with MSE rising from 2.5077×10^{3} to 6.4307×10^{3}, an increase of about 156%. Despite the high percentage, this gap is acceptable because the absolute testing error remains significantly lower than in Model A, indicating that the model achieves better real-world prediction accuracy even with a larger proportional difference. Model C, by comparison, experienced a sharp rise in error from 3.1628×10^{3} to 1.8151×10^{4}, corresponding to a 474% generalisation gap. Unlike Model B, this high gap is not acceptable because both the percentage increase and the absolute test error are considerably larger, suggesting severe overfitting and poor generalisation to unseen data. These results highlight that a low generalisation gap alone does not guarantee strong performance; rather, a balance between error magnitude and generalisation is critical. Based on both error levels and generalisation behaviour, Model B stands out as the most effective model configuration.

4.4.2 Error Analysis and Model Robustness

Figures 4.1 to 4.3 display the error histograms for the three model configurations, providing a visual representation of the distribution and magnitude of prediction residuals across training, validation, and testing datasets. These plots facilitate an evaluation of model precision by showing how frequently errors fall near

zero and where larger deviations occur. The interpretation of these error histograms follows the approach described in Section 3.4.2, where the shape, concentration, and spread of errors relative to the zero-error line provide insight into each model's predictive accuracy and generalisation capability.

In general, an effective predictive model tends to produce an error distribution that is centred near the zero-error line, indicating that predicted values are close to the actual observations. A strong concentration of bars around zero suggests high prediction accuracy, whereas a wider spread of errors indicates larger deviations between predicted and actual outputs. Therefore, examining the spread and symmetry of the histogram provides insight into the stability and reliability of the model's predictions.

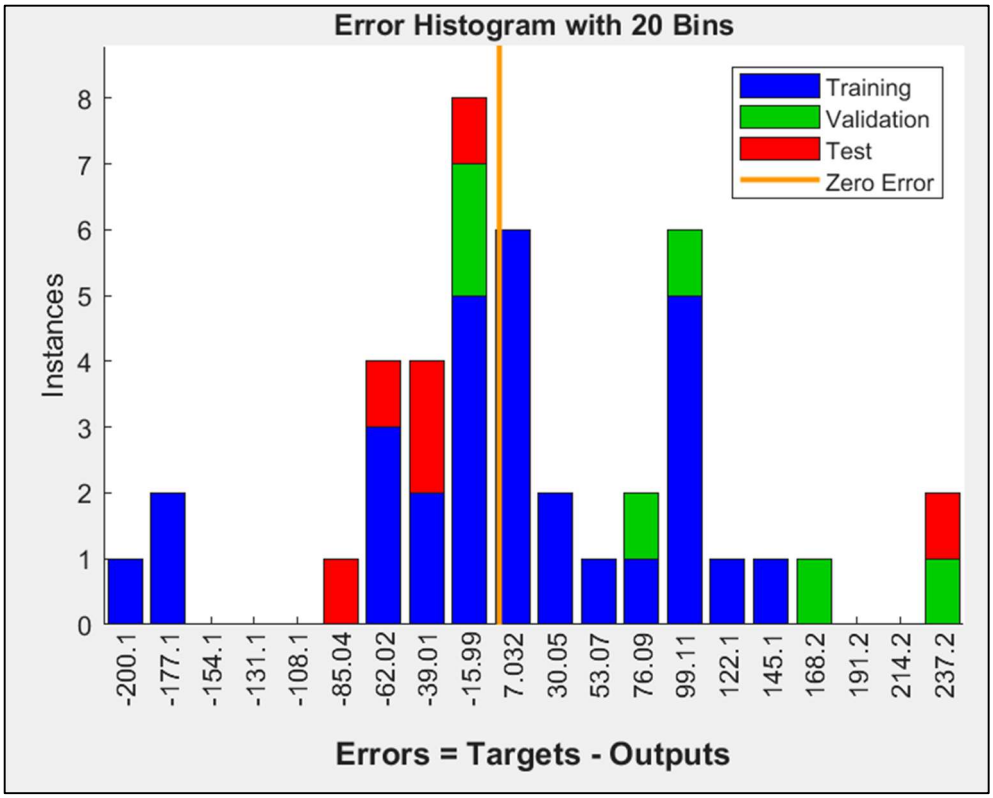

Figure 4.1 Error Histogram (Model A)

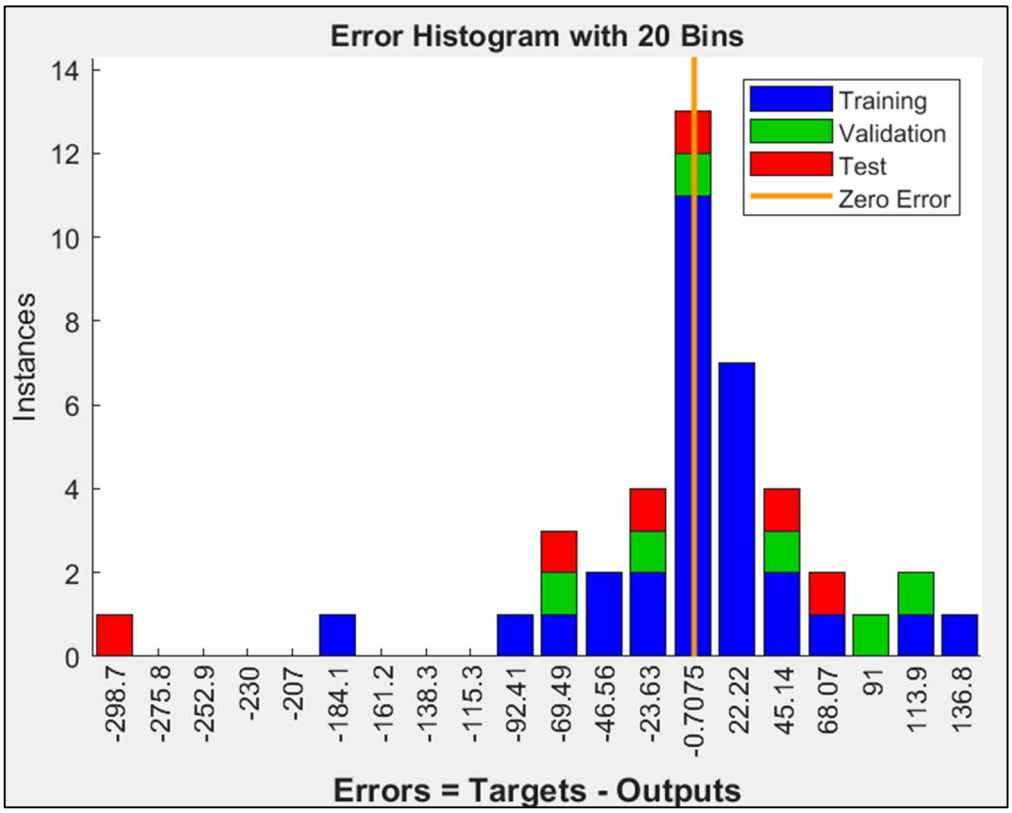

Figure 4.2 Error Histogram (Model B)

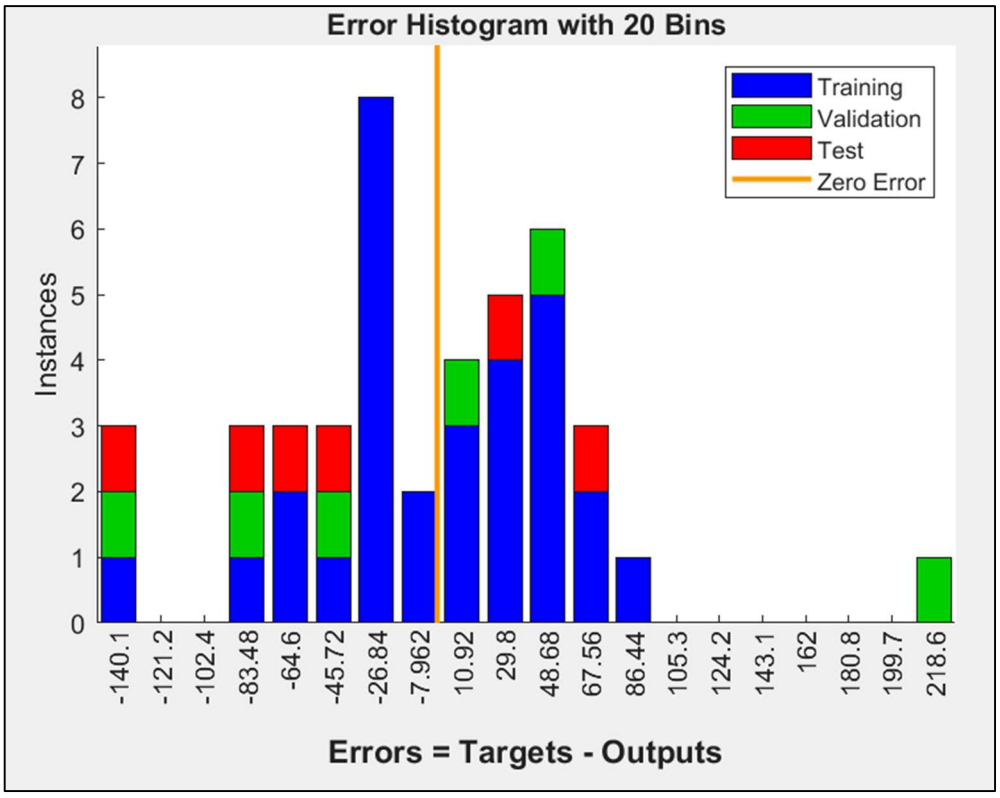

Figure 4.3 Error Histogram (Model C)

The error histogram for Model A in Figure 4.1, presents a moderately wide distribution, particularly within the validation and testing datasets. While most training errors are concentrated between -50 and 50, unseen data errors extend up to approximately -200 on the negative side and +237 on the positive side, indicating a broader deviation range. This dispersion suggests that the model was able to learn underlying patterns in the training set but exhibited limited robustness when predicting unseen data. The wider spread and lower bar concentration around the zero-error line reflect weaker generalisation ability and the presence of significant outliers.

The histogram also shows a slightly irregular distribution rather than a perfectly smooth pattern. This behaviour is likely influenced by the relatively small dataset used in this study, particularly the limited number of validation and testing observations. When the number of samples is small, the distribution of residuals may appear uneven, as individual prediction errors have a stronger influence on the overall histogram shape.

When Population was included as an additional input feature, the error histogram in Model B, as illustrated in Figure 4.2, became more compact compared to Model A. Most errors for training, validation, and testing datasets are concentrated between -30 and 90, though a few large outliers extend to -140 and +219. The increased density of bars around the zero-error line reflects improved predictive accuracy and better generalisation performance. This improvement suggests that the inclusion of Population provided additional explanatory power, capturing demographic influences on hydroelectric generation and enhancing the stability of the forecasting model.

Compared with Model A, the errors in Model B are more tightly grouped near the zero-error line, showing that the predictions deviate less from the actual values. Although the distribution is not perfectly symmetrical, the reduced spread of residuals suggests that the model achieved more consistent predictive behaviour across the different datasets.

