Taylor & Francis
Taylor & Francis Group

Engineering Applications of Computational Fluid Mechanics

ISSN: 1994-2060 (Print) 1997-003X (Online) Journal homepage: www.tandfonline.com/journals/tcfm20

Optimisation of energy performance of axial-flow pump based on conditional tabular generative adversarial network enhancement learning method

Kan Kan , Siqi Gao , Ye Li , Zhaodan Fei , Hui Xu , Jinbo Chen & Mosè Rossi

To cite this article: Kan Kan , Siqi Gao , Ye Li , Zhaodan Fei , Hui Xu , Jinbo Chen & Mosè Rossi (2026) Optimisation of energy performance of axial-flow pump based on conditional tabular generative adversarial network enhancement learning method, Engineering Applications of Computational Fluid Mechanics, 20:1, 2679809, DOI: 10.1080/19942060.2026.2679809

To link to this article: https://doi.org/10.1080/19942060.2026.2679809

Published online: 29 May 2026.

ENGINEERING APPLICATIONS OF COMPUTATIONAL FLUID MECHANICS
2026, VOL. 20, NO. 1, 2679809
https://doi.org/10.1080/19942060.2026.2679809

Optimisation of energy performance of axial-flow pump based on conditional tabular generative adversarial network enhancement learning method

Kan Kan[a,b], Siqi Gao[a], Ye Li[c], Zhaodan Fei[b,d], Hui Xu[b,e], Jinbo Chen[f] and Mosè Rossi[g]

[a]School of Electrical and Power Engineering, Hohai University, Nanjing, People's Republic of China; [b]National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing, People's Republic of China; [c]Gezhouba Hydropower Station, China Yangtze Power Co., Ltd., Yichang, People's Republic of China; [d]Institute of Water Science and Technology, Hohai University, Nanjing, People's Republic of China; [e]College of Agricultural Science and Engineering, Hohai University, Nanjing, People's Republic of China; [f]College of Engineering and Technology, Department of Engineering, East Carolina University, Greenville, USA; [g]Department of Industrial Engineering and Mathematical Sciences (DIISM), Marche Polytechnic University, Ancona, Italy

ABSTRACT

In the optimisation of axial-flow pumps, common challenges include the need for large training datasets, high numerical simulation costs, and limited accuracy in fast performance prediction. To overcome these issues, this work proposes an integrated optimisation framework combining a conditional tabular generative adversarial network (CTGAN), a backpropagation neural network (BPNN), the rime optimisation algorithm (RIME), and a standard genetic algorithm (SGA) to improve pump energy performance. Key geometric parameters of an axial-flow pump with a specific rotational speed of 1030 are extracted from numerical simulations to build the parametric model. An initial dataset of 4000 samples is generated through random sampling and evaluated via numerical simulation. CTGAN is then used for data augmentation, producing 1912 high-quality synthetic samples. The entire data-generation and filtering process is completed within seconds, reducing computational effort by approximately 3824 h on 36 CPU cores. The augmented dataset enables the construction of a more accurate energy-performance prediction model using BPNN coupled with RIME, reducing the mean squared error from 1.015 × 10−4 to 7.88 × 10−5. Finally, global optimisation using SGA further enhances pump performance. Compared with the preliminary optimised design, the final optimised axial-flow pump achieves a 0.88% increase in hydraulic efficiency, along with improved internal flow structures and a marked reduction in vortex intensity and turbulent kinetic energy near guide-vane walls. This framework demonstrates an efficient and reliable approach for intelligent design, performance prediction, and optimisation of axial-flow pumps, significantly reducing computational cost while improving accuracy and hydraulic performance.

Highlights

- An integrated framework combining CTGAN, BPNN, RIME, and SGA is proposed.
- The integrated framework significantly reduces the usage of computing resources.
- The optimized pump achieves 0.88% higher efficiency with improved internal flow.
- The improvement mechanisms are revealed by streamlines, vorticity, and TKE.

ARTICLE HISTORY
Received 8 January 2026
Accepted 16 May 2026

KEYWORDS
Axial-flow pump optimisation; data augmentation; standard genetic algorithm; hydraulic efficiency improvement; performance prediction

Nomenclature

f	Airfoil chord length
t	Cascade pitch
r	The radius of the current leaf grid
Z	The number of blades.
σ	Cascade solidity
σ_r, σ_s	Cascade solidity at *the* impeller blade and the guide vane blade
V_1, V_2, V_3	Absolute velocity at the impeller inlet, impeller outlet and guide vane outlet
W_1, W_2, W_3	Relative velocity at the impeller inlet, impeller outlet and guide vane outlet
U_1, U_2	Circumferential velocity at the impeller inlet and outlet
V_{u2}	Absolute velocity circumferential component at the impeller outlet

CONTACT Kan Kan kankan@hhu.edu.cn School of Electrical and Power Engineering, Hohai University, Nanjing, 211100, People's Republic of China; National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing, 210098, People's Republic of China

$\beta_1, \beta_2, \beta_3$	Relative angles of the relative velocity component at the impeller inlet, impeller outlet and guide vane outlet
α_2	Relative angles of the absolute velocity component at the impeller outlet
$R_1 \sim R_6$	Column radius 1–6
a	Vortex coefficient
β_{r1}, β_{r2}	Inlet angle and outlet angle at the impeller blade camber line
$\varphi_{r1}, \varphi_{r2}$	Inlet and outlet auxiliary angle at the impeller blade camber line
γ_r	Setting angle at the impeller blade
C_{rsa}, C_{ssa}	Setting angle coefficient at the impeller blade and the guide vane blade
$\mathrm{P}_{r1} \sim \mathrm{P}_{r9}$	Control points of impeller blade camber line
$c_{r1} \sim c_{r4}$	Control coefficients of impeller blade camber line
T_r, T_s	Maximum thickness coefficient at the impeller blade and the guide vane
x_s, y_s	Coordinates of airfoil suction surface
x_p, y_p	Coordinates of airfoil pressure surface
x_c, y_c	Coordinates of airfoil camber line
$a_0 \sim a_4$	Prescribe factors
BC	Original length of the two-dimensional airfoil in the top view
A	The point on the airfoil that needs to be projected
S_0	The central projection point of the point A
S_1	The point of the fully adherent projection point
S_2	Moving point on the arc $\mathrm{S}_0\mathrm{S}_1$
P_s	Projection coefficient at the guide vane blade
P_r	Projection coefficient at the impeller blade
θ_1	Variable projection angle
x_l, y_l	Coordinates of point S_2 after the variable projection

Abbreviations

CFD	Computational Fluid Dynamics
DOE	Design of Experiments
ESG	Environmental, Social and Governance
RAV	Roof-Attached Vortex
JAVD	Joint Anti-Vortex Device
SA	Simulated Annealing
PSO	Particle Swarm Optimisation
GAN	Generative Adversarial Network
CTGAN	Conditional Tabular Generative Adversarial Network
SST	Shear-Stress Transport
SIMPLEC	Semi-Implicit Method for Pressure-Linked Equations-Consistent
GCI	Grid Convergence Index
BPNN	Back Propagation Neural Network
RIME	Rime Optimisation Algorithm
SGA	Standard Genetic Algorithm
MSE	Mean Square Error
KPI	Key Performance Indicator
TKE	Turbulent Kinetic Energy

1. Introduction

Axial-flow pumps are the core mechanical equipment in China's large low-head pumping stations, which operate for exceptionally long annual durations and therefore consume substantial amounts of electricity. According to the Ministry of Water Resources of the People's Republic of China (2022)'s annual bulletins, many of these stations operate 3,500–5,000 h per year, and their electricity use represents one of the largest operational expenditures in water-resources management. The South-to-North Water Diversion Project alone reports over 2 TWh/year of electricity consumption for the Eastern Route pumping stations, highlighting how even small efficiency improvements translate into significant national energy savings (Jia et al., 2025). This is consistent with the China Water Affairs Group's 2023 ENVIRONMENTAL, SOCIAL and GOVERNANCE (ESG) report, which shows that electricity consumption for water supply averages 0.257 MWh/m^3, with energy costs accounting for 14.27% of total operating expenses (China Water Affairs Group Limited, 2025). In this context, improving the hydraulic efficiency of axial-flow pumps by even 1% could save tens of GWh of electricity annually and meaningfully reduce operating costs across China's water-transfer infrastructure. Consequently, the optimisation of axial-flow pump performance represents a key pathway for energy conservation, financial savings, and carbon-reduction goals in China's water-sector.

Compared with low specific-speed pumps such as centrifugal pumps (Chang et al., 2025), axial-flow pumps tend to incur a significantly higher fraction of hydraulic losses due to the strong three-dimensional secondary flows, tip-leakage vortices, and wall-bounded shear layers that develop inside the impeller and guide-vane passages. Their blades also feature highly twisted and spatially distorted geometries, making the control of incidence angles and circulation distribution far more complex than in centrifugal machines. Such geometric and flow-field complexities make improving hydraulic efficiency challenging. Zhang et al. (2025) showed that harmful sump vortices, especially the roof-attached vortex (RAV), strongly interact with axial-flow pump blades and generate unstable broadband pressure fluctuations that reduce efficiency. High-speed visualisation identifies a three-stage blade-vortex interaction process that amplifies vortex growth and energy loss. Installing a joint anti-vortex device (JAVD) significantly improves performance, reducing hydraulic loss by 0.18 m and increasing peak efficiency by 2.2%, while markedly suppressing roof pressure fluctuations. Al-Obaidi et al. (2025) showed that adjusting the blade angle from −3° to +3° strongly affects axial-flow pump stability, with +3° producing the highest pressure pulsations at rated conditions. As flow decreases from optimal to deep part-load, pulsation amplitudes first weaken and then intensify due to local unsteadiness and stronger rotor-stator interaction, with the distributor region experiencing the largest fluctuations. Blade-angle increases amplify instability at rated flow, while deep part-load operation causes more random pulsation behaviour. Further insights into the complex flow mechanisms governing pump performance have been gained through advanced analysis techniques, such as energy flow density and entropy production for cavitation characterisation (Yu et al., 2025), and reduced-order variational mode decomposition for understanding energy dissipation in cavitation vortices (Shen et al., 2025).

Two primary approaches exist for axial-flow blade design: the direct and inverse methods. The direct method designs the blade from predefined geometric parameters and can be formulated through three classical theories: 1D flow theory, 2D planar-cascade theory, and 3D S1/S2 surface theory (Dixon & Hall, 2013; Stepanoff, 1957; Wu, 1952). Among these, the 2D planar-cascade framework is most widely used in engineering practice because it provides a good balance between computational cost and hydraulic accuracy.

The inverse method, by contrast, prescribes the desired flow or pressure distribution and reconstructs the blade geometry, typically through back-propagation, ensuring that the resulting shape meets target performance requirements. Modern axial-flow pump design often combines both direct and inverse strategies, with particular emphasis on the 2D planar-cascade inverse problem method, which offers simple implementation and reliable performance (Borges et al., 1996; Tan et al., 1984). This method solves the velocity distribution on the airfoil surface based on preset hydraulic targets and is therefore adopted in this work.

During the design of axial-flow pumps, it is difficult to directly obtain accurate energy-performance characteristics, making further optimisation necessary. Existing optimisation strategies fall into two categories: empirical methods, which rely heavily on designer experience, and de-empirical methods, which use computational tools to guide parameter adjustments. With advances in computing, de-empirical approaches have become dominant and can be grouped into three types: (i) statistical methods such as Design of Experiments (DOE) (Uy & Telford, 2009), (ii) surrogate or approximate models, e.g. response surfaces (Khuri & Mukhopadhyay, 2010), Kriging (Kleijnen, 2009), hybrid models (Abate et al., 2010), and neural networks (Wu & Feng, 2018), and (iii) intelligent optimisation algorithms such as genetic algorithms (GA), simulated annealing (SA), and particle swarm optimisation (PSO) (Bertsimas & Tsitsiklis, 1993; Lambora et al., 2019; Wang et al., 2018; Yang, 2020).

Surrogate models reduce the need for repeated Computational Fluid Dynamics (CFD) simulations but require large, high-quality training datasets, while intelligent optimisation techniques often involve substantial computational cost due to iterative global searches. With the rapid development of machine learning, neural-network-based surrogate models have shown strong ability to capture complex nonlinear relationships in pump performance (Hinton & Salakhutdinov, 2006; Rumelhart et al., 1986). Therefore, this work integrates machine-learning-based surrogate modelling with intelligent optimisation to achieve more efficient and accurate axial-flow pump performance enhancement.

Recent advances in related fields have further demonstrated the potential of combining data-driven methods with optimisation techniques. For instance, synthetic data generation via generative adversarial networks has proven effective in augmenting limited datasets (Xu et al., 2019), while systematic hyperparameter optimisation has been shown to significantly improve predictive accuracy in medical image classification (Li et al., 2025; Nabeel et al., 2025) and image steganalysis (Bohang et al., 2025). These successes motivate the integration of similar strategies into pump performance optimisation, where data scarcity and model accuracy are critical challenges. In machine learning, data augmentation is crucial for improving model generalisation, particularly when datasets are limited or imbalanced. Common techniques include Mixup (Zhang et al., 2018), SMOTE (Chawla et al., 2002), GANs (Goodfellow et al., 2014), and variational auto encoders (VAEs) (Kingma & Welling, 2014). Among these, generative adversarial networks (GANs) generate high-quality synthetic samples through adversarial training. The conditional tabular GAN (CTGAN) (Xu et al., 2019), specifically designed for tabular data, learns the underlying statistical structure of datasets and produces realistic samples that capture complex feature relationships.

For axial-flow pump energy-performance prediction, CFD simulations (Chitrakar et al., 2020; Rong et al., 2024) typically yield tabular datasets influenced by nonlinear interactions among design parameters. Compared with traditional GANs, CTGAN introduces modal-adaptive normalisation and a variational Gaussian mixture mechanism, enabling accurate modelling of continuous, multimodal distributions, making it well suited for augmenting CFD-based performance datasets.

However, existing studies rarely address two critical limitations simultaneously: (i) the scarcity and high cost of generating high-quality CFD datasets for axial-flow pumps, and (ii) the limited accuracy of existing surrogate models in capturing complex, multi-peak performance landscapes. To close these gaps, this work proposes a novel intelligent optimisation framework for axial-flow pump energy performance that integrates CTGAN-based data augmentation, a hybrid BPNN-RIME predictor, and SGA global optimisation. The key innovations are:

1. CTGAN-enhanced data generation to expand limited CFD datasets while preserving multi-modal distributions.
2. A hybrid BPNN-RIME surrogate model, which improves prediction accuracy and reduces MSE compared with traditional models.
3. Global optimisation with SGA, enabling efficient exploration of multi-parameter design spaces.

The proposed framework integrates physical understanding with data-driven optimisation: the geometric design parameters derived from axial-flow pump theory (cylindrical–layer theory, velocity triangle analysis, and Bézier curve parameterisation) serve as inputs to the machine learning models. CTGAN captures their complex distributions, BPNN predicts performance, and SGA optimises the parameters, with the optimised design validated through CFD analysis demonstrating physical improvements in flow structures. This logical progression ensures that the machine learning components are built upon a solid hydraulic foundation, seamlessly linking pump flow physics with advanced computational methods. This framework achieves significant computational savings, reducing CFD time demand by more than 3,800 h while simultaneously improving energy-performance prediction.

The remainder of the paper is organised as follows: the second section introduces the axial-flow pump design theory; the third section describes the numerical calculation setup and carries on the experimental validation; the fourth section is based on the CTGAN-based axial-flow pump energy performance prediction model; the fifth section trains and re-optimises the CTGAN-based axial-flow pump energy performance prediction model; and the sixth section reports the conclusions of the work.

2. Axial-Flow pump blade design theory

In this work, a unified design methodology is applied to both the impeller and the guide vane of the axial-flow pump, as the aerodynamic and hydraulic principles governing their geometry are closely related. For clarity, this section presents the design theory using the impeller as the representative component (Adding "r" to the subscript of the relevant variables to represent the impeller section), while the same procedures and formulations are subsequently applied to the guide vanes. The following subsections introduce the fundamental assumptions, two-dimensional cascade transformations, and associated design strategies that form the basis for constructing high-performance axial-flow pump blades.

2.1. Cylindrical-Layer unfolding for 2D blade design

In this study, the cylindrical-layer irrelevance hypothesis is adopted, whereby the three-dimensional flow inside an axial-flow pump is approximated as a set of concentric cylindrical flows rotating around the axis, and fluid particles on adjacent cylindrical surfaces are assumed not to interact. This assumption decomposes the complex three-dimensional flow field into several independent two-dimensional blade-to-blade flow problems, substantially reducing computational complexity and improving the efficiency of hydraulic analysis. Based on this hypothesis, six concentric cylindrical surfaces with radii R_1-R_6 are defined, as shown in Figure 1. The selection of six surfaces represents a balance between computational cost and design accuracy, ensuring sufficient resolution to capture the spanwise geometric variation while maintaining efficiency. Each cylindrical surface is then cut and unfolded into a planar blade-to-blade (streamwise) view, generating a series of equally spaced airfoil sections along the circumferential direction of the unfolded cylindrical surface. These sections form a two-dimensional planar cascade at each radial location and serve as the basis for subsequent blade design and flow-field analysis.

Each airfoil has a specific geometry and hydrodynamic properties in this cascade as shown in Figure 2, where f is the airfoil chord length and t is the cascade pitch. The t is expressed as follows:

$$t = \frac{2\pi r}{Z} \tag{1}$$

Where r is the radius of the current cascade and Z is the number of blades.

The ratio of f to t is used to characterise the degree of blade densification, which is defined as the cascade

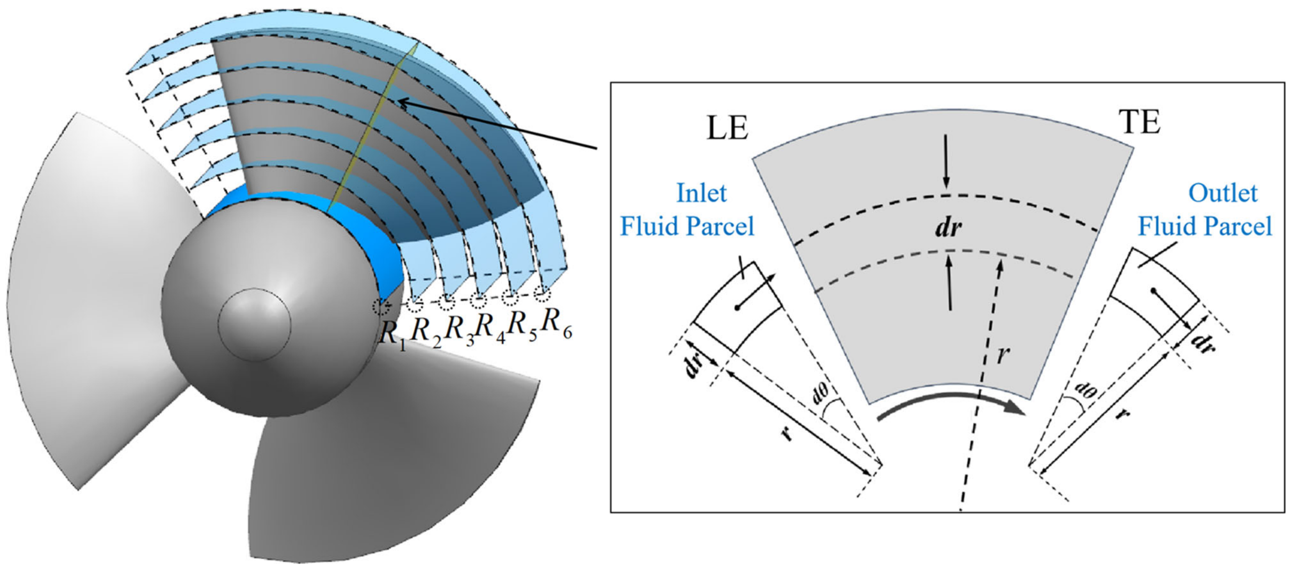

Figure 1. Schematic of the cylindrical-layer irrelevance hypothesis used for 2D cascade construction.

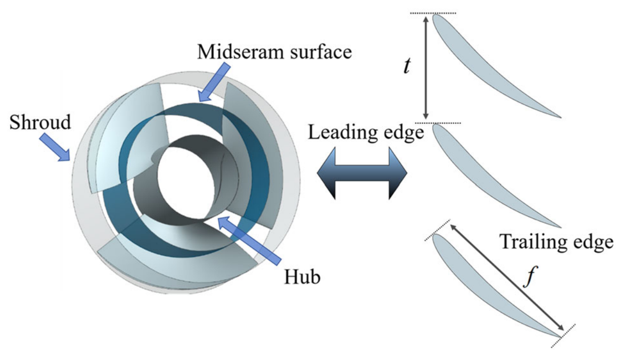

Figure 2. Schematic of blade-to-blade grid density distribution in the 2D planar cascade.

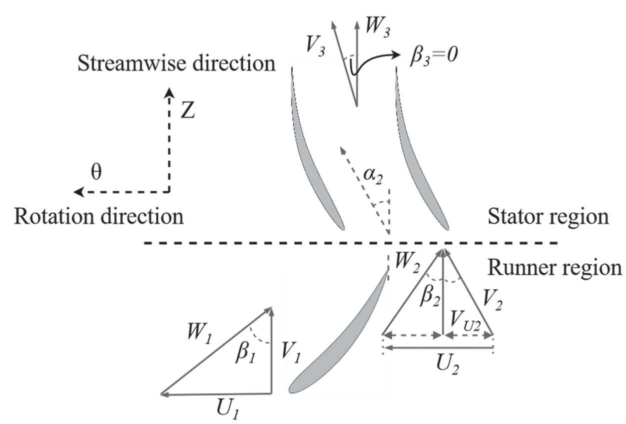

Figure 3. Schematic of the velocity triangles in the unfolded 2D blade cascade.

solidity σ with the expression:

$$\sigma = \frac{f}{t} \tag{2}$$

The cascade solidity σ is one of the key geometric parameters in the design of impeller, which has a direct impact on the hydraulic performance and cavitation performance of the axial-flow pump.

2.2. *Velocity triangle and variable vortex assumption*

The flow on the cylindrical surface of the axial-flow pump impeller is a compound motion, and the absolute velocity $\vec{V}$ can therefore be decomposed into Eq.(3) such as (Thin et al., 2008):

$$\vec{V} = \vec{W} + \vec{U} \tag{3}$$

Where $\vec{W}$ is the relative velocity and $\vec{U}$ is the circumferential velocity.

To further analyze the flow characteristics, the cylindrical layer with column radius R_i (i is $1 \sim 6$) is expanded in the plane of cylindrical coordinate system to obtain the velocity distribution characteristics of the two-dimensional planar cascade as shown in Figure 3, where 1, 2, and 3 represent the velocity at the impeller inlet, impeller outlet, and guide vane outlet, respectively. It is assumed that the direction of $\vec{W}$ becomes tangent to the airfoil surface in the limit of an infinite number of blades.

There are three kinds of vortex assumptions for the calculation of initial velocity: the forced vortex assumption, the constant vortex assumption, and the free vortex assumption. By combining the three vortex assumptions described above, this study adopts the variable-vortex hypothesis for the initial velocity design, which is defined by the following rules:

$$r^a C_{u2} = \text{constant} \tag{4}$$

Where r is the radius of the calculation point, C_{u2} is the circumferential component of impeller outlet absolute velocity, and a is the vortex coefficient, $-1 \le a \le 1$.

In particular, when $a = -1$, Eq.(4) is the forced vortex assumption of the flow regime; when $a = 0$, Eq.(4) is the constant vortex assumption of the flow regime; and when $a = 1$, Eq.(4) is the free vortex assumption of the flow regime. The adjustment of the vortex coefficient facilitates the transition between the free vortex and forced vortex assumptions, thereby enabling the selection of the most suitable vortex assumption.

2.3. Enhanced parametric method for airfoil camber line design

Considering the strong effect of camber-line flexibility and maximum-camber position on airfoil performance (Zaffar et al., 2018), this work proposes an improved Bézier curve camber line design method (Kan et al., 2025a; Kan et al., 2025b), named as the fourth-order two-segment Bézier camber curve line design method, which allows explicit and independent control of the key maximum camber position via a dedicated control point (P_{r5}), combining targeted geometric flexibility with optimisation efficiency. As shown in Figure 4, β_1 is the inlet angle of the impeller blade camber line; β_2 is the outlet angle of the impeller blade camber line; γ_r is the impeller blade setting angle, i.e. the airfoil's angle of attack relative to the incoming flow; and φ_{r1} and φ_{r2} are the inlet and outlet auxiliary angles, respectively. To further parameterise and precisely control the amplitude of the impeller blade setting angle γ_r, two adjustment coefficients are introduced: the impeller blade setting angle coefficient C_{rsa} and the guide vane blade setting angle coefficient C_{ssa}, both ranging from 0 to 1. The expression of the blade setting angle γ_r is defined as:

$$\gamma_r = \left(\frac{\pi}{2} - \beta_{r2}\right) + Crsa(\beta_{r2} - \beta_{r1}) \tag{5}$$

$$\frac{\pi}{2} - \beta_{r2} < \gamma_r < \frac{\pi}{2} - \beta_{r1} \tag{6}$$

The angular relationship can be seen in Eqs.(7)-(8) and Figure 4:

$$\varphi_{r1} = \gamma_r + \beta_{r1} - \frac{\pi}{2} \tag{7}$$

$$\varphi_{r2} = \frac{\pi}{2} - \gamma_r - \beta_{r2} \tag{8}$$

The airfoil camber line is calculated using fourth-order Bézier curve and divided by two segments. The Segment 1 extends from the inlet point $P_{r1}(0,0)$ to the point P_{r5}, while the Segment 2 lengthens from the point P_{r5} to the outlet point $P_{r9}(1,0)$. The point P_{r5} marks the maximum

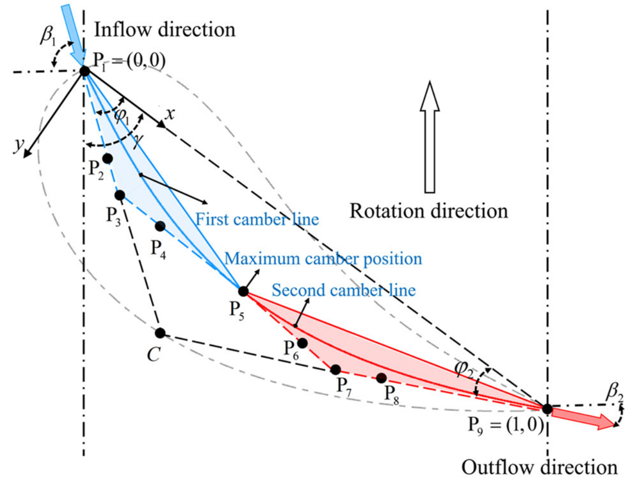

Figure 4. Construction of a fourth-order, two-stage Bézier curve.

camber position, and the coordinates of the point C are obtained from Eq.(8):

$$C = \left(\frac{\tan(\varphi_{r2})}{\tan(\varphi_{r1}) + \tan(\varphi_{r2})}, \frac{\tan(\varphi_{r1})\tan(\varphi_{r2})}{\tan(\varphi_{r1}) + \tan(\varphi_{r2})}\right) \tag{9}$$

Where the point P_{r3} is on the line segment $P_{r1}C$; therefore, the point P_{r3} is expressed as:

$$P_{r3} = (C - P_{r1})c_{r1} + P_{r1} \tag{10}$$

Where c_{r1} is the control factor for the point P_{r3}.

To ensure that the point P_{r5} marks the maximum camber position, the point P_{r7} is expressed as:

$$P_{r7} = (C - P_{r9})c_{r1} + P_{r9} \tag{11}$$

The point P_{r5} related to the maximum camber position is located on the line segment $P_{r3}P_{r7}$; therefore, the point P_{r5} is expressed as:

$$P_{r5} = (P_{r3} - P_{r7})c_{r2} + P_{r7} \tag{12}$$

Where c_{r2} is the control factor for the point P_{r5}.

To increase the flexibility of the camber lines of Segment 1 and Segment 2, it is required to incorporate two additional control points, designated P_{r2} and P_{r4}, into the camber line of Segment 1. A similar approach is used for Segment 2, with the addition of two control points, named P_{r6} and P_{r8}, to the camber line of Segment 2. The expressions for the four points of P_{r2}, P_{r4}, P_{r6}, and P_{r8} are shown below:

$$P_{r2} = (P_{r3} - P_{r1})c_{r3} + P_{r1} \tag{13}$$

$$P_{r4} = (P_{r3} - P_{r5})(1 - c_{r3}) + P_{r5} \tag{14}$$

$$P_{r6} = (P_{r7} - P_{r5})c_{r4} + P_{r5} \tag{15}$$

$$P_{r8} = (P_{r7} - P_{r9})(1 - c_{r4}) + P_{r9} \tag{16}$$

Where c_{r3} is the control coefficient for points P_{r2} and P_{r4}, and c_{r4} is the control coefficient for points P_{r6} and P_{r8}. Segment 1 consists of five control points, P_{r1}, P_{r2}, P_{r3}, P_{r4}, and P_{r5}, and the camber line in the form of a Bézier curve is expressed as:

$$B_{r1}(x) = (1-x)^4 P_{r1} + 4(1-x)^3 x P_{r2} + 6(1-x)^2 x^2 P_{r3} + 4(1-x) x^3 P_{r4} + x^4 P_{r5} \tag{17}$$

Similarly, the Segment 2 includes five control points, such as P_{r5}, P_{r6}, P_{r7}, P_{r8}, and P_{r9}, and the camber line is expressed as:

$$B_{r2}(x) = (1-x)^4 P_{r5} + 4(1-x)^3 x P_{r6} + 6(1-x)^2 x^2 P_{r7} + 4(1-x) x^3 P_{r8} + x^4 P_{r9} \tag{18}$$

It is worth noting that the value of x in Eq.(17) ranges from 0 to 1. Additionally, all four control coefficients are defined within the range of 0–1. For the purposes of this study, each control coefficient is held constant at 0.5, and their variation is not considered in the subsequent analysis.

2.4. Thickness design method for 2D planar airfoils

In this work, the thickness distribution method of the NACA 4-series airfoil is adopted as the basis function (Asr et al., 2016; Kim et al., 2012). The classical NACA 4-series thickness distribution is adopted as the basis function due to its well-documented aerodynamic characteristics, extensive validation in turbomachinery contexts, and its suitability for parametric integration within the present optimisation framework. While modern, pump-specific distributions may offer potential gains, the NACA 4-series airfoil provides a robust and standardised baseline that can help to clearly highlight the influence of the primary geometric parameters under investigation. For each cylindrical section, the airfoil chord length is independently normalised to unit length, and its thickness distribution function y_t is expressed as follows:

$$y_t = 5T_r\left(a_0\sqrt{x} + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4\right) \tag{19}$$

Where T_r is the maximum thickness of the impeller blade (e.g. the maximum thickness of the guide vane is indicated by T_s); $a_0 \sim a_4$ is the prescribe factors, $a_0 = 0.2926$, $a_1 = -0.126$, $a_2 = -0.3516$, $a_3 = 0.2834$, and $a_4 = -0.1015$.

As reported in Figure 5, a schematic representation of the thickness distribution is superimposed on the airfoil camber line of the axial-flow pump. The thickness distribution function of the NACA 4-series airfoil is superimposed on the axial-flow pump airfoil camber line equation to obtain a representation of the axial-flow pump airfoil surface equation. The coordinate equation for the side of the suction surface is as follows:

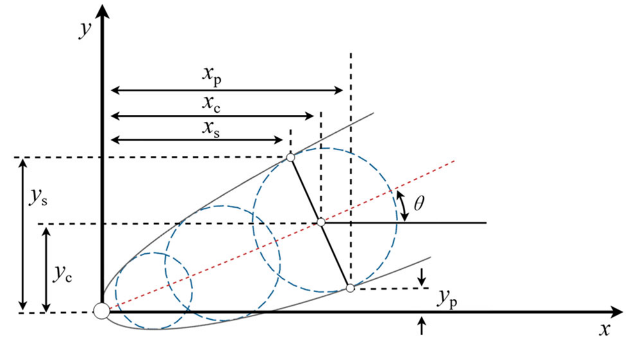

Figure 5. Schematic of the airfoil thickness superposition method.

$$\begin{aligned} x_s &= x_c - y_t \sin\alpha \\ y_s &= y_c + y_t \cos\alpha \end{aligned} \tag{20}$$

The coordinate equation of one side of the pressure surface is as follows:

$$\begin{aligned} x_p &= x_c + y_t \sin\alpha \\ y_p &= y_c - y_t \cos\alpha \end{aligned} \tag{21}$$

In Eqs.(20) and (21), (x_c, y_c) is expressed as the coordinate of the airfoil camber line, and α is the tangent angle at the camber line coordinate point.

To ensure a smooth and manufacturable trailing-edge geometry, a cubic Bézier curve is implemented to blend the pressure and suction surfaces near the trailing edge, replacing the finite-thickness cusp from the baseline NACA 4-series distribution.

2.5. Projection of 2D planar airfoils onto 3D cylindrical flow surfaces

Through the variable projection method, the obtained two-dimensional shape of the blade is projected back to the cylindrical flow plane as shown in Figure 6.

As shown in Figure 6, the straight line BC represents the initial length of the two-dimensional airfoil in the top view. The point A denotes the airfoil location to be projected. The point S_1 represents the fully adherent projection point, where the arc length BS_1 equals the straight-line length BA. The point S_0 represents the central projection point of the point A. The point S_2 a movable point on the arc S_0S_1, representing the variable

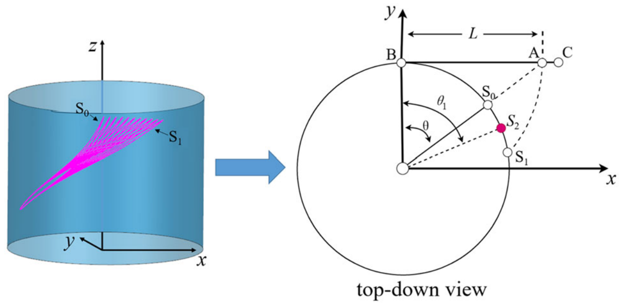

Figure 6. Principle of the variable projection method for mapping 2D airfoils onto cylindrical surfaces.

projection of the point A between the positions of S_0 and S_1. In ΔOBA, the projection angle θ is expressed as follows:

$$\theta = \arctan\left(\frac{L}{R}\right) \tag{22}$$

Where L is the length of the line BA before its projection, and R is the radius of the unfolded column surface. The arc BS_0 is the central projection of the line BA with length L_1:

$$L_1 = \theta \times R \tag{23}$$

The arc BS_2 is a variable projection of the straight line BA_1, whose length L_2 is between L_1 and L:

$$L_2 = L_1 + (L - L_1) \times P = \theta \times R + (L - \theta \times R) \times P \tag{24}$$

Where P is the projection coefficient, $0 \leq P \leq 1$. When $P = 0$, the arc BS_2 is centered to the line BA; and when $P = 1$, the arc BS_2 is a perfect fit projection. In the present design, both the impeller and the guide vane are assigned a constant value of projection coefficient along the radial direction, denoted as P_r and P_s respectively. This value remains uniform across all radial stations from hub to tip. The variable projection angle θ_1 is:

$$\theta_1 = \frac{L_2}{R} = \theta + \left(\frac{L}{R} - \theta\right) \times P \tag{25}$$

After the variable projection of the point A, the coordinates of the point S_2 are:

$$\begin{cases} x_l = R \times \sin(\theta_1) = R \times \sin\left(\theta + \left(\frac{L}{R} - \theta\right) \times P\right) \\ y_l = R \times \cos(\theta_1) = R \times \cos\left(\theta + \left(\frac{L}{R} - \theta\right) \times P\right) \end{cases} \tag{26}$$

It should be noted that, in this study, only the impeller projection coefficient P_r is considered as a variable in the subsequent optimisation, while the guide vane projection coefficient P_s is held constant at 0.5.

After the two-dimensional wing projection of each cylindrical surface, based on the three-dimensional wing obtained from the projection of each cylindrical surface, the smooth transition connection between the blade hub and the tip is realised by curve fitting, thus obtaining the complete three-dimensional solid model of the axial-flow pump blades. The impeller modelling process is shown in Figure 7(a), and the final three-dimensional model of the impeller and the guide vane is shown in Figure 7(b).

3. Numerical setup

Based on the axial-flow pump design methodology described in Section 2, this section presents the numerical simulation approach and experimental validation

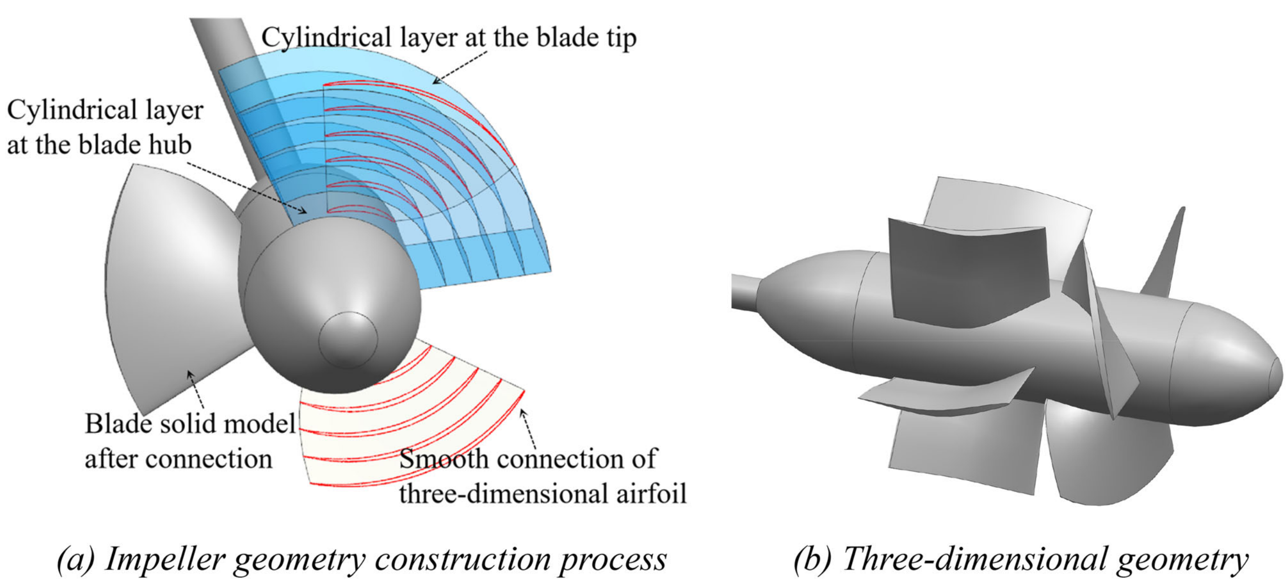

Figure 7. Overall geometric modelling process for the impeller and guide vane.

Table 1. Basic parameters of model pump.

Parameter	Value	Parameter	Value
n (r/min)	1450	D_{tip} (mm)	300
Q_m (kg/s)	401.64	D_{hub} (mm)	12
H (m)	4.8	Z_r	3
n_s	1030	Z_s	5

used to evaluate pump performance and internal flow characteristics. CFD simulations are performed to predict hydraulic efficiency, pressure distribution, and flow behaviour under different operating conditions.

3.1. Three-Dimensional model and boundary condition setup of the axial-flow pump

This work focuses on an axial-flow pump with a specific speed n_s of 1030, a rated rotational speed of 1450 rpm, a rated mass flow rate Q_m of 401.64 kg/s, and a rated head H of 4.8 m. From a design point of view, the impeller has got a blade number Z_r of 3 and guide vane blade number Z_s of 5. The specific parameters of the model pump are listed in Table 1.

The axial-flow pump model and the detailed mesh configuration are presented in Figure 8. The computational domain is discretized by using hexahedrons. Both impeller and guide vanes' meshes are generated with Turbo-Grid, while the remaining domain are meshed using ICEM. To ensure the inflow design condition, an inlet straight pipe with a diameter D_1 of 300 mm is used. The outlet flow channel features a 60° bend. The pressure monitoring planes at the pump segment's inlet and outlet are positioned $2D_1$ far from the impeller. In addition, the effect of the tip clearance is not considered since at the initial design and optimisation stage the energy performance of the axial-flow pump is predominantly influenced by the main blade geometry parameters considered in the present work.

3.2. Turbulence model, boundary conditions, and solution methods

Inlet and outlet boundary conditions are set as mass-flow ($Q_m = 401.64$ kg / s) and static pressure outlet, respectively, where the latter is imposed equal to 0 (gauge pressure). Boundary condition tests (varying outlet pressure and switching to pressure-inlet/flow-outlet) showed efficiency variations below 0.1%, confirming numerical robustness. The no-slip condition is applied to wall boundaries. The impeller is the rotational domain, while the others are stationary. The Shear-Stress Transport (SST) k-ω turbulence model is employed to simulate the complex turbulent flow inside the axial-flow pump. This model is widely adopted in the simulation of turbomachinery due to its comprehensive advantages in near-wall accuracy, separation flow prediction, and adaptability to rotational flows, making it well-suited for capturing the intricate three-dimensional flow structures within axial-flow pumps. This model combines the near-wall accuracy of the k-ω formulation with the robustness of the k-ϵ model in the free-stream region, enabling reