Comparative Analysis of Machine Learning Models for Forecasting Hydroelectric Generation using Socioeconomic Indicators

Aliaa Aqilah Nor Azman, Mohd Azri Abdul Aziz, Noorfadzli Abd Razak and Mahanijah Md Kamal

***Abstract*—Hydropower plays a significant role in Malaysia's renewable energy mix, particularly in regions with abundant water resources such as Sarawak. Accurate forecasting of hydroelectric generation is increasingly important to support effective energy planning and the country's sustainability objectives. This study explores the performance of four machine learning models: Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, and XGBoost in forecasting Malaysia's hydroelectric power output using socioeconomic indicators, including Gross Domestic Product (GDP), energy consumption, and population. ANN demonstrated the most promising results among these models, achieving a testing Mean Squared Error (MSE) of 1.1541×10^4 and a correlation coefficient (R) of 0.9962. These results suggest that ANN can capture the underlying patterns within the data and may offer a valuable tool for improving the reliability of hydropower generation forecasts, thereby contributing to Malaysia's ongoing efforts toward renewable energy development.**

***Index Terms*— Artificial Neural Network, economic indicators, energy consumption, energy forecasting, GDP, hydroelectric power, machine learning, population, Random Forest, Support Vector Machine, XGBoost**

I. INTRODUCTION

Hydropower has long contributed to Malaysia's renewable energy development, particularly in regions with abundant water resources such as Sarawak. According to the Malaysia Renewable Energy Roadmap (MyRER), it is identified as one of four key areas for expanding clean energy, alongside solar, bioenergy, and emerging technologies [1]. The roadmap outlines targets to increase renewable energy's share in the national installed capacity mix to 40% by 2035, with hydropower remaining a relevant and established component. As of 2023, hydropower accounted for approximately 15% of Malaysia's total energy mix, reflecting its continued role in national energy planning. In support of this, initiatives such as the Feed-in Tariff (FiT) and Low Carbon Power Generation (LCPG) have been introduced to promote small-scale hydro projects and facilitate a shift toward low-carbon energy development [1].

Beyond electricity generation, hydropower can potentially deliver broader social and environmental benefits, especially in rural and less-developed areas. Recent reviews indicate that small-scale hydropower is receiving growing attention in Peninsular Malaysia as a means to meet local energy needs while reducing reliance on fossil fuels [2]. Although Malaysia is estimated to have over 29,000 MW of gross hydropower potential, only a portion of this capacity has been utilised to date [2]. This indicates an opportunity for further development in the sector, which could support energy diversification and local development objectives.

Accurate forecasting of hydropower generation is becoming increasingly crucial as Malaysia pursues a more sustainable and balanced energy mix. However, traditional forecasting methods often face limitations in modelling the complex, nonlinear interactions that characterise energy systems, particularly when socioeconomic variables are involved. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) techniques have gained attention for improving forecasting accuracy. Models such as ANN, SVM, Random Forest, and XGBoost have shown promising results in energy-related studies due to their capacity to learn patterns from large and complex datasets [3], [4].

Socioeconomic factors such as GDP, energy consumption, and population growth are recognised as influential drivers of energy production and demand. Their interrelationships can shape trends in hydropower generation, particularly in developing economies. For instance, Bawazir et al. [5] found that long-term patterns in renewable energy use, including hydropower, are closely tied to economic performance in Pakistan, suggesting the presence of feedback between energy development and economic growth.

This study compares four AI-based models, ANN, SVM, Random Forest, and XGBoost, to forecast Malaysia's hydroelectric generation using historical socioeconomic data. By evaluating model performance under consistent conditions, the study offers insights into the potential of data-driven approaches to support more informed and effective energy planning. The remainder of this paper is organised as follows: Section 2 presents the literature review, followed by Sections 3 and Section 4, which cover the methodology and the results and

This manuscript is submitted on 19th June 2025, revised on 8th August 2025, accepted on 17th September 2025 and published on 31st October 2025.
Aliaa Aqilah Nor Azman is a student at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia (e-mail: qlhazman@gmail.com).

Mohd Azri Abdul Aziz is with Faculty of Electrical Engineering, UiTM, and a fellow at the Innovative Electromobility Research Lab (ITEM), UiTM (e-mail: azriaziz@uitm.edu.my).

Noorfadzli Abdul Razak is with Faculty of Electrical Engineering, UiTM, and Head of the ITEM Research Lab (e-mail: noorfadzli@uitm.edu.my).

Mahanijah Md Kamal is with the Faculty of Electrical Engineering, UiTM (e-mail: mahani724@uitm.edu.my).

*Corresponding author: azriaziz@uitm.edu.my

discussion, respectively. The last two sections provide the concluding remarks and recommendations.

II. Literature Review

This literature review examines previous studies on hydroelectric power forecasting. It discusses the importance of accurate forecasting and the main challenges involved. The review also explores the role of socioeconomic factors such as GDP and population, and compares traditional forecasting methods with modern machine learning approaches. The goal is to provide an understanding of current practices and potential areas for improvement.

A. Hydroelectric Generation Forecasting: Importance and Challenges

Hydroelectric power plays a crucial role in the renewable energy mix of many countries. In Indonesia, for example, hydropower accounted for 57% of all electricity generated from renewable sources by the end of 2021, underscoring its dominance over more intermittent sources such as wind and solar [6]. The reliability, flexibility, and efficiency of hydropower often exceed 90% making it a promising option for countries looking to reduce dependence on fossil fuels [6]. Additionally, its ability to provide ancillary services, stabilise fluctuations in other renewable energy sources, and respond quickly to demand changes strengthens its potential to support energy security [7]. Furthermore, hydropower is considered an environmentally friendly alternative to conventional fossil-fuel-based generation, offering a cleaner means of electricity production [7].

However, forecasting hydroelectric generation involves several challenges due to various environmental and external factors. Seasonal rainfall, temperature variations, and changes in inflow patterns introduce a high degree of uncertainty in predicting hydropower output [8]**.** For instance, in the Indian Himalayan region, extreme silt buildup during monsoon seasons has led to frequent shutdowns of hydro units to avoid turbine damage, resulting in significant capacity losses [9]. These environmental factors are compounded by socio-economic and regulatory influences, such as competing water demands and operational constraints, which further complicate prediction accuracy [7]. The non-linear relationships between generation and external factors, such as the effects of climate change on water levels and melting rates, also add complexity to forecasting models [10].

Reliable forecasting of hydroelectric generation is critical for effective energy planning and decision-making. Accurate predictions of inflows and generation capacity are essential for optimising resources, ensuring the economic viability of hydropower projects, and preventing overuse or underuse of infrastructure [11]. Poor forecasting can lead to operational inefficiencies, such as unutilized capacity or excessive wear on equipment, which have financial consequences [12]. Additionally, in the context of modern power systems, reliable forecasts of hydropower generation are vital for balancing grid stability, especially when integrated with variable energy sources like wind and solar [4]. The growing adoption of AI-based forecasting methods highlights the increasing recognition of the need for accurate, timely insights to manage hydropower assets effectively [12], [13].

B. Socioeconomic Indicators and Their Relevance in Energy Forecasting

Including socioeconomic indicators such as GDP, population, and energy consumption in hydroelectric energy forecasting is grounded in their direct influence on energy demand and development planning. Energy is a key driver of national development, influencing industrial growth and overall economic progress [13]. Energy consumption patterns, influenced by household occupancy and environmental conditions, are fundamental when predicting demand. For instance, combining occupancy data with energy usage and weather variables has been shown to improve prediction accuracy, with R^2 values surpassing 0.85 in some models [14]. These findings suggest that incorporating socioeconomic variables such as population behaviour provides essential context for understanding energy consumption trends.

Several studies have confirmed that electricity demand is closely linked to economic growth and demographic changes. As population sizes increase and urbanisation spreads, electricity demand rises due to higher technological adoption and infrastructure development [15]. For example, emerging energy systems in smart cities have highlighted the importance of integrating socioeconomic factors, although this integration remains challenging due to the diverse nature of the data involved [14]. The increasing importance of population and economic variables in forecasting reflects a broader trend where socioeconomic progress drives energy demand growth.

In hydroelectric energy, socioeconomic factors can influence both energy demand and the design of energy systems. In Iran, for instance, forecasts of energy consumption have been significantly improved by incorporating GDP and population data alongside traditional energy consumption measures [16]. Hydroelectric power plants serve multiple purposes beyond electricity generation, including flood control, irrigation, and water storage, all of which align with broader socioeconomic needs [13]. However, factors such as community resistance to displacement, fiscal constraints, and a lack of skilled labour in some developing countries complicate the deployment of hydroelectric systems [7]. These challenges demonstrate the importance of considering socioeconomic variables when evaluating the feasibility and effectiveness of hydroelectric power projects.

C. Traditional and Statistical Approaches in Hydropower Prediction

Historically, traditional time series models such as Autoregressive Integrated Moving Average (ARIMA), regression-based techniques, and grey models have been widely adopted for hydropower forecasting due to their straightforward mathematical structure and interpretability. These models have shown reliable performance, particularly in capturing linear trends and periodic behaviours in time-series data. ARIMA, in

particular, has been favoured for its ability to handle temporal dependencies, and has been a standard for modelling hydrological time series [17]. Regression techniques, such as multiple linear regression and k-Nearest Neighbour, have also been used for short-term water level prediction because of their simplicity [18]. However, while these statistical tools perform well under stable, linear conditions, they may struggle when dealing with nonlinearity and variable interactions that characterise real-world hydropower systems [4].

Despite their historical significance, traditional methods have limitations that may affect their performance in more complex forecasting scenarios. Multivariate regression models are effective for analysing cross-sectional data but struggle to capture dynamic temporal patterns in time series forecasting. This limitation becomes more apparent in hydrological systems, where relationships between variables like rainfall, inflow, and power generation can change over time. ARIMA models help capture stationarity and seasonality, but rely on linear assumptions, which can limit their performance when dealing with complex or highly variable inputs such as fluctuating climatic conditions or sudden demand spikes [19]. Furthermore, these models can fail to predict extreme hydropower production or peak inflows, which are crucial for operational planning in energy generation [18]. Traditional models are also sometimes prone to overfitting, mainly when employed in simpler single-model approaches, where sensitivity to hyperparameter tuning may affect their predictive accuracy [19].

Given these challenges, researchers have been exploring more advanced techniques, including machine learning models and hybrid approaches, which have shown promise in capturing nonlinear patterns and multi-factor interactions more effectively than traditional methods. Such techniques can provide improved performance in forecasting complex hydropower systems, especially in environments characterised by significant variability [20].

D. Machine Learning in Energy Forecasting: A Growing Trend

In recent years, machine learning (ML) has been used for energy forecasting, gaining increasing attention, driven by the need for accurate and adaptive models that can handle complex, nonlinear data. Although traditional models have their merits, they often fail to capture the intricate relationships within modern energy systems. As a result, ML has emerged as a promising alternative, offering flexibility and the ability to learn from data without relying on rigid assumptions. The integration of ML into areas like hydropower and renewable energy forecasting has shown potential in improving system efficiency and grid stability, although its widespread impact is still being explored [13], [21].

Several ML models have been widely applied in energy forecasting, with ANN, SVM, Random Forests, and Gradient Boosting models such as XGBoost being among the most commonly used. ANN is favoured for its ability to model nonlinear relationships and temporal dependencies in data [22], while SVM is often preferred for its generalisation capabilities, particularly in situations with smaller datasets [4]. Random Forest and XGBoost, on the other hand, have gained recognition for their robustness and high predictive accuracy, particularly when dealing with high-dimensional or noisy data [15], [23]. Additionally, hybrid models that combine multiple ML approaches are being explored to overcome the limitations of single models and improve forecasting accuracy [20].

Unlike traditional statistical methods, ML models provide several advantages, including handling nonlinearities, feature interactions, and temporal dependencies in a way that conventional methods cannot. For instance, an ANN can automatically extract complex patterns from data, which improves forecasting accuracy and enables the handling of time-series data more effectively [22], [24]. Ensemble methods like Random Forest and XGBoost offer robust performance and can mitigate the effects of overfitting, leading to more reliable predictions. These strengths help explain why ML is rapidly becoming an essential tool for energy forecasting, providing a more flexible and data-driven approach compared to traditional models [21], [25].

E. Artificial Neural Networks (ANN) in Hydropower and Energy Forecasting

ANNs have demonstrated their utility in energy forecasting, primarily due to their capacity to model complex, nonlinear relationships in data. For instance, ANN models have been applied to predict energy consumption in buildings, where they can capture nonlinear patterns that traditional statistical methods may not fully address [20]. In hydropower forecasting, ANN models have been explored to predict reservoir inflows and energy production. While they have shown potential, performance varies depending on the specific model configuration and the data used, indicating the importance of proper setup and data selection [18]. Furthermore, ANN models have been successfully implemented in predicting electricity demand based on temperature variations, where they outperformed some existing methods, highlighting their effectiveness in handling multivariate inputs [26].

The strength of ANNs lies in their ability to model complex, nonlinear behaviours, making them particularly well-suited for systems like energy forecasting, where such relationships are common. For example, ANNs can effectively capture the dynamic and nonstationary behaviour of electricity markets and hydropower systems, adapting to changing conditions like seasonal variations [17]. This ability to approximate continuous functions allows ANN models to perform well in renewable energy forecasting and other areas with fluctuating patterns [16]. In wind speed forecasting, ANN models have also demonstrated their ability to provide more accurate predictions than traditional models such as ARIMA, especially when dealing with high variability in the data [22].

However, despite their strengths, ANNs come with challenges. One of the main concerns is the risk of overfitting, especially when models are trained on small or noisy datasets. This can limit their generalisation capabilities, making them less effective on unseen data [21]. Additionally, ANNs require substantial computational resources and large datasets for practical training [8]. Another limitation is their lack of

interpretability, which can be problematic in applications where understanding the decision-making process is essential. Furthermore, optimising ANN models requires careful parameter selection and substantial trial-and-error, which can be time-consuming [3]. Despite these challenges, ANNs remain a powerful tool in energy forecasting when applied with appropriate care and resources.

F. Ensemble Methods (Random Forest and XGBoost) in Energy Modelling

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 Extreme Gradient Boosting (XGBoost) are particularly noted for their robustness and ability to handle complex, nonlinear relationships in data. By combining multiple decision trees, Random Forest helps reduce variance and overfitting, thus enhancing prediction reliability. It has been successfully applied in energy forecasting tasks, including predicting building energy consumption 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 [10], [27].

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. Studies have shown that XGBoost can be highly effective when integrated into hybrid models. For instance, a hybrid ensemble model combining Random Forest, XGBoost, and time series components like Seasonal Naive achieved an R^2 value of 0.95941 and significantly improved forecasting accuracy and robustness compared to standalone models [19], [20].

In conclusion, Random Forest and XGBoost are valuable tools in energy modelling. Random Forest is praised for its robustness and ability to reduce overfitting, making it suitable for handling diverse input data. On the other hand, XGBoost is recognised for its high predictive accuracy and flexibility, particularly in complex datasets. When these methods are used in hybrid models, they can complement each other's strengths, leading to improved overall performance in energy consumption and generation forecasting [21], [27].

III. METHODOLOGY

This chapter presents the methodology to develop and evaluate four machine learning models for forecasting hydroelectric power generation based on key socioeconomic indicators. The main goal is to identify the most reliable model for predicting hydroelectric generation using variables such as GDP, Energy Consumption, and Population.

The overall methodological framework outlined in the flowchart, as shown in Fig. 1, is structured into five main stages: (i) data collection and preprocessing, (ii) model architecture and configuration, (iii) training and validation process, (iv) testing and performance evaluation, and (v) result analysis and comparison. Each stage is designed to ensure consistency and fairness in model comparison using the same dataset, performance metrics, and evaluation conditions.

By examining the performance of each machine learning technique under the same experimental setup, this study aims to offer insights into model suitability for forecasting hydroelectric generation using socioeconomic inputs under identical experimental conditions.

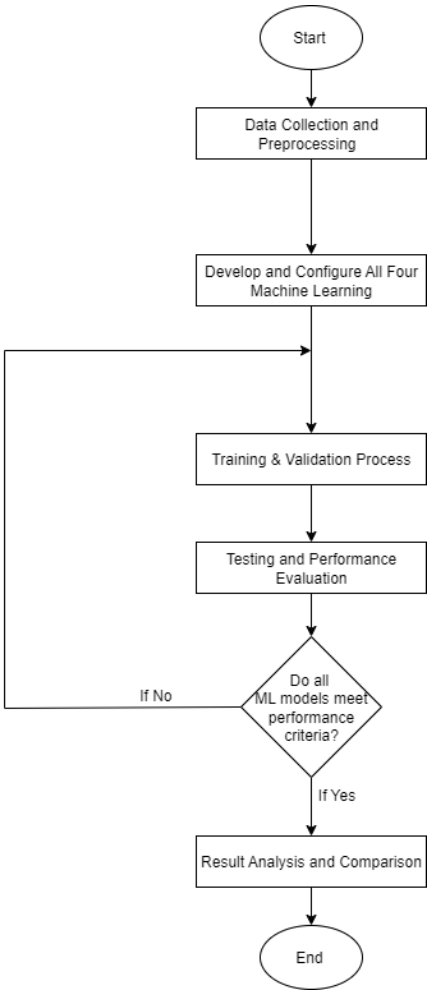


Fig. 1. Flowchart of the Project

A. Data Collection and Preprocessing

This study used annual data from 1980 to 2021, sourced from the Malaysia Energy Information Hub and Macrotrends, ensuring reliable coverage of key socioeconomic and energy indicators. The selected input variables were GDP, Energy Consumption, and Population, while the target variable was Hydroelectric Power Generation.

Before model training and evaluation, the dataset underwent preprocessing steps to ensure consistency and eliminate biases

caused by differences in data scales. Min-max normalisation was used for the ANN model as shown in (1)[28], required by the Neural Net Fitting App. This technique typically scales each feature to a fixed range [0, 1]. This ensures that all input values are within the same scale, essential for neural networks to converge efficiently during training.

$$x_{norm} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (1)$$

For SVM, Random Forest, and XGBoost, z-score normalisation as shown in (2) was applied programmatically to align with each algorithm's internal assumptions [28]. Where μ is the mean and σ is the standard deviation of the feature. Z-score normalisation benefits algorithms like SVM, which are sensitive to feature scaling when computing distances or defining decision boundaries. This standardisation method transforms the data to have a mean of 0 and a standard deviation of 1.

$$x_{std} = \frac{(x - \mu)}{\sigma} \quad (2)$$

Any missing values in the dataset were handled using appropriate imputation methods to ensure completeness without sacrificing data integrity. After preprocessing, the data was divided into training and testing sets in a 70-15-15 ratio. This split allowed for balanced model development, tuning, and unbiased performance evaluation across all models.

B. Model Selection and Configuration

To evaluate the comparative performance of various machine learning approaches in forecasting hydroelectric generation, this study employed four distinct models: ANN, SVM, Random Forest, and XGBoost. These models were selected due to their individual strengths and proven reliability in regression and energy-related forecasting applications, offering a balanced representation of both traditional and ensemble-based learning techniques.

The ANN was chosen for its ability to capture complex, nonlinear interactions between input variables. Due to its capacity to model nonlinear relationships, the ANN can perform effectively even with datasets of modest size. In this study, the ANN was implemented using MATLAB's Neural Net Fitting App, chosen for its convenience and built-in tools for preprocessing, partitioning, and visualising training performance, regression diagrams, and error histograms.

The ANN model architecture consisted of three input neurons (GDP, Energy Consumption, and Population), one hidden layer with ten neurons, and a single output neuron for hydroelectric generation. The model was trained using the Levenberg–Marquardt algorithm with a maximum of 1000 epochs, a performance goal of zero, and early stopping after six validation failures. Training was halted at epoch 13 in practice when no further improvement was observed in validation performance.

SVM was selected based on its strong performance in high-dimensional spaces and suitability for small datasets. Using the fitrsvm function in MATLAB with a Radial Basis Function (RBF) kernel, the SVM model was designed to capture non-linear relationships with minimal manual configuration. The input data were standardised using z-score normalisation before training to align with the algorithm's assumptions.

Random Forest was included in the study due to its robustness against overfitting and ability to manage datasets with irrelevant or noisy features. Implemented using MATLAB's fitrensemble function with the Bagging method, the model was allowed to use default parameters to reflect its base-level performance.

Similarly, the XGBoost model was incorporated due to its reputation for high accuracy and computational efficiency in structured data problems. Built using the same fitrensemble function but with the LSBoost (Least Squares Boosting) method, XGBoost constructs trees sequentially to minimise residual errors, making it particularly effective in refining predictions over iterations.

Although the ANN was developed through a graphical interface, while the other models were coded via scripting, all were trained on the same preprocessed data using identical inputs and evaluation metrics. To ensure transparency and reproducibility, the specific configurations and training parameters used for each model are summarised in Table 1.

TABLE I. MODEL CONFIGURATIONS AND TRAINING PARAMETERS

Model	MATLAB Tool / Function	Learning Method	Key Parameters (Values)
ANN	Neural Net Fitting App	Levenberg–Marquardt	3 input neurons (GDP, Energy, Population) 1 hidden layer with 10 neurons Performance goal: 0 Max iterations: 1000 Validation checks: 6 Initial Mu: 0.001 Early stopping at epoch:13
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

C. Training and Validation Process

Each model was trained using the same 70% training subset, with the remaining 30% split equally into validation and test sets. This partitioning was controlled using MATLAB's partition function for the SVM, Random Forest, and XGBoost models, ensuring randomised and consistent splits. For the ANN model, data division was configured directly within the

Neural Net Fitting App, allowing internal training, validation, and test ratios to be set as 70%, 15%, and 15%, respectively.

During the training phase, each model learned the functional mapping between the selected input features (GDP, Energy Consumption, and Population) and the target output (Hydroelectric Generation). The validation subset was used to monitor the model's generalisation ability and detect any signs of overfitting or underfitting. While the Neural Net Fitting App offered built-in visualisation tools, the remaining models required explicit validation handling via code and plotting.

In addition to the standard train-validation-test split, a five-fold cross-validation procedure was applied to all models to assess their generalisation performance. This involved dividing the dataset into five equal parts and iteratively training on four parts while testing the remaining one. The process was repeated five times so that each fold served as the test set once. All five iterations' performance metrics (MSE and R-value) were averaged to provide a more reliable predictive accuracy and robustness evaluation.

D. Performance Evaluation and Comparison

Model performance was assessed using two key evaluation metrics: Mean Squared Error (MSE) and the correlation coefficient (R-value). These were applied to training, validation, and testing datasets to assess learning behaviour and generalisation performance. Equation (3) shows the formula of MSE, where, $\hat{y}_i$ represents the predicted values, y_i is the actual value, and n is the number of test samples [29]. A lower MSE indicates that predictions are closer to the actual outputs. At the same time, (4) is the formula for the R-value where $\hat{y}$ is the mean of the actual values. A higher R-value (closer to 1) reflects a stronger linear correlation between predicted and actual values [29].

Regression plots were generated to visualise the relationship between predicted and actual values, where a tighter alignment with the 45-degree reference line indicated more accurate predictions. Additionally, error histograms were produced to assess residual distribution across all subsets, highlighting consistency and potential bias.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\, y_i \tag{3}$$

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2}} \tag{4}$$

IV. Results And Discussion

This study evaluates the performance of four machine learning models, ANN, SVM, Random Forest, and XGBoost, in forecasting hydroelectric power generation using economic indicators (Population, GDP, and Energy Consumption). Each model was trained and tested on data from 1980 to 2021, and evaluated using MSE and R-values, regression plots and error histograms.

A. Performance Evaluation of Machine Learning Models

This subsection presents a comparative assessment of model performance based on the computed MSE and R-values for the training, validation, and testing phases. A detailed summary of the performance metrics for ANN, SVM, Random Forest, and XGBoost is provided in Table 2.

TABLE II. Performance Evaluation of ANN, SVM, Random Forest and XGBoost

Models		Training	Validation	Testing
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	MSE	1.0423e+05	7.6374e+04	6.2457e+04
Forest	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

ANN demonstrated the highest predictive accuracy among the evaluated models with the lowest MSE across all phases. During training, ANN achieved an MSE of 5.0355×10^3, significantly lower than the other models, and a substantial R-value of 0.9941. This trend continued into validation (MSE = $9.3132\text{x}10^4$, R = 0.9788) and testing (MSE = 1.1541×10^4, R = 0.9962), indicating excellent generalisation and minimal error. In contrast, the SVM model exhibited the highest error, with an MSE of 1.7091×10^5 during training. Although its correlation improved in validation (R = 0.9865), its high testing error (MSE = 1.3031×10^5) made it less reliable. Random Forest and XGBoost offered moderate performance, with Random Forest showing slightly lower error values than SVM, and XGBoost performing marginally better in testing but struggling in the training phase.

ANN consistently showed the highest correlation between predicted and actual values, with R-values exceeding 0.99 in training and testing, confirming its strong fit to data trends. Despite having a high R-value of 0.9703 during training, SVM exhibited a weaker correlation (R = 0.9250) in testing, suggesting inconsistency in performance across phases. Random Forest and XGBoost displayed moderate correlation strengths, with XGBoost having the lowest training R-value (0.8487), which indicates difficulties in learning underlying data patterns. Although Random Forest improved validation and testing, it never reached the correlation levels observed in ANN.

A key aspect of evaluating model performance is how well each model generalises to unseen data. The ANN model showed consistently strong results from training to testing, suggesting a good balance between model complexity and generalisation ability. While Random Forest did not surpass ANN, it showed relatively better generalisation than SVM, as reflected in the decrease in error from training to validation. The SVM model maintained relatively stable correlation values, though the overall error levels remained high, indicating some limitations in adapting to new data. XGBoost, on the other hand, appeared to face challenges with generalisation, showing less consistent performance, particularly during the training phase, where its correlation was comparatively lower.

Regarding overall performance, ANN emerged as the most suitable model for this forecasting task due to its low errors and high predictive correlation. With its high MSE, SVM was the least reliable, as it struggled to capture non-linear relationships effectively. While better than SVM in some aspects, Random Forest still lagged behind ANN due to higher error rates. XGBoost showed some promise but suffered from inconsistencies, particularly in model fitting. These findings suggest that ANN is the best choice for forecasting hydroelectric power generation, while other models may require further optimisation to improve their reliability.

B. Error Distribution and Model Robustness

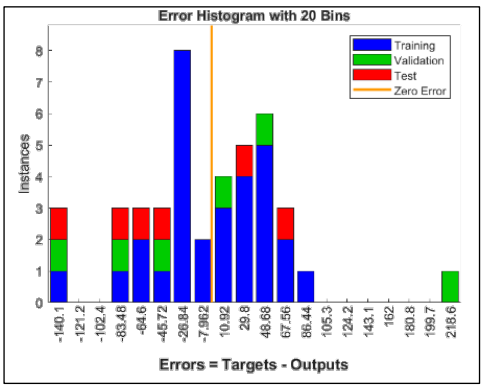
This subsection explores the distribution of prediction errors to assess how consistently each model performs across different datasets. Error histograms offer a visual perspective on the spread of prediction errors, which can help indicate the presence of outliers, potential overfitting, and the general behaviour of each model across datasets. The comparative error distributions of all four models are shown in Fig. 2.

Compared to the other models, ANN exhibited the most balanced error distribution, with errors predominantly concentrated within a narrow range of -90 to 100. A well-defined peak near zero error suggests strong predictive accuracy and minimal deviation from actual values. In contrast, SVM displayed the broadest error distribution, with errors extending up to ±1500 and clustering around -500 and 500, indicating weaker predictive precision. Random Forest and XGBoost showed moderate error concentration, with most errors falling within -500 to 500, but lacked the tight distribution seen in ANN.

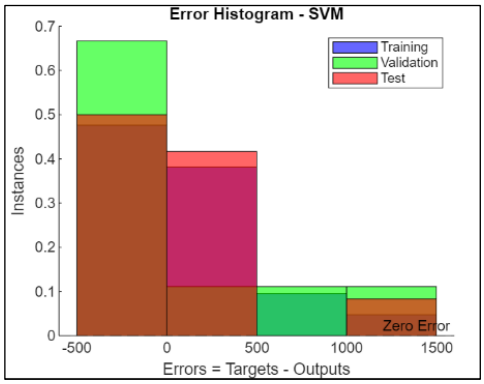
ANN demonstrated a stronger generalisation capability, as evidenced by its consistent error patterns across training, validation, and testing phases. The low presence of extreme outliers further supports its robustness in capturing patterns with minimal overfitting. Random Forest showed slight improvements over SVM but still contained significant outliers, reducing its overall reliability. SVM struggled with generalisation, as its validation and test errors exhibited high variance, leading to unreliable predictions. XGBoost provided better robustness than SVM and Random Forest, but its error spread remained wider than ANN, suggesting some inconsistencies in generalisation.

The high concentration of errors around zero in ANN indicates a superior ability to capture complex relationships within the data. In this regard, XGBoost performed better than SVM and Random Forest, but had a lower density in the zero-error bin, reflecting less precision in modelling the data's underlying structure. While slightly better than SVM, Random Forest still showed a long-tailed error distribution, suggesting occasional large deviations from actual values. These results indicate that ANN is the most precise model, followed by XGBoost, while SVM and Random Forest exhibit lower accuracy in capturing data trends.

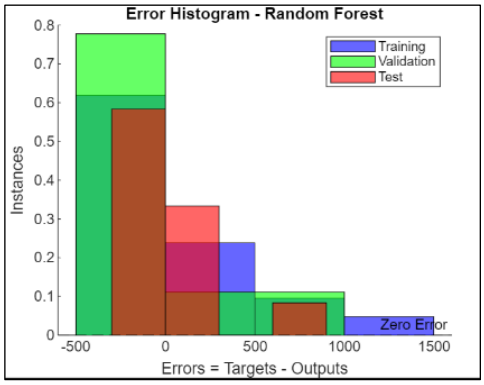
Considering error distribution and robustness, ANN emerged as the most reliable model, maintaining low error spread and strong generalisation across datasets. XGBoost demonstrated moderate performance but lacked the precision to match ANN's effectiveness. Random Forest showed slight improvements over SVM, but its inconsistent error distribution limited its reliability. SVM exhibited the weakest performance, with high prediction errors and poor generalisation, making it the least suitable choice for this forecasting task.



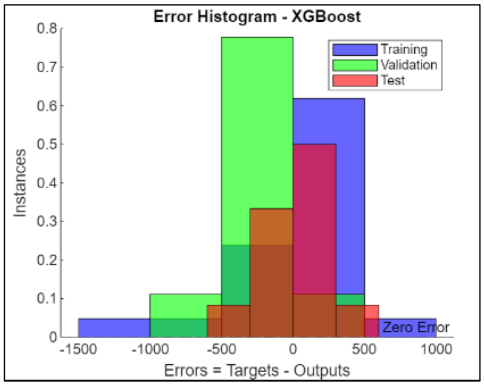
(a) ANN



(b) SVM

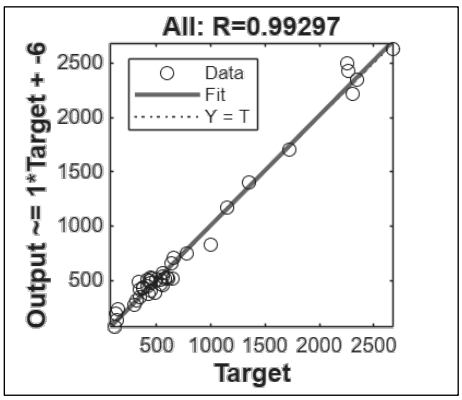


(c) Random Forest

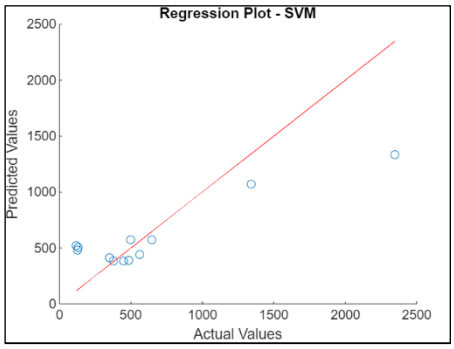


(d) XGBoost

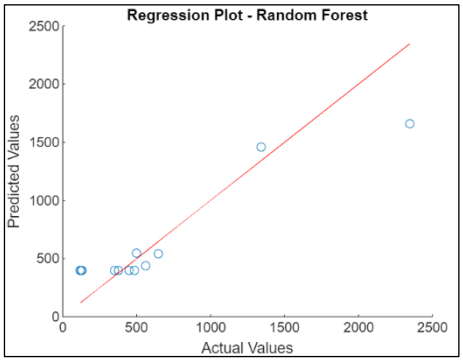
Fig. 2. Error Histogram: (a) ANN, (b) SVM, (c) Random Forest and (d) XGBoost



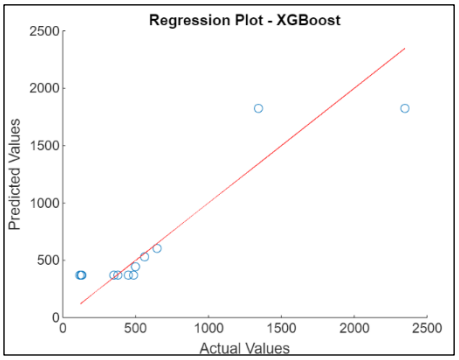
(a) ANN



(b) SVM



(c) Random Forest



(d) XGBoost

Fig. 3. Regression Plot: (a) ANN (b)SVM (c) Random Forest (d) XGBoost

C. Insights into Regression Plots

To further assess the model's ability to predict, regression plots are used to visualise the alignment between expected and actual hydroelectric generation values. These plots reveal how well each model captures underlying data patterns. Fig. 3 illustrates the regression performance of all models.

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

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 handling nonlinear patterns but fell short of ANN's precision and consistency.

Among all models, ANN emerged as the most reliable, with high alignment to the perfect-fit line, minimal scatter, and strong generalisation across different values of hydroelectric generation. XGBoost provided better predictive alignment than SVM and Random Forest but lacked ANN's precision. Random

Forest showed moderate accuracy but struggled with higher generation values, while SVM exhibited the weakest predictive performance due to significant deviations from the regression line.

V. Conclusion

This study used four artificial intelligence-based models, ANN, SVM, Random Forest, and XGBoost, to forecast Malaysia's hydroelectric power generation using historical socioeconomic data from 1980 to 2021. The models were assessed using multiple performance metrics, including MSE, R, regression plots, error histograms, and learning curves. Among the models evaluated, the ANN model showed the most promising results, achieving a testing MSE of $1.1541 \times 10^4$and an R value of 0.9962. These outcomes suggest that ANNs offer stronger predictive capabilities in capturing the nonlinear relationships between socioeconomic factors and hydropower generation. While XGBoost and Random Forest also demonstrated reasonable performance, SVM showed lower accuracy in this context. Overall, the findings support the potential of AI-based models, particularly ANN, as valuable tools to assist in forecasting tasks related to renewable energy, offering a data-driven perspective that could complement traditional planning methods.

VI. Future Recommendations

Future research could benefit from expanding the range of input variables to include climatic and hydrological factors such as rainfall, river discharge, and reservoir levels, which are directly linked to hydropower generation. Incorporating these variables may help optimise model performance and improve predictive accuracy. Additionally, exploring hybrid or ensemble modelling approaches that combine the strengths of multiple algorithms may offer further improvements in forecasting reliability. Regional or plant-specific datasets are also recommended to yield more localised insights, which can be valuable for site-level energy planning. From a policy standpoint, energy planners are encouraged to adopt AI-based forecasting tools as part of the decision-making process. These tools can support more accurate energy projections and facilitate strategic planning, contributing to Malaysia's broader goals of energy diversification, low-carbon development, and long-term sustainability in the power sector.

References

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

[2] M. H. Hazmin, F. Mustapha "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, vol. 11, no. 1, pp. 1–24, 2024.

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

[4] J. Rajarajeswaran, "Applications of Artificial Intelligence and Machine Learning for Accurate Forecasting and Optimization of Renewable Energy Generation," in 3rd International Conference on Electronics and Renewable Systems, ICEARS 2025 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 1886–1889. doi: 10.1109/ICEARS64219.2025.10940813.

[5] F. Muhammad, A. Qayyum, A. A. Bawazir, M. Jan, and N. Ahmed, "Assessing the Tri-Dimensional Nexus of Energy, Environment, and Economic Growth in Pakistan: An Empirical Study," International Journal of Energy Economics and Policy, vol. 14, no. 4, pp. 329–343, Jul. 2024, doi: 10.32479/ijeep.16128.

[6] R. N. Hasanah, H. Suyono, J. Kim, Y. Muharram, and E. Muljadi, "Hydropower Development Towards a Full-Renewable Energy Grid in Indonesia," in 2023 IEEE Industry Applications Society Annual Meeting, IAS 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/IAS54024.2023.10406491.

[7] M. Alshammari, "Hydroelectric Energy: Challenges, Solutions and Future Trends," in International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICECET55527.2022.9873025.

[8] M. Ersan and E. Irmak, "Governor Control Systems in Hydroelectric Power Plants: Overview, Challenges, and Recommendations," in 11th International Conference on Smart Grid, icSmartGrid 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/icSmartGrid58556.2023.10170958.

[9] D. Kumar, A. Kumar, S. Bhattacharya, S. Lakra, M. P. Singh, and N. Roy, "Impact of Hydro Generation Outage due to High Silt in Northern Region of India," in 2024 IEEE International Conference on Power and Energy, PECon 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 145–150. doi: 10.1109/PECON62060.2024.10827293.

[10] S. Di Grande, R. Gueli, M. Berlotti, and S. Cavalieri, "Harnessing multivariate AI to enhance hydropower generation forecasting," in Proc. 2024 AEIT Int. Annu. Conf. (AEIT), Sept. 2024, pp. 1–6. doi: 10.23919/AEIT63317.2024.10736754.

[11] H. Apaydin, H. Feizi, M. T. Sattari, M. S. Colak, S. Shamshirband, and K. W. Chau, "Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting," Water (Switzerland), vol. 12, no. 5, May 2020, doi: 10.3390/w12051500.

[12] V. Velasquez and W. Flores, "Machine Learning Approach for Predictive Maintenance in Hydroelectric Power Plants," in 2022 IEEE Biennial Congress of Argentina, ARGENCON 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ARGENCON55245.2022.9939782.

[13] G. Shahgholian, M. Moazzami, S. M. Zanjani, A. Mosavi, and A. Fathollahi, "A Hydroelectric Power Plant Brief: Classification and Application of Artificial Intelligence," in SACI 2023 - IEEE 17th International Symposium on Applied Computational Intelligence and Informatics, Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 141–146. doi: 10.1109/SACI58269.2023.10158597.

[14] A. Dubey, A. Tiwari, A. Bhardwaj, S. Sivamohan, and S. Krishnaveni, "Global Energy Consumption Patterns and Optimization using Big Data," in Proceedings of the 9th International Conference on Communication and Electronics Systems, ICCES 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1025–1028. doi: 10.1109/ICCES63552.2024.10859370.

[15] S. Chowdhury and D. K. Das, "Renewable Energy Trends in Future Prospect Analyzed by XGBoost- A Time Series Analysis," in 4th International Conference on Sustainable Expert Systems, ICSES 2024 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 134–138. doi: 10.1109/ICSES63445.2024.10763177.

[16] S. N. Fard and M. M. Ardehali, "Long Term Forecasting of Electrical Energy Consumption for Iran Based on Optimized Artificial Neural Networks and Socio-Economic Indicators Data," in 2023 5th International Conference on Optimizing Electrical Energy Consumption, OEEC 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 40–44. doi: 10.1109/OEEC58272.2023.10135403.

[17] S. Kumari, S. Sreekumar, S. Singh, and D. P. Kothari, "Comparison among ARIMA, ANN, and SVR Models for Wind Power Deviation Charge Reduction," in 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 551–557. doi: 10.1109/COM-IT-CON54601.2022.9850913.

[18] T. Paul et al., "Forecasting of Reservoir Inflow by the Combination of Deep Learning and Conventional Machine Learning," in IEEE International Conference on Data Mining Workshops, ICDMW, IEEE Computer Society, 2021, pp. 558–565. doi:

10.1109/ICDMW53433.2021.00074.

[19] S. Wang and J. Ma, "A Novel Ensemble Model for Load Forecasting: Integrating Random Forest, XGBoost, and Seasonal Naive Methods," in Proceedings - 2023 2nd Asian Conference on Frontiers of Power and Energy, ACFPE 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 114–118. doi: 10.1109/ACFPE59335.2023.10455494.

[20] H. Luo and X. Li, "A hybrid model of CNN-BiLSTM and XGBoost for HVAC systems energy consumption prediction," in 2023 5th International Conference on Industrial Artificial Intelligence, IAI 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/IAI59504.2023.10327594.

[21] R. Gao, A. Zielinski, and T. Hong, "Regularized probabilistic forecasting of electricity wholesale price and demand," IEEE Trans. Power Syst., vol. 36, no. 3, pp. 2301–2312, May 2021. doi: 10.1109/TPWRS.2020.3038256.

[22] L. Chen and X. Lai, "Comparison between ARIMA and ANN models used in short-term wind speed forecasting," in Proc. Asia-Pacific Power and Energy Engineering Conf. (APPEEC), Wuhan, China, Mar. 2011, pp. 1–4. doi: 10.1109/APPEEC.2011.5748446.

[23] S. Wang, F. Liu, and X. Niu, "Research and Application of Random Forest and ARIMA Based Modelling," in 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering, ICSECE 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1469–1473. doi: 10.1109/ICSECE61636.2024.10729344.

[24] N. Sato, Y. Fukuyama, T. Iizaka, and T. Matsui, "A Correntropy-Based Artificial Neural Network using Early Stopping for Daily Peak Load Forecasting," in Proc. 59th Annu. Conf. Soc. Instrum. Control Eng. (SICE), Chiang Mai, Thailand, Sept. 2020, pp. 581–586. doi: 10.23919/SICE48898.2020.9240336.

[25] R. Mathumitha, P. Rathika, and K. Manimala, "SVM-based regression for forecasting building power energy consumption using smart meter data," in 2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCNT56998.2023.10307674.

[26] Q. Wang, H. Wang, C. Gupta, A. R. Rao, and H. Khorasgani, "A Non-linear Function-on-Function Model for Regression with Time Series Data," in Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 232–239. doi: 10.1109/BigData50022.2020.9378087.

[27] H. Li, Q. Zhou, J. Tian, and X. Lin, "Energy demand forecasting for an office building based on random forests," in 2020 IEEE 4th Conference on Energy Internet and Energy System Integration: Connecting the Grids Towards a Low-Carbon High-Efficiency Energy System, EI2 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 29–32. doi: 10.1109/EI250167.2020.9347021.

[28] Y. S. Kim, M. K. Kim, N. Fu, J. Liu, J. Wang, and J. Srebric, "Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models," Sustain Cities Soc, vol. 118, Jan. 2025, doi: 10.1016/j.scs.2024.105570.

[29] X. Du, "Prediction of Power Consumption of Hydroelectric Power Station by Levenberg-Marquardt-BP Algorithm," in Proceedings - 2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 43–46. doi: 10.1109/AINIT54228.2021.00017.