Figure 4.3, which shows an error histogram for Model C, where the Inflation Rate replaced the Population as an input variable, the error histogram shows a noticeably wider spread than the previous configurations. Validation and testing errors extend up to approximately -300 on the negative side and +137 on the positive side, creating the broadest error range among the three models. Although training errors remain largely concentrated near zero, the presence of several extreme outliers indicates weaker generalisation and reduced reliability on unseen data. This dispersed distribution reflects instability in the model's predictions, potentially caused by the high

volatility and noise introduced by the Inflation Rate as an input feature. The wider dispersion of errors show that the model experienced greater difficulty maintaining stable prediction accuracy. This outcome may also be influenced by the limited number of observations in the dataset, which can amplify the impact of individual prediction errors and produce a more irregular distribution.

Nevertheless, when interpreted together with the performance metrics and the K-fold cross-validation results presented earlier, the findings still provide meaningful insight into the predictive capability of the models.

4.4.3 Insight from Learning Curve and Regression Plot

To provide a clearer understanding of each model's predictive behaviour, Figures 4.4 to 4.6 display the corresponding model performance metrics, and Figures 4.7 to 4.9 present the regression plots for each ANN model. Model A, B and C correspond to the three input scenarios: GDP with Energy Consumption, the addition of Population, and the substitution with Inflation Rate, respectively.

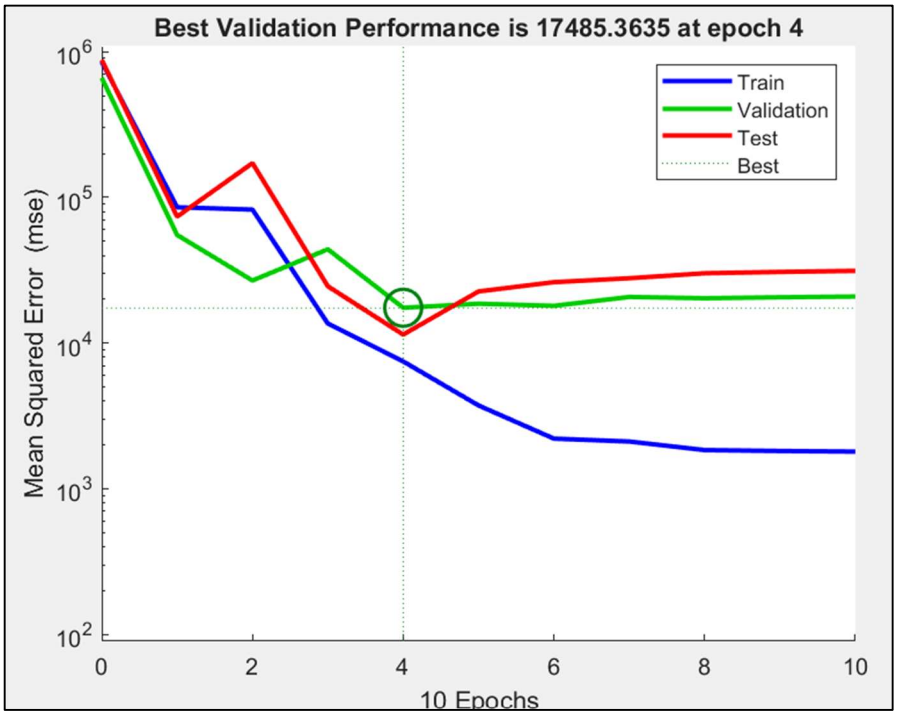

Figure 4.4 Model Performance (Model A)

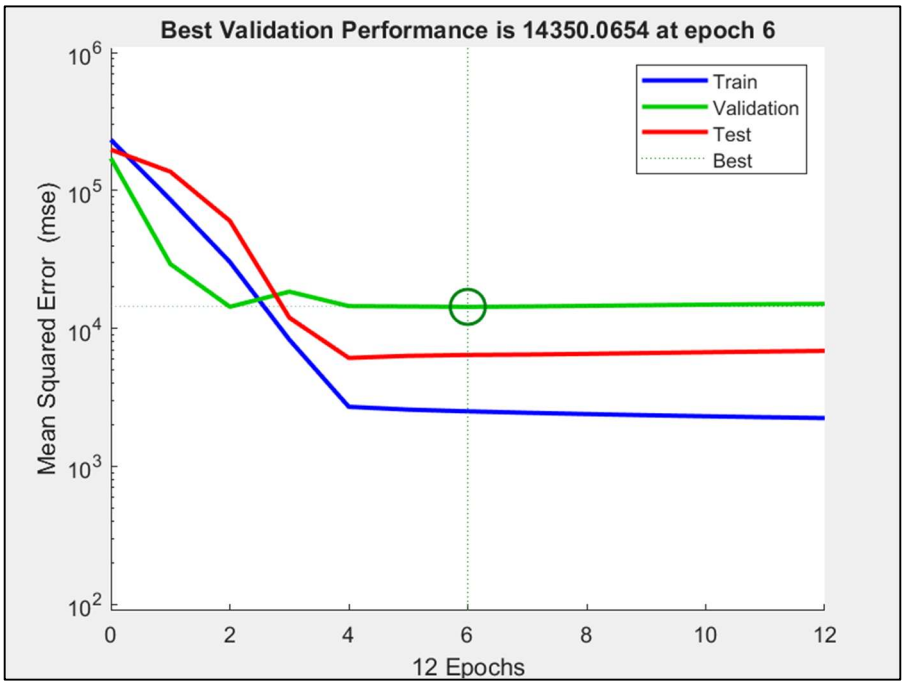

Figure 4.5 Model Performance (Model B)

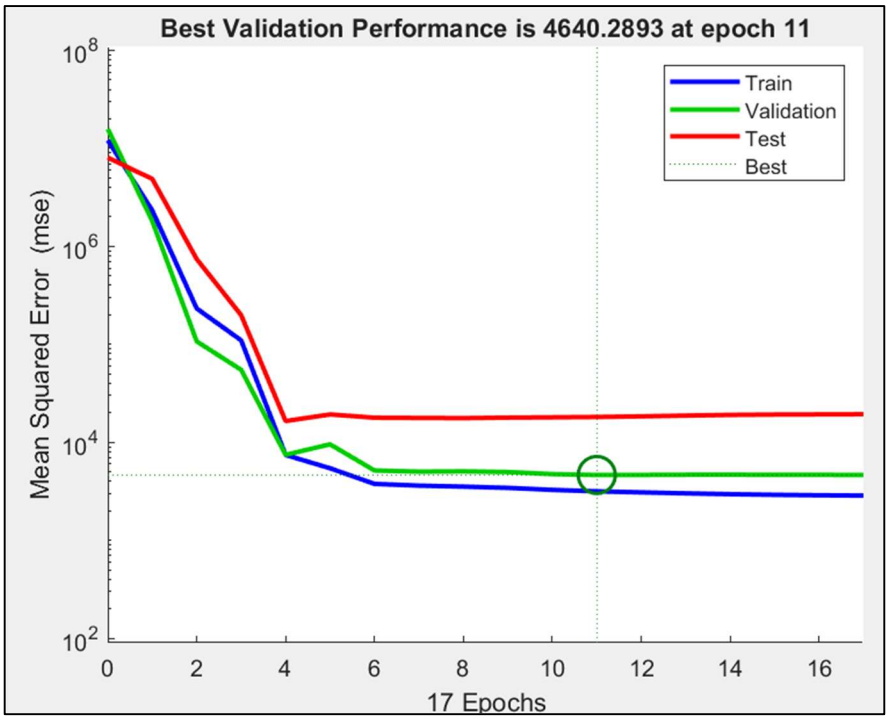

Figure 4.6 Model Performance (Model C)

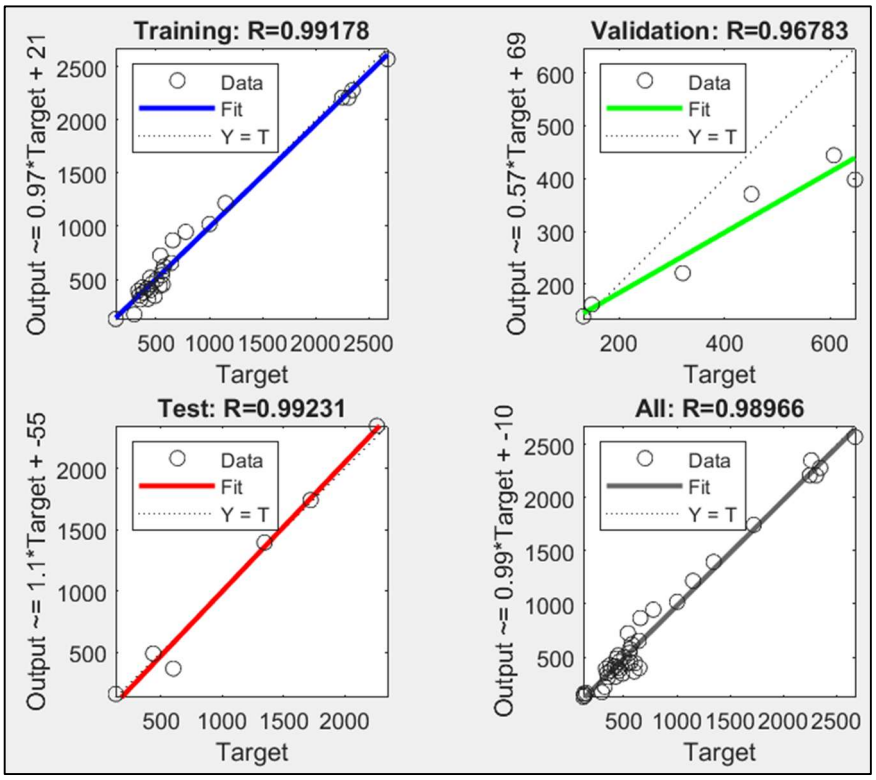

Figure 4.7 Regression Plot (Model A)

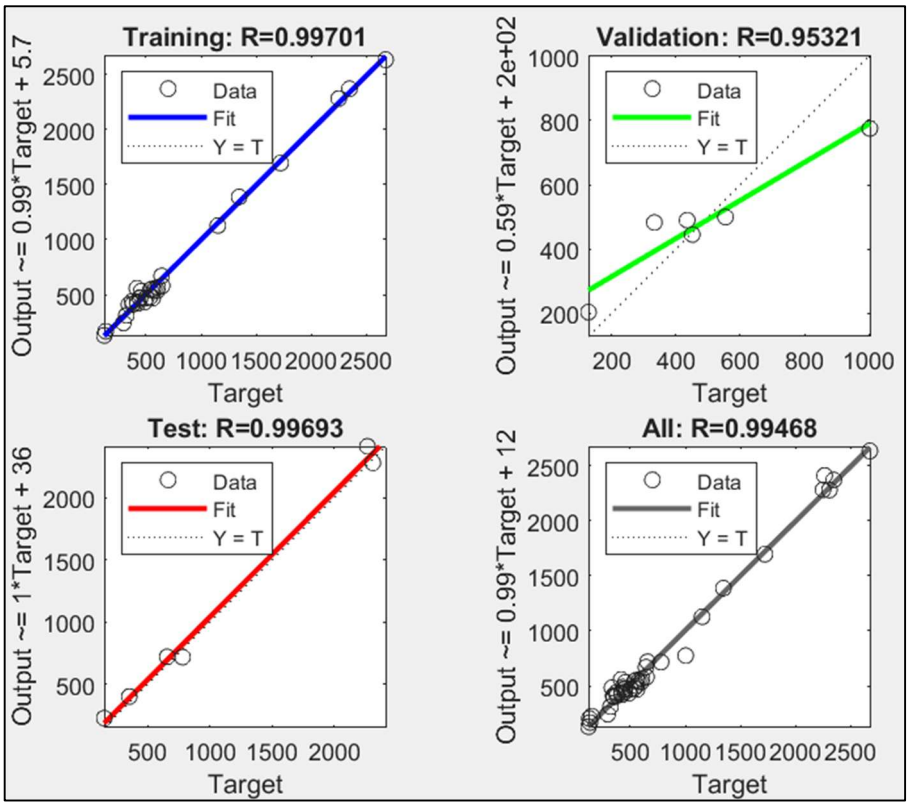

Figure 4.8 Regression Plot (Model B)

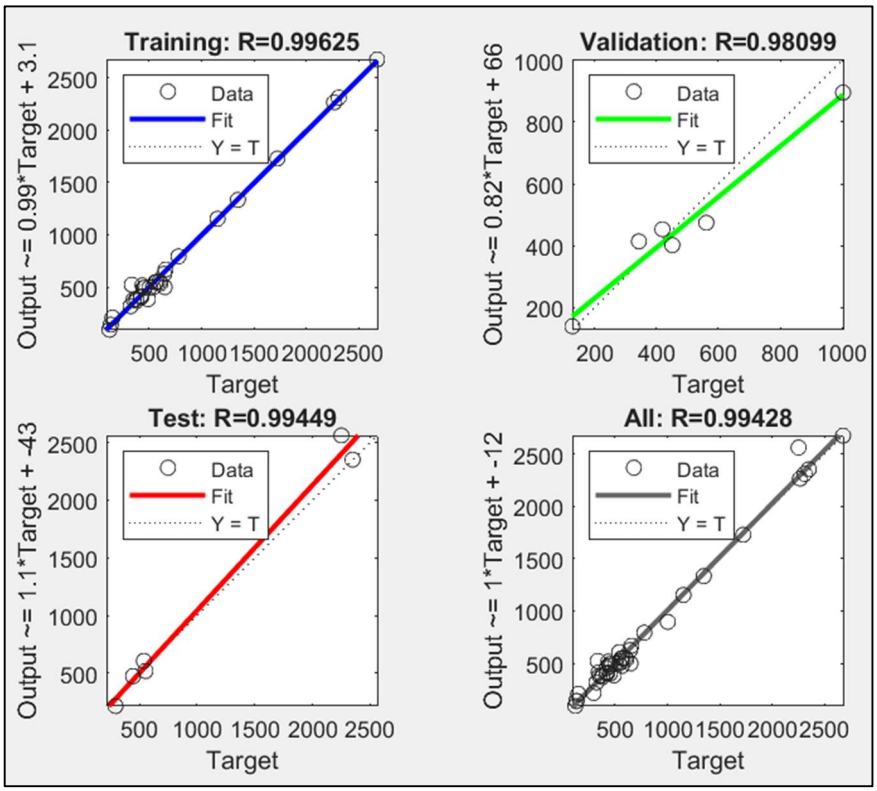

Figure 4.9 Regression Plot (Model C)

The learning curve presented in Figure 4.4 indicates that the model achieves its optimal validation performance at epoch 4, with a minimum MSE of 1.7485×10^4. This early convergence suggests that the model was able to quickly capture dominant patterns in the training data. However, the consistently higher validation MSE compared to the test MSE (1.1441×10^4) reveals a moderate degree of overfitting. The model adapts well to the training data but exhibits difficulty in generalising to unseen data, likely due to the limited complexity of the input variables, GDP and Energy Consumption. The regression plot in Figure 4.7 further supports this observation. While there is a generally positive linear trend in the predictions, the validation set demonstrates more noticeable scatter from the perfect-fit line (R-value: 0.9678), indicating variability in prediction accuracy. These patterns suggest that although the model captures the overall structure of the data, its predictive consistency is limited, especially under conditions not represented in the training phase.

In contrast, Figure 4.5 shows the learning curve of Model B, which reveals enhanced model performance and stronger generalisation. The model attains its best validation result at epoch 6, achieving an MSE of 1.4350×10^4. Notably, the gap between validation and test MSE is narrow (test MSE: 1.1541×10^4), implying a more

balanced learning process with minimal overfitting. This improved alignment indicates that the model retains its predictive capacity even when applied to new, unseen data. The regression plots in Figure 4.9, exhibit tighter clustering of predicted values along the ideal diagonal line, with R-values of 0.9532 (validation) and 0.9969 (test), reflecting a strong linear relationship between predicted and actual outputs. The addition of Population as an input variable appears to enrich the model's capacity to detect nuanced patterns related to demographic trends. Compared to Model A, this configuration enhances both accuracy and stability, suggesting that the inclusion of Population improves the explanatory power of the input set and contributes to greater robustness in forecasting hydroelectric generation.

Figure 4.6 presents that the model for Model C requires 11 epochs to reach its best validation performance, with an MSE of 4.6430×10^3, indicating slower convergence and potentially increased training complexity. While the test R-value remains relatively high at 0.9945, the validation R-value of 0.9810 is accompanied by more dispersion in the regression plot, particularly at the extremes of the target range. Figure 4.6 shows that predictions deviate more from the ideal line for higher values, indicating weaker predictive reliability in such cases. Additionally, the learning curve in Figure 4.9 exhibits a persistent gap between validation and test MSEs, suggesting that the model struggles to generalise well despite achieving low error on specific datasets. This pattern may be attributed to the volatility and less stable nature of Inflation Rates as a macroeconomic indicator, which can introduce noise into the learning process. As a result, while the model performs acceptably, its robustness and overall reliability appear diminished compared to the configuration in Model B, reinforcing the importance of input variable stability in model development. While ANN demonstrated strong performance in modelling the selected socioeconomic indicators, further analysis is required to examine how alternative machine learning models perform under the same input conditions. This comparison is presented in the following section.

4.5 Comparative Performance of Machine Learning Models for Forecasting Hydroelectric Generation

Building upon the findings obtained from the Artificial Neural Network (ANN) modelling phase, this section extends the analysis by conducting a comparative

evaluation of multiple machine learning models. While ANN was initially employed to investigate the relationship between socioeconomic indicators and hydroelectric power generation, further validation using additional modelling techniques was undertaken to strengthen the reliability and robustness of the predictive analysis.

In addition to ANN, three widely used machine learning algorithms, Support Vector Machine (SVM), Random Forest, and XGBoost, were implemented in this study. These models were selected due to their established effectiveness in regression-based forecasting tasks and their distinct learning mechanisms. SVM is capable of modelling complex non-linear relationships through kernel-based transformations, Random Forest improves predictive stability through ensemble averaging of multiple decision trees, while XGBoost utilises gradient boosting to iteratively refine predictions and reduce model error.

To ensure a fair comparison, all models were trained using the same input variables, GDP, Energy Consumption, and Population, and evaluated under identical experimental conditions, including data partitioning and performance metrics. This standardised approach allows differences in model performance to be attributed primarily to the modelling techniques themselves rather than variations in data preprocessing or experimental design. Consequently, the comparative analysis provides insight into how different machine learning approaches respond to the same socioeconomic predictors when forecasting hydroelectric power generation.

4.5.1 Performance Evaluation of Machine Learning

This subsection presents a comparative evaluation of the four machine learning models, ANN, SVM, Random Forest, and XGBoost, based on their performance across the training, validation, and testing datasets. The assessment relies on two standard metrics: Mean Squared Error (MSE), which measures average prediction error magnitude, and the correlation coefficient (R-value), which quantifies the strength of the linear relationship between actual and predicted values. A detailed summary of these metrics is presented in Table 4.4.

Table 4.4
Performance Evaluation of Each Machine Learning

Models		Training	Validation	Test
ANN	MSE	5.0355e+03	9.3132e+03	1.1541e+04
	R	0.9941	0.9788	0.9962
SVM	MSE	1.7091e+05	1.6818e+05	1.3031e+05
	R	0.9703	0.9865	0.9250
Random Forest	MSE	1.0423e+05	7.6374e+04	6.2457e+04
	R	0.9158	0.9453	0.9438
XGBoost	MSE	1.2657e+05	1.1475e+05	5.8948e+04
	R	0.8487	0.8952	0.9196

Among the models assessed, the ANN achieved the best overall performance across the evaluation phases. With the lowest MSE values in both training (5.0355×10^3) and testing (1.1541×10^4), and consistently high R-values close to 1.0, ANN demonstrated strong predictive accuracy compared with the other models. Its validation R-value of 0.9788, though slightly lower than the training and testing values, suggests that the model maintained stable predictive behaviour when applied to unseen data.

The Random Forest model delivered moderate performance across the evaluation phases. Its testing MSE of 6.2457×10^4, while higher than ANN, was substantially lower than the testing error produced by SVM, showing comparatively better predictive accuracy. The correlation coefficients for Random Forest ranged from 0.9158 to 0.9438, reflecting a reasonably strong relationship between predicted and actual values. Overall, these results suggest that Random Forest was able to capture general patterns in the dataset, although with less precision than ANN.

In comparison, the SVM model recorded comparatively higher prediction errors. Its training MSE of 1.7091×10^5 was the largest among the evaluated models. Although the validation R-value was relatively high (0.9865), the drop in testing correlation to 0.9250 suggests some difficulty in maintaining consistent predictive performance on unseen data. This behaviour may reflect limitations in capturing the underlying structure of the dataset.

XGBoost demonstrated mixed results across the evaluation phases. It achieved a testing MSE of 5.8948×10^{4}, which was slightly lower than that of Random Forest, showing competitive predictive performance during testing. However, its comparatively high training error (1.2657×10^{5}) and lower training correlation (0.8487) suggest challenges during the learning stage. This inconsistency between training and testing behaviour may reflect sensitivity to parameter settings or the characteristics of the dataset.

In addition to the standard evaluation metrics, the generalisation gap, defined as the percentage change between training and testing MSE, was examined to provide further insight into model stability. The ANN model showed a generalisation gap of approximately 156%, which, although relatively large, remained acceptable considering its low testing error and strong correlation values. Random Forest exhibited a larger gap of approximately 235%, suggesting greater variation between training and testing performance. In contrast, SVM and XGBoost produced negative gaps of −23.7% and −53.4%, respectively. Such behaviour may occur when the models do not fully capture the structure of the training data or when regularisation influences the learning process.

Overall, the comparative evaluation suggests that ANN provided the most consistent predictive behaviour for forecasting hydroelectric power generation using the selected socioeconomic indicators. Random Forest and XGBoost demonstrated moderate predictive capability, while SVM showed comparatively higher variability in prediction performance. These findings highlight the importance of selecting suitable machine learning techniques when modelling complex relationships in energy forecasting applications.

4.5.2 Error Distribution and Model Robustness

This subsection examines the distribution of prediction errors as a means of evaluating the consistency, accuracy, and generalisation ability of each machine learning model. Error histograms, illustrated in Figure 4.10 for ANN, figure 4.11 for SVM, Figure 4.12 for Random Forest, and Figure 4.13 for XGBoost, provide a visual interpretation of how prediction residuals are distributed across the training, validation, and testing datasets. These distributions are instrumental in identifying outliers, assessing the degree of overfitting, and understanding how reliably each model captures

the underlying relationships within the data. These figures should be interpreted in accordance with the guidelines outlined in Section 3.5.3.

It should be noted that the error distributions observed in this study do not form a perfectly symmetrical pattern. This is primarily due to the relatively small dataset used in the experiment, particularly the limited number of testing observations. With fewer samples, the residual distribution may appear uneven because individual prediction errors contribute more strongly to the overall histogram shape. Despite this limitation, the consistency observed in the evaluation metrics and the K-fold validation results suggests that the models still capture the general relationship between the input variables and hydroelectric generation.

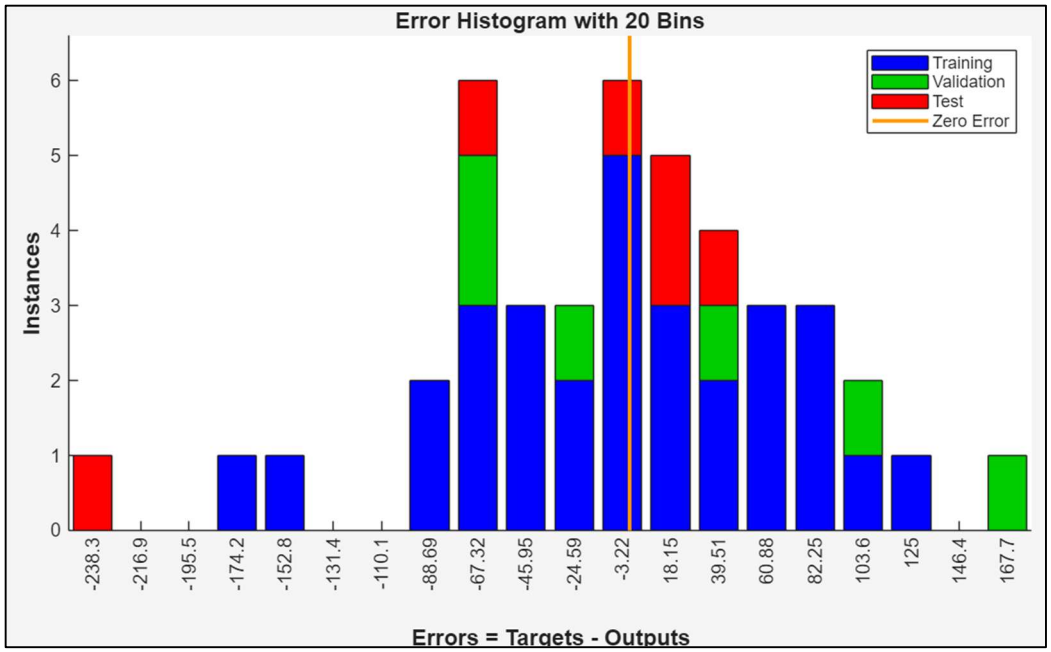

Figure 4.10 Error Histogram for ANN

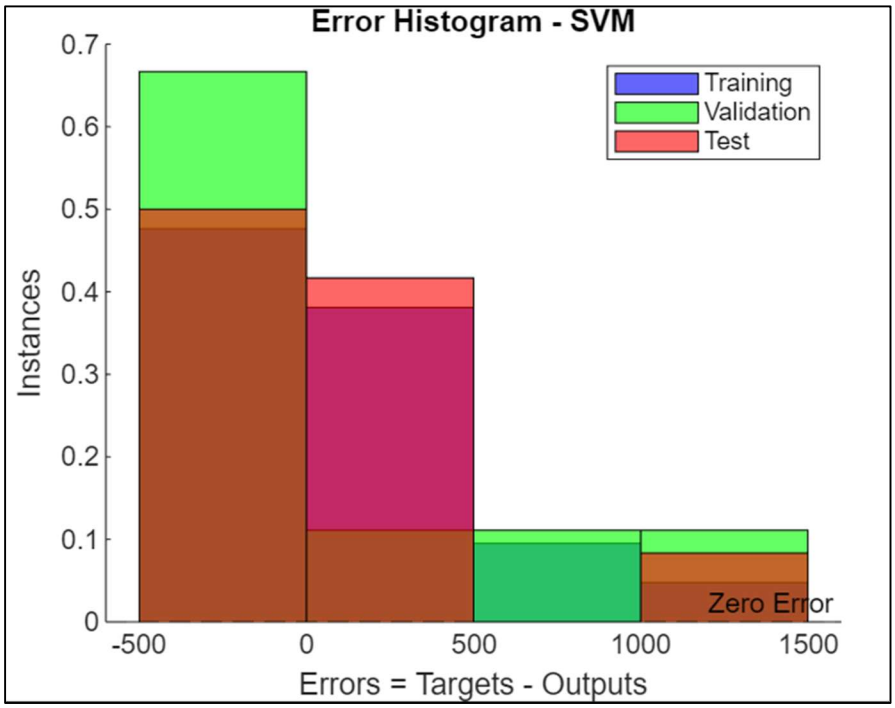

Figure 4.11 Error Histogram for SVM

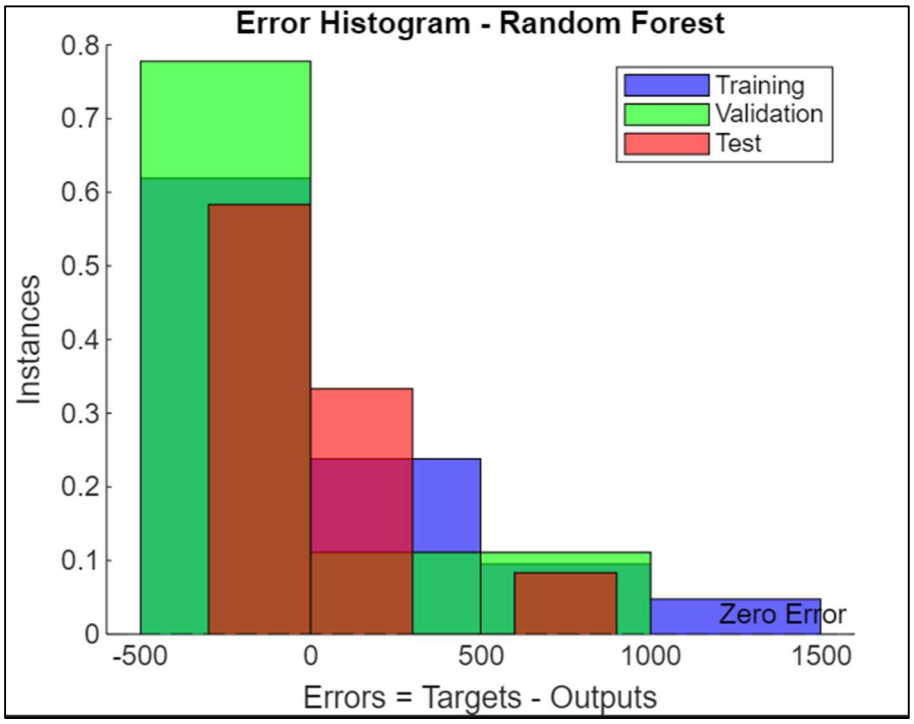

Figure 4.12 Error Histogram for Random Forest

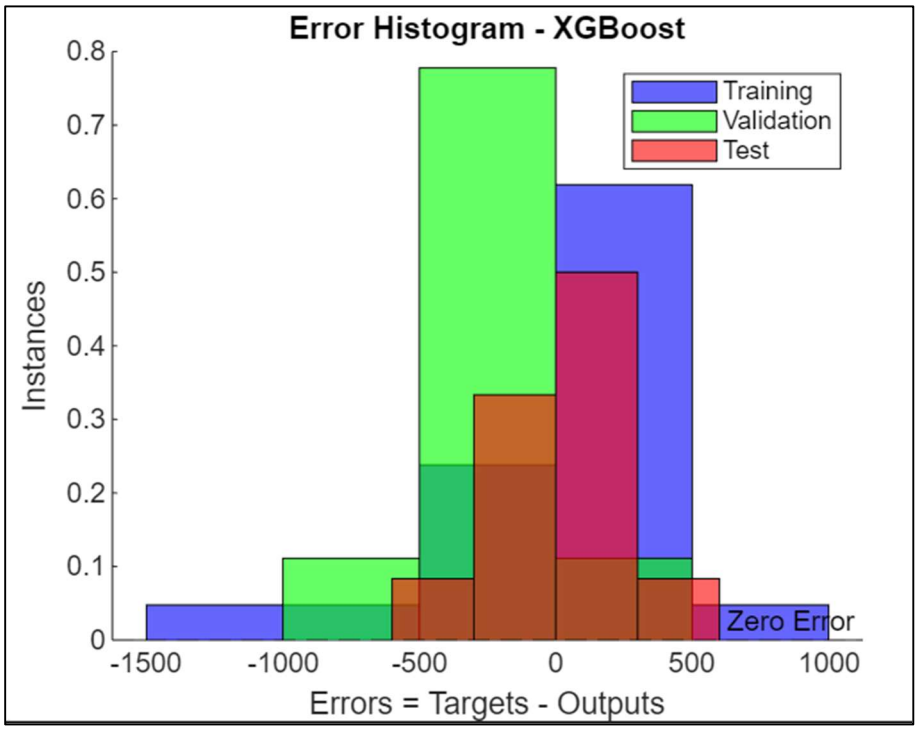

Figure 4.13 Error Histogram for XGBoost

As shown in Figure 4.10, the ANN exhibited the most compact and symmetrical error distribution. The majority of prediction errors were confined within a narrow range of approximately -90 to 100, with a pronounced peak near the zero-error line. This indicates a high level of predictive accuracy, suggesting that the ANN model consistently approximated actual values with minimal deviation. Although the visual distribution is affected by the finite sample size (n=42), the high density of residuals at the zero-error bin demonstrates a central tendency that approximates a normal distribution, confirming that the model is unbiased. Furthermore, the even distribution across training, validation, and test subsets supports the model's capacity to generalise well without overfitting.

In contrast, the SVM model (Figure 4.11) displayed the broadest error distribution, with residuals extending up to ±1500 and clustering around ±500. This wide dispersion suggests a higher level of variance in prediction outcomes and points to a reduced ability to learn complex patterns accurately. The presence of multiple peaks and a less centred distribution also imply potential issues with overfitting or model instability, particularly in extrapolating from training to unseen data.

The Random Forest error histogram (Figure 4.12) demonstrated moderate error concentration, with most residuals distributed between -500 and 500. While its

performance was more stable than SVM, the histogram revealed a longer tail and scattered large errors, especially in the test dataset. These characteristics suggest that although Random Forest was better at capturing general patterns than SVM, it occasionally produced large deviations that undermine its reliability.

The XGBoost model (Figure 4.13) performed better than both SVM and Random Forest. Its error distribution was more concentrated around zero, although not as sharply peaked as ANN. The model maintained reasonable consistency across data partitions but showed a slightly broader spread than ANN, indicating some fluctuations in prediction precision. Despite this, XGBoost showed fewer extreme outliers compared to SVM and Random Forest, reflecting a more balanced learning process and moderate robustness.

Taken together, the error histograms provide valuable insights into each model's strengths and limitations. ANN demonstrated the most consistent predictive behaviour among the evaluated models, offering minimal error spread and consistent generalisation. XGBoost followed closely, performing well but with slightly more variation in error magnitudes. Random Forest showed intermediate performance but was hampered by occasional large residuals. SVM, by contrast, exhibited the weakest performance, characterised by high variance and reduced accuracy in both validation and testing phases.

Overall, the comparative analysis based on error distribution confirms the capability of ANN in modelling the complex relationships between socioeconomic indicators and hydroelectric power generation. It highlights the importance of both tight error concentration and consistency across data subsets in evaluating model effectiveness for real-world forecasting tasks.

4.5.3 Insight into Regression Plots

To further evaluate the predictive capability and generalisation of each model, regression plots were analysed for training, validation, and testing datasets. These plots provide a graphical representation of the relationship between predicted and actual hydroelectric generation values, offering valuable insights beyond numerical metrics. An ideal regression plot displays data points tightly clustered along the 45-degree diagonal line, indicating that predictions closely match true values. Figure 4.14 presents

the regression plots for ANN, Figure 4.15 for SVM, Figure 4.16 for Random Forest, and Figure 4.17 for XGBoost.

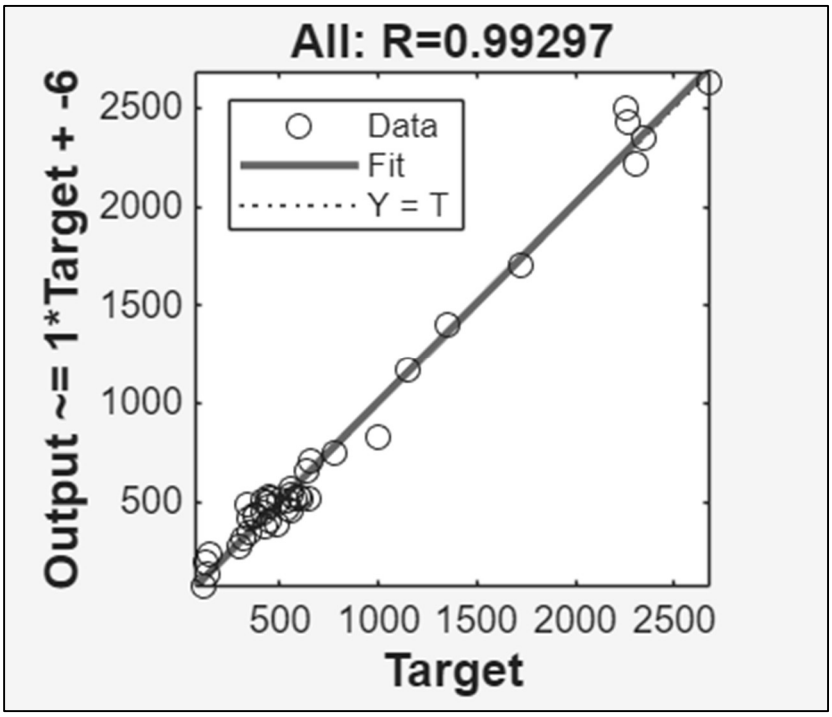

Figure 4.14 Regression Plot for ANN

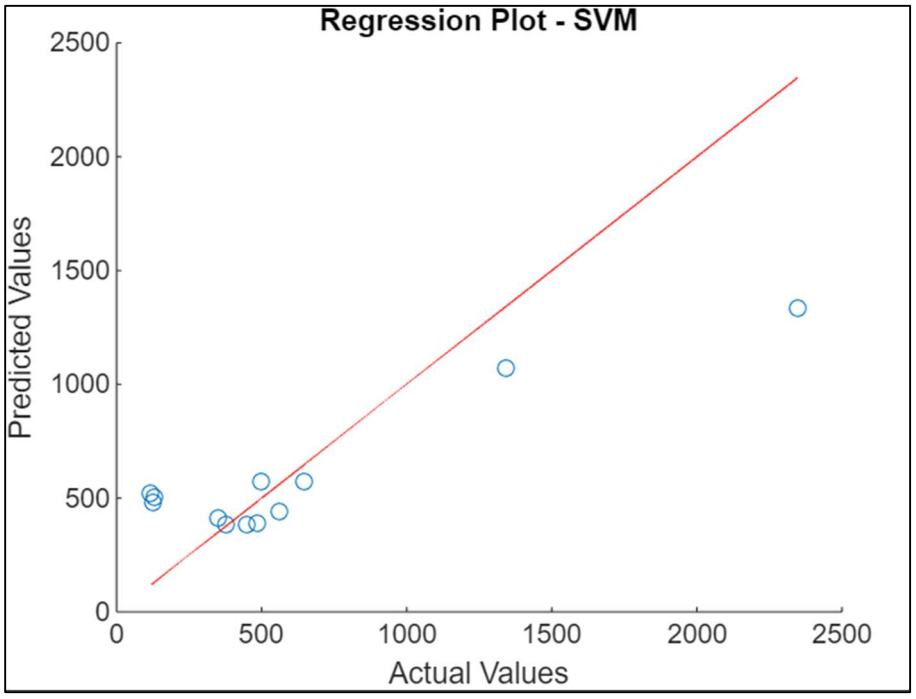

Figure 4.15 Regression Plot for SVM

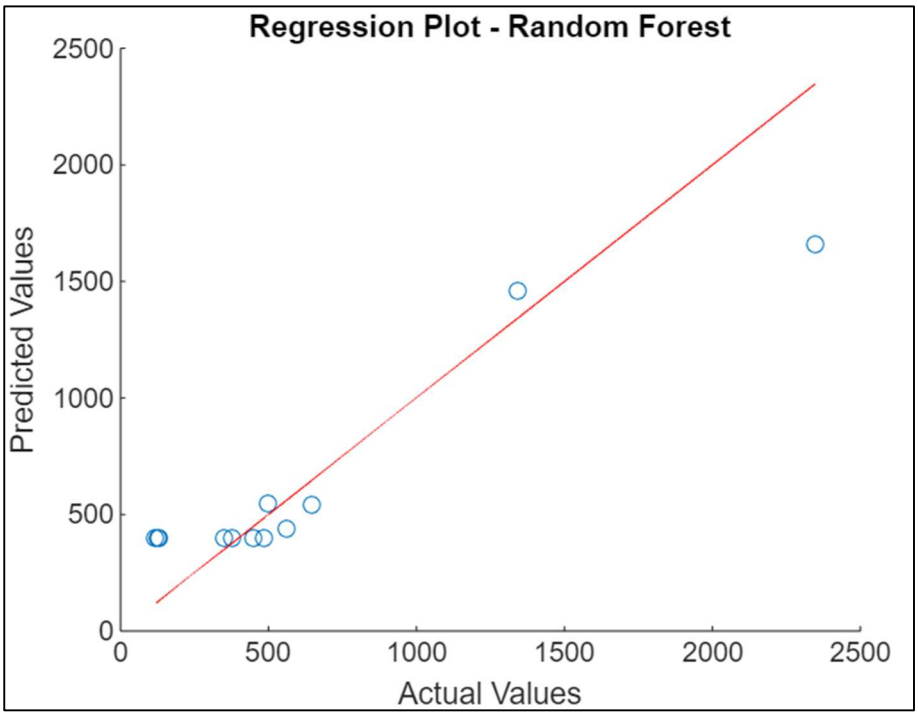

Figure 4.16 Regression Plot for Random Forest

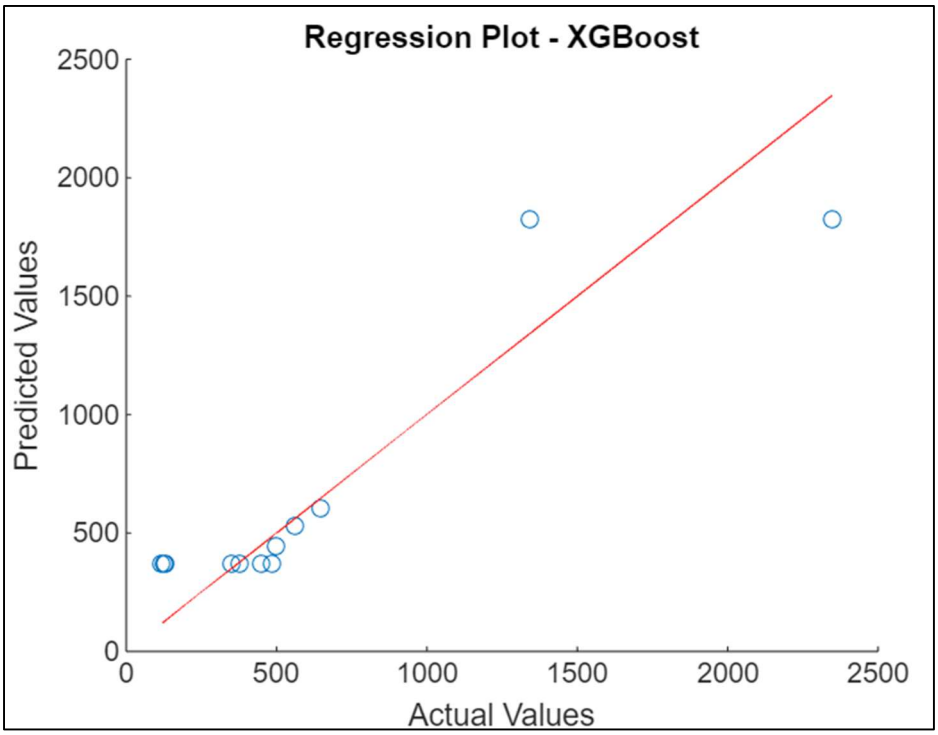

Figure 4.17 Regression Plot for XGBoost

ANN demonstrated the strongest predictive performance, with its regression plot showing a near-perfect alignment between predicted and actual values, as shown

in Figure 4.14. The data points closely followed the perfect-fit line, indicating the model's ability to accurately capture complex relationships between Population, GDP, and Energy Consumption in predicting hydroelectric generation. In contrast, SVM exhibited noticeable deviations from the perfect-fit line, with several scattered data points, particularly at higher values of hydroelectric generation, as shown in Figure 4.15. This suggests that SVM struggled to generalise well across the dataset.

The minimal scatter in the ANN's regression plot highlights its strong generalisation capability, making it the most reliable model tested. Random Forest displayed moderate performance in this regard, with a reasonable alignment to the regression line but increased scatter for larger hydroelectric generation values. This suggests that while Random Forest effectively handles nonlinear relationships, it does not fully capture intricate dependencies, leading to occasional mispredictions. XGBoost showed better consistency than SVM and Random Forest but still exhibited deviations, particularly for higher generation levels, limiting its predictive accuracy.

From a visual perspective, ANN's regression plots exhibit a dense clustering of points along the diagonal line, especially in the test dataset, indicating minimal bias and well-calibrated predictions. In contrast, the plots for SVM and Random Forest display visible fan-like patterns, especially for higher hydroelectric generation values, suggesting increasing error as output magnitude increases, a sign of underfitting or limited model flexibility. XGBoost, while closer to the ideal than SVM, still showed asymmetrical dispersion around the fit line, hinting at residual bias in the mid-to-high output range. These discrepancies are critical in real-world applications where peak prediction accuracy is necessary during periods of high energy demand. An underperforming model at these ranges could lead to underestimated supply needs or misaligned policy planning. Thus, regression plots not only validate performance numerically but also visually confirm where models succeed or fail in capturing real-world dynamics.

ANN effectively captured the nonlinear dependencies among predictors, ensuring high accuracy in predictions. SVM, however, failed to do so, leading to reduced accuracy and misalignment in the regression plot. Random Forest demonstrated some ability to manage nonlinear relationships but struggled with intricate dependencies, as seen in the increased scatter. XGBoost improved upon SVM and Random Forest in handling nonlinear patterns, but still fell short of ANN's precision and consistency.

4.6 Summary

Chapter 4 presented the results of data preprocessing, cross-validation, ANN model evaluation, and comparative machine learning analysis for forecasting hydroelectric generation. The five-fold cross-validation results showed that the input combination consisting of GDP, Energy Consumption, and Population produced the most favourable preliminary performance, with an average MSE of 1.8955×10^5 and an R-value of 0.7731, supporting its selection for further modelling. In the ANN analysis, Model B generally produced better predictive performance than the other input configurations, while the inclusion of Inflation in Model C was associated with less stable behaviour. The comparative evaluation further showed that ANN performed more consistently than SVM, Random Forest, and XGBoost under the experimental setting used in this study. However, these observations should be interpreted with caution because the dataset was limited, particularly in the testing phase, which contained only a small number of observations. Even with this limitation, the results provide useful insight into how input selection and model choice may influence hydroelectric generation forecasting, and they suggest that GDP, Energy Consumption, and Population form a more suitable input combination for the dataset used in this study.

CHAPTER 5
CONCLUSION

5.1 Conclusion

This study investigated the use of machine learning techniques to forecast hydroelectric power generation in Malaysia using selected socioeconomic indicators covering the period from 1980 to 2021. The research focused on analysing the relationship between socioeconomic variables and hydropower output, as well as evaluating the performance of several machine learning models in predicting hydroelectric generation.

The first objective of this study was to analyse the influence of socioeconomic factors, including Gross Domestic Product (GDP), energy consumption, population, and inflation, on hydroelectric power generation. Three ANN input configurations were evaluated to examine how different combinations of these variables influence prediction performance. The results indicate that the model incorporating GDP, energy consumption, and population produced the most stable and accurate predictions. This model achieved a testing Mean Squared Error (MSE) of 6.4307×10^3 with a correlation coefficient (R) of 0.9969, suggesting a strong predictive relationship between the selected variables and hydroelectric generation. In comparison, the configuration that replaced population with inflation showed slightly higher prediction variability, indicating that inflation may have a weaker predictive contribution within the dataset used in this study. These findings suggest that demographic and economic activity indicators provide useful explanatory signals for modelling hydroelectric generation trends.

The second objective was to develop, validate, and evaluate the performance of several machine learning models, namely Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost), in forecasting hydroelectric power generation. All models were trained using the same dataset and input configuration to ensure fair comparison. Based on the evaluation metrics, the ANN model demonstrated the best overall performance, followed by Random Forest and XGBoost, while the SVM model showed comparatively lower

predictive accuracy. The model evaluation was conducted using multiple indicators, including MSE, correlation coefficient (R), regression analysis, and error histogram interpretation. In addition, K-fold cross-validation produced an average validation performance of approximately 77%, suggesting that the selected modelling framework demonstrates reasonable generalisation capability despite the relatively limited dataset size. Overall, the findings indicate that ANN is a suitable modelling approach for capturing the nonlinear relationship between socioeconomic variables and hydroelectric generation within the context of this study.

While the results demonstrate the potential of machine learning models for hydropower forecasting, the limited number of observations in the testing dataset suggests that the findings should be interpreted cautiously. Nevertheless, the study provides useful insights into the role of socioeconomic indicators in energy forecasting and offers a methodological reference for future research exploring data-driven approaches in renewable energy planning.

5.2 Recommendation for Future Work

Future research may improve the robustness of hydroelectric forecasting models by expanding the available dataset. The present study relies on annual data from 1980 to 2021, which results in a relatively small number of training samples for machine learning models. Increasing the dataset size, for example by incorporating longer historical records, plant-level data, or higher temporal resolution such as monthly observations, could help improve prediction stability and model reliability.

Another possible improvement is the inclusion of additional explanatory variables related to hydrological and environmental conditions. Hydroelectric generation is directly influenced by factors such as rainfall patterns, river flow, reservoir water levels, and climate variability. Integrating these variables alongside socioeconomic indicators could allow future forecasting models to capture both environmental and economic influences on hydropower generation.

Future studies may also explore techniques for addressing the limitations of small datasets. Data balancing and augmentation approaches such as the Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) could be investigated to generate additional synthetic samples when

appropriate. These methods may help improve model training performance when the available dataset is limited.

Finally, further research could investigate ensemble or hybrid modelling approaches that combine multiple machine learning algorithms. Such strategies may allow different models to capture complementary patterns within the data and potentially improve prediction stability. Expanding comparative experiments across different modelling frameworks may also provide additional insight into suitable forecasting techniques for renewable energy planning in Malaysia

REFERENCES

Regularized_Probabilistic_Forecasting_of_Electricity_Wholesale_Price_and_Demand . (n.d.).

2018 International Conference on Power System Technology (POWERCON). (2018). IEEE.

2024 AEIT International Annual Conference (AEIT). (2024). IEEE.

Abbasian Hamedani, E., & Talebi, S. (2025). Modeling and long-term forecasting of CO2 emissions in Asia: An optimized Artificial Neural Network approach with consideration of renewable energy scenarios. Energy Conversion and Management: X, 26. https://doi.org/10.1016/j.ecmx.2025.101030

Ahmad, A., Dey, M., Raghuwanshi, S. S., Mishra, S., & Shrivastava, V. K. (2021, March 25). Operational Experience in Management of Renewable Energy in Western Regional Grid, India. 12th International Symposium on Advanced Topics in Electrical Engineering, ATEE 2021. https://doi.org/10.1109/ATEE52255.2021.9425235

Almashayikh, Y., Zawaydeh, S., Abdelsalam, E., Alkasrawi, M., Albdour, R., & Salameh, T. (2022). Pumped Hydro Storage Contributions to Achieve Jordan Energy Strategy 2020-2030. 2022 Advances in Science and Engineering Technology International Conferences, ASET 2022. https://doi.org/10.1109/ASET53988.2022.9734311

Alsamraee, S. A., & Khanna, S. (2025). High-resolution energy consumption forecasting of a university campus power plant based on advanced machine learning techniques. Energy Strategy Reviews, 60. https://doi.org/10.1016/j.esr.2025.101769

Alshammari, M. (2022). Hydroelectric Energy: Challenges, Solutions and Future Trends. International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022. https://doi.org/10.1109/ICECET55527.2022.9873025

Apaydin, H., Feizi, H., Sattari, M. T., Colak, M. S., Shamshirband, S., & Chau, K. W. (2020). Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. Water (Switzerland), 12(5). https://doi.org/10.3390/w12051500

Barzola-Monteses, J., Gómez-Romero, J., Espinoza-Andaluz, M., & Fajardo, W. (2025). Time series forecasting techniques applied to hydroelectric generation systems. In International Journal of Electrical Power and Energy Systems (Vol. 164). Elsevier Ltd. https://doi.org/10.1016/j.ijepes.2024.110424

Barzola-Monteses, J., Mite-León, M., Espinoza-Andaluz, M., Gómez-Romero, J., & Fajardo, W. (2019). Time series analysis for predicting hydroelectric power production: The ecuador case. Sustainability (Switzerland), 11(23). https://doi.org/10.3390/su11236539

Benti, N. E., Chaka, M. D., & Semie, A. G. (2023). Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects. In Sustainability (Switzerland) (Vol. 15, Number 9). MDPI. https://doi.org/10.3390/su15097087

Bespalyy, S., & Bespalaya, Y. (2025). Balancing electricity and integrating renewable energy facilities into the Unified Energy System of Kazakhstan: Challenges and opportunities. E3S Web of Conferences, 623. https://doi.org/10.1051/e3sconf/202562303001

Dahal, S., Mishra, S., Øyvang, T., Hegglid, G. J., Jha, S. K., & Chhetri, B. B. (2024). Climate Change and Seasonal Trends in Water-Energy Synergy of Hydropower Plants in the Himalayan Region. 2024 IEEE International Conference on Power System Technology, PowerCon 2024 - Proceedings. https://doi.org/10.1109/PowerCon60995.2024.10870549

Datsyuk, P., Dinçer, H., Yüksel, S., Mikhaylov, A., & Pinter, G. (2024). Comparative analysis of hydro energy determinants for European Economies using Golden Cut-oriented Quantum Spherical fuzzy modelling and causality analysis. Heliyon, 10(5). https://doi.org/10.1016/j.heliyon.2024.e26506

Dubey, A., Tiwari, A., Bhardwaj, A., Sivamohan, S., & Krishnaveni, S. (2024). Global Energy Consumption Patterns and Optimization using Big Data. Proceedings of the 9th International Conference on Communication and Electronics Systems, ICCES 2024, 1025–1028. https://doi.org/10.1109/ICCES63552.2024.10859370

Dubey, N. K., Chawla, M. P. S., & Malviya, L. (2021). 0 th IEEE International Conference on Communication Systems and Network Technologies An Artificial Neural Network Based Forecasting Strategy for Estimating Weather parameters: Application for Sizing Stand-alone Renewable Power System. 2021

10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). https://doi.org/10.1109/CSNT.2021.50

Ersan, M., & Irmak, E. (2023). Governor Control Systems in Hydroelectric Power Plants: Overview, Challenges, and Recommendations. 11th International Conference on Smart Grid, IcSmartGrid 2023. https://doi.org/10.1109/icSmartGrid58556.2023.10170958

Fard, S. N., & Ardehali, M. M. (2023). Long Term Forecasting of Electrical Energy Consumption for Iran Based on Optimized Artificial Neural Networks and Socio-Economic Indicators Data. 2023 5th International Conference on Optimizing Electrical Energy Consumption, OEEC 2023, 40–44. https://doi.org/10.1109/OEEC58272.2023.10135403

Feng, W., Lei, X., Wang, C., & Huang, H. (2021). Study on water level prediction method of Shaping Hydropower Station based on BP neural network. 2021 7th International Conference on Hydraulic and Civil Engineering and Smart Water Conservancy and Intelligent Disaster Reduction Forum, ICHCE and SWIDR 2021, 1713–1716. https://doi.org/10.1109/ICHCESWIDR54323.2021.9656344

Gunasinghe, L., Gunawardhana, L., & Rajapakse, L. (2023). Predictive Analysis of Landslide Susceptibility under Hydrological Aspects of Climate Change in Kegalle District, Sri Lanka. Moratuwa Engineering Research Conference, MERCon, 119–124. https://doi.org/10.1109/mERCon60487.2023.10355394

Gupta, K. (n.d.). Navigating the Challenge of Small Data: Optimizing Utility Company Net Revenue Forecasting with Robust Modeling Techniques. https://doi.org/10.1109/AIC.2024.67

Hartono, J., Surya, A. A., Harsono, B., Tambunan, H. B., & Purnomoadi, A. P. (n.d.). Long Term Load Demand Forecasting in Bali Province Using Deep Learning Neural Network.

Hasanah, R. N., Suyono, H., Kim, J., Muharram, Y., & Muljadi, E. (2023). Hydropower Development Towards a Full-Renewable Energy Grid in Indonesia. 2023 IEEE Industry Applications Society Annual Meeting, IAS 2023. https://doi.org/10.1109/IAS54024.2023.10406491

Hashemizadeh, A., Ju, Y., & Abadi, F. Z. B. (2024). Policy design for renewable energy development based on government support: A system dynamics model. Applied Energy, 376. https://doi.org/10.1016/j.apenergy.2024.124331

Hazmin, M. H., Mustapha, F., & Author, C. (2024). An outlook on hydropower in Malaysia: Policies, conditions, and the potential of small hydropower in Malaysian rivers as a new norm in renewable energy. Malaysian Journal of Sustainable Environment, 11(1), 1–24.

Hodoroski, D., & Jiang, Y. L. (2023). Impact of Climate Change on Long-Term Load Forecasting: Case Studies in New York State. 2023 North American Power Symposium, NAPS 2023. https://doi.org/10.1109/NAPS58826.2023.10318650

Hooshyar, D., Azevedo, R., & Yang, Y. (2024). Augmenting Deep Neural Networks with Symbolic Educational Knowledge: Towards Trustworthy and Interpretable AI for Education. Machine Learning and Knowledge Extraction, 6(1), 593–618. https://doi.org/10.3390/make6010028

IEEE Power & Energy Society., & Institute of Electrical and Electronics Engineers. (n.d.). 16th International Conference on the European Energy Market (EEM) : 18-20 September 2019, Ljubljana, Slovenia.

IEEE Staff, . (2011). 2011 Asia-Pacific Power and Energy Engineering Conference. IEEE.

Jahanger, A., Yu, Y., Awan, A., Chishti, M. Z., Radulescu, M., & Balsalobre-Lorente, D. (2022). The Impact of Hydropower Energy in Malaysia Under the EKC Hypothesis: Evidence From Quantile ARDL Approach. SAGE Open, 12(3). https://doi.org/10.1177/21582440221109580

Jāmiʻat al-Imārāt al-ʻArabīyah al-Muttaḥidah, & Institute of Electrical and Electronics Engineers. (n.d.). 5th International Conference on Renewable Energy: Generation & Application : Mon to Wed, Feb 26-28, 2018, CIT Building.

Kaewarsa, S., & Kongpaseuth, V. (2024). Hydropower Plant Available Energy Forecasting Using Artificial Neural Network and Particle Swarm Optimization. Electricity, 5(4), 751–769. https://doi.org/10.3390/electricity5040037

Kim, Y. S., Kim, M. K., Fu, N., Liu, J., Wang, J., & Srebric, J. (2025). Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models. Sustainable Cities and Society, 118. https://doi.org/10.1016/j.scs.2024.105570

Kongpaseuth, V., & Kaewarsa, S. (2023). Nam Theun 2 Hydropower Plant Energy Prediction Using Artificial Neural Network and Genetic Algorithm. Asia-Pacific Power and Energy Engineering Conference, APPEEC. https://doi.org/10.1109/APPEEC57400.2023.10561951

Kumar, D., Kumar, A., Bhattacharya, S., Lakra, S., Singh, M. P., & Roy, N. (2024). Impact of Hydro Generation Outage due to High Silt in Northern Region of India. 2024 IEEE International Conference on Power and Energy, PECon 2024, 145–150. https://doi.org/10.1109/PECON62060.2024.10827293

Kumari, S., Sreekumar, S., Singh, S., & Kothari, D. P. (2022). Comparison among ARIMA, ANN, and SVR Models for Wind Power Deviation Charge Reduction. 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022, 551–557. https://doi.org/10.1109/COM-IT-CON54601.2022.9850913

Lamidi, T. A., Nounangnonhou, C. T., Agbomahena, B. M., Richy Aza-Gnandji, M., & Lopez, I. (2024). Impacts of Climate Change on the Hydroelectric Potential of the Ouémé River (Benin) under Different Climatic Flow Scenarios. 2024 4th International Conference on Electrical Engineering and Informatics (ICon EEI), 124–129. https://doi.org/10.1109/IConEEI64414.2024.10748232

Lee, D., Ng, J. Y., Galelli, S., & Block, P. (2022). Unfolding the relationship between seasonal forecast skill and value in hydropower production: a global analysis. Hydrology and Earth System Sciences, 26(9), 2431–2448. https://doi.org/10.5194/hess-26-2431-2022

Li, H., Zhou, Q., Tian, J., & Lin, X. (2020). Energy demand forecasting for an office building based on random forests. 2020 IEEE 4th Conference on Energy Internet and Energy System Integration: Connecting the Grids Towards a Low-Carbon High-Efficiency Energy System, EI2 2020, 29–32. https://doi.org/10.1109/EI250167.2020.9347021

Luo, H., & Li, X. (2023). A hybrid model of CNN-BiLSTM and XGBoost for HVAC systems energy consumption prediction. 2023 5th International Conference on Industrial Artificial Intelligence, IAI 2023. https://doi.org/10.1109/IAI59504.2023.10327594

Matrenin, P. V., Atabaeva, L. S., & Sergeev, N. N. (2022). Limitations and Perspectives of Short-Term Renewable Energy Generation Forecasting Methods. 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences, SIBIRCON 2022, 770–774. https://doi.org/10.1109/SIBIRCON56155.2022.10017051

Meng, J., Wang, Y., Guo, H., & Ding, Y. (2023). Application of Wavelet Denoising Algorithm in Monthly Runoff Series of Fuchun River Hydropower Station.

2023 International Seminar on Computer Science and Engineering Technology (SCSET), 719–723. https://doi.org/10.1109/SCSET58950.2023.00162

Menon, S., Sultanova, N., Jayabalan, M., & Mustafina, J. (2023). A Study on Data-Driven Energy Forecasting: a Machine Learning Perspective. Proceedings - International Conference on Developments in ESystems Engineering, DeSE, 702–705. https://doi.org/10.1109/DeSE60595.2023.10469027

Muhammad, F., Qayyum, A., Bawazir, A. A., Jan, M., & Ahmed, N. (2024). Assessing the Tri-Dimensional Nexus of Energy, Environment, and Economic Growth in Pakistan: An Empirical Study. International Journal of Energy Economics and Policy, 14(4), 329–343. https://doi.org/10.32479/ijeep.16128

Nanayakkara, N., Mohamed, F. A., & Nakamura, M. (2009). System Modeling for Affordable and Sustainable Development of Small Hydro Projects in Rwanda. www.hydropowerinternational.com

National Energy Transition Roadmap_0. (n.d.).

Nwobi-Okoye, C. C., Takim, S. A., & Okiy, S. (2024). Power Planning in Hydropower Plants Using Artificial Intelligence: A Case Study of Nigeria. 2024 IEEE PES/IAS PowerAfrica, PowerAfrica 2024. https://doi.org/10.1109/PowerAfrica61624.2024.10759516

Parvathareddy, S., Yahya, A., Amuhaya, L., Samikannu, R., & Suglo, R. S. (2025). A hybrid machine learning and optimization framework for energy forecasting and management. Results in Engineering, 26. https://doi.org/10.1016/j.rineng.2025.105425

Paul, T., Raghavendra, S., Ueno, K., Ni, F., Shin, H., Nishino, K., & Shingaki, R. (2021). Forecasting of Reservoir Inflow by the Combination of Deep Learning and Conventional Machine Learning. IEEE International Conference on Data Mining Workshops, ICDMW, 2021-December, 558–565. https://doi.org/10.1109/ICDMW53433.2021.00074

Predicting_World_Energy_Consumption_Comparison_of_ANN_and_Regression_Analysis. (n.d.).

Priya, J. S., Rengaraj, R., & Thennarasu, S. (2025). Wind Power Forecasting with Machine Learning: A Comprehensive Overview. 2025 International Conference on Computing and Communication Technologies, ICCCT 2025. https://doi.org/10.1109/ICCCT63501.2025.11019417

Proceedings, 2020 11th IEEE Control & System Graduate Research Colloquium (ICSGRC 2020) : 8th August 2020, Shah Alam, Malaysia. (2020). IEEE.

Rajarajeswaran, J. (2025). Applications of Artificial Intelligence and Machine Learning for Accurate Forecasting and Optimization of Renewable Energy Generation. 3rd International Conference on Electronics and Renewable Systems, ICEARS 2025 - Proceedings, 1886–1889. https://doi.org/10.1109/ICEARS64219.2025.10940813

Rajnish, Saini, M. K., & Saroha, S. (2023). Solar Power Forecasting using Machine Learning Approaches : A Review. India International Conference on Power Electronics, IICPE. https://doi.org/10.1109/IICPE60303.2023.10474992

Rasku, T., Miettinen, J., Rinne, E., & Kiviluoma, J. (2020). Impact of 15-day energy forecasts on the hydro-thermal scheduling of a future Nordic power system. Energy, 192. https://doi.org/10.1016/j.energy.2019.116668

Ren, Y., Xia, L., & Wang, Y. (2023). Forecasting China's hydropower generation using a novel seasonal optimized multivariate grey model. Technological Forecasting and Social Change, 194. https://doi.org/10.1016/j.techfore.2023.122677

S. D.Grande, R. Gueli, M. Berlotti, & S. Cavalieri. (2024). Harnessing Multivariate AI to Enhance Hydropower Generation Forecasting. AEIT International Annual Conference (AEIT). https://doi.org/10.23919/AEIT63317.2024.10736754

Sahin, M. E., & Ozbay Karakus, M. (2024). Smart hydropower management: utilizing machine learning and deep learning method to enhance dam's energy generation efficiency. Neural Computing and Applications, 36(19), 11195–11211. https://doi.org/10.1007/s00521-024-09613-1

Salem, K. M., Rey-Hernández, J. M., Elgharib, A. O., & Rey-Martínez, F. J. (2025). Optimizing Energy Forecasting Using ANN and RF Models for HVAC and Heating Predictions. Applied Sciences (Switzerland), 15(12). https://doi.org/10.3390/app15126806

Sapitang, M., Ridwan, W. M., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A. (2020). Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. Sustainability (Switzerland), 12(15). https://doi.org/10.3390/su12156121

Sarpong, S. A., & Agyei, A. (2022). Forecasting Hydropower Generation in Ghana Using ARIMA Models. International Journal of Statistics and Probability, 11(5), 30. https://doi.org/10.5539/ijsp.v11n5p30

Sattar Hanoon, M., Najah Ahmed, A., Razzaq, A., Oudah, A. Y., Alkhayyat, A., Feng Huang, Y., kumar, P., & El-Shafie, A. (2023a). Prediction of hydropower generation via machine learning algorithms at three Gorges Dam, China. Ain Shams Engineering Journal, 14(4). https://doi.org/10.1016/j.asej.2022.101919

Sauhats, A., Petrichenko, R., Broka, Z., Baltputnis, K., & Sobolevskis, D. (2016). ANN-Based Forecasting of Hydropower Reservoir Inflow. 57th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON 2016). https://doi.org/10.1109/RTUCON.2016.7763129

Scorzato, L. (2024). Reliability and Interpretability in Science and Deep Learning. Minds and Machines, 34(3). https://doi.org/10.1007/s11023-024-09682-0

Shahgholian, G., Moazzami, M., Zanjani, S. M., Mosavi, A., & Fathollahi, A. (2023). A Hydroelectric Power Plant Brief: Classification and Application of Artificial Intelligence. SACI 2023 - IEEE 17th International Symposium on Applied Computational Intelligence and Informatics, Proceedings, 141–146. https://doi.org/10.1109/SACI58269.2023.10158597

Shamout, M. D., Khamkar, K. A., Lal, A., Danaiah, P., Mukasheva, A., & Kaushik, N. (2022). Hydropower technology as a renewable energy source of power generation and its effect on environment sustainability. International Interdisciplinary Humanitarian Conference for Sustainability, IIHC 2022 - Proceedings, 1017–1020. https://doi.org/10.1109/IIHC55949.2022.10059855

Shoaga, G. O., Ikuzwe, A., & Gupta, A. (2022). Forecasting of Monthly Hydroelectric and Solar Energy in Rwanda using SARIMA. 2022 IEEE PES/IAS PowerAfrica, PowerAfrica 2022. https://doi.org/10.1109/PowerAfrica53997.2022.9905311

Sustainable Energy Development Authority (SEDA) Malaysia, Annual Report. (2021). https://doi.org/https://www.seda.gov.my/

Teodorescu, V., & Obreja Brașoveanu, L. (2025). Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost. Computation, 13(5). https://doi.org/10.3390/computation13050127

Villeneuve, Y., Séguin, S., & Chehri, A. (2023). AI-Based Scheduling Models, Optimization, and Prediction for Hydropower Generation: Opportunities, Issues, and Future Directions. In Energies (Vol. 16, Number 8). MDPI. https://doi.org/10.3390/en16083335

Wang, Q., Wang, H., Gupta, C., Rao, A. R., & Khorasgani, H. (2020). A Non-linear Function-on-Function Model for Regression with Time Series Data. Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020, 232–239. https://doi.org/10.1109/BigData50022.2020.9378087

Wang, S., & Ma, J. (2023). A Novel Ensemble Model for Load Forecasting: Integrating Random Forest, XGBoost, and Seasonal Naive Methods. Proceedings - 2023 2nd Asian Conference on Frontiers of Power and Energy, ACFPE 2023, 114–118. https://doi.org/10.1109/ACFPE59335.2023.10455494

Wang, Y. (2024). Deep Learning Techniques in Renewable Energy Forecasting: Solving Intermittency and Instability in Solar and Wind Energy. 2024 4th International Conference on Energy Engineering and Power Systems, EEPS 2024, 143–146. https://doi.org/10.1109/EEPS63402.2024.10804395

Zha, L., & Jiang, D. (2022). An Ultra-Short-Term and Short-Term Wind Power Forecasting Approach Based on Optimized Artificial Neural Network with Time Series Reconstruction. 2022 4th International Conference on Smart Power and Internet Energy Systems, SPIES 2022, 2068–2073. https://doi.org/10.1109/SPIES55999.2022.10082352

Zhang, X., & Liu, C. A. (2023). Model averaging prediction by K-fold cross-validation. Journal of Econometrics, 235(1), 280–301. https://doi.org/10.1016/j.jeconom.2022.04.007

Zhang, Z., Wang, X., Wang, J., & Yuan, X. (2023). Power Load Forecasting Model Based on LSTM and Prophet algorithm. 2023 3rd International Conference on Consumer Electronics and Computer Engineering, ICCECE 2023, 156–159. https://doi.org/10.1109/ICCECE58074.2023.10135304

APPENDICES

APPENDIX 1

Codes for SVM, Random Forest and XGBoost

```
%% Load and Prepare Data
clc; clear; close all;

% Load dataset
data = readtable('complete data.xlsx');

% Extract input features and target variable
X = table2array(data(:, 1:end-1)); % Assuming last column is target
Y = table2array(data(:, end));

% Split Data (70% Training, 15% Validation, 15% Testing)
cv = cvpartition(size(X,1),'HoldOut',0.30);
train_idx = training(cv,1);
test_idx = test(cv,1);

X_train = X(train_idx, :);
Y_train = Y(train_idx, :);
X_test = X(test_idx, :);
Y_test = Y(test_idx, :);

```

```
% Further split train set into training (70%) and validation (30%)
cv2 = cvpartition(length(Y_train), 'HoldOut', 0.30);
X_train_final = X_train(training(cv2,1), :);
Y_train_final = Y_train(training(cv2,1), :);
X_val = X_train(test(cv2,1), :);
Y_val = Y_train(test(cv2,1), :);

% Normalize Features
[X_train_final, mu, sigma] = zscore(X_train_final);
X_val = (X_val - mu) ./ sigma;
X_test = (X_test - mu) ./ sigma;
% Display mean and std dev for each feature
featureNames = data.Properties.VariableNames(1:end-1);

fprintf('Z-Score Normalization Stats:\n');
for i = 1:length(mu)
    fprintf('%s: Mean = %.4f, Std = %.4f\n', featureNames{i}, mu(i), sigma(i));
end
```

```
%% Train Models

% Train SVM
Mdl_SVM = fitrsvm(X_train_final, Y_train_final, 'KernelFunction', 'rbf');
Y_pred_SVM_train = predict(Mdl_SVM, X_train_final);
Y_pred_SVM_val = predict(Mdl_SVM, X_val);
Y_pred_SVM_test = predict(Mdl_SVM, X_test);

% Train Random Forest
Mdl_RF = fitrensemble(X_train_final, Y_train_final, 'Method', 'Bag');
Y_pred_RF_train = predict(Mdl_RF, X_train_final);
Y_pred_RF_val = predict(Mdl_RF, X_val);
Y_pred_RF_test = predict(Mdl_RF, X_test);

% Train XGBoost
Mdl_XGB = fitrensemble(X_train_final, Y_train_final, 'Method', 'LSBoost');
Y_pred_XGB_train = predict(Mdl_XGB, X_train_final);
Y_pred_XGB_val = predict(Mdl_XGB, X_val);
Y_pred_XGB_test = predict(Mdl_XGB, X_test);

```

```
%% Performance Evaluation
models = {'SVM', 'Random Forest', 'XGBoost'};
train_preds = {Y_pred_SVM_train, Y_pred_RF_train, Y_pred_XGB_train};
val_preds = {Y_pred_SVM_val, Y_pred_RF_val, Y_pred_XGB_val};
test_preds = {Y_pred_SVM_test, Y_pred_RF_test, Y_pred_XGB_test};

for i = 1:length(models)
    Y_pred_train = train_preds{i};
    Y_pred_val = val_preds{i};
    Y_pred_test = test_preds{i};

    % Calculate errors
    errors_train = Y_train_final - Y_pred_train;
    errors_val = Y_val - Y_pred_val;
    errors_test = Y_test - Y_pred_test;

    % Mean Squared Error (MSE)
    mse_train = mean(errors_train.^2);
    mse_val = mean(errors_val.^2);
    mse_test = mean(errors_test.^2);

```

```
        % Compute R-values
        R_train = corr(Y_train_final, Y_pred_train, 'Rows', 'complete');
        R_val = corr(Y_val, Y_pred_val, 'Rows', 'complete');
        R_test = corr(Y_test, Y_pred_test, 'Rows', 'complete');

        % Handle NaN cases if standard deviation is 0 (constant predictions)
        if isnan(R_train), R_train = 0; end
        if isnan(R_val), R_val = 0; end
        if isnan(R_test), R_test = 0; end

        % Display results
        fprintf('%s Model:\n', models{i});
        fprintf('Training MSE: %.4f | R: %.4f\n', mse_train, R_train);
        fprintf('Validation MSE: %.4f | R: %.4f\n', mse_val, R_val);
        fprintf('Test MSE: %.4f | R: %.4f\n\n', mse_test, R_test);

```

```
        % Regression Plot
        figure;
        scatter(Y_test, Y_pred_test);
        hold on;
        plot(Y_test, Y_test, 'r'); % Perfect fit line
        xlabel('Actual Values');
        ylabel('Predicted Values');
        title(['Regression Plot - ', models{i}]);
        hold off;

        % Enhanced Error Histogram (Training, Validation, and Test)
        figure;
        hold on;
        histogram(errors_train, 'Normalization', 'probability', 'FaceColor', 'b');
        histogram(errors_val, 'Normalization', 'probability', 'FaceColor', 'g');
        histogram(errors_test, 'Normalization', 'probability', 'FaceColor', 'r');
        yline(0, '--k', 'Zero Error');
        title(['Error Histogram - ', models{i}]);
        xlabel('Errors = Targets - Outputs');
        ylabel('Instances');
        legend('Training', 'Validation', 'Test');
        hold off;
    end

```

AUTHOR'S PROFILE

'Aliaa Aqilah Binti Nor Azman obtained her Bachelor of Engineering (Hons.) in Electrical Engineering in 2023 from Universiti Teknologi MARA (UiTM) and MSc in Electrical Engineering from Universiti Teknologi MARA (UiTM). Her MSc research focuses on the application of artificial intelligence and machine learning techniques for forecasting hydroelectric power generation in Malaysia, with particular emphasis on the relationship between socioeconomic indicators and renewable energy output. Her study evaluates the performance of Artificial Neural Networks (ANN) in comparison with Support Vector Machines (SVM), Random Forest, and XGBoost, highlighting the potential of data-driven models to support energy planning and policy development.

LIST OF PUBLICATIONS

A. A. Nor Azman, M. A. Abdul Aziz, N. Abd Razak, and M. Md Kamal, "Evaluating AI-Based Models for Energy Economy Nexus Analysis: A Focus on Hydroelectric Power," in Proc. 2025 IEEE 8th International Conference on Electrical, Control and Computer Engineering (InECCE), Kuantan, Malaysia, Aug. 2025, pp. 1-6.

A. A. Nor Azman, M. A. Abdul Aziz, N. Abd Razak, and M. Md Kamal, "Comparative Analysis of Machine Learning Models for Forecasting Hydroelectric Generation Using Socioeconomic Indicators," Journal of Electrical & Electronic Systems Research (JEESR), vol. 27, no. 1, pp. 26-35, Oct. 2025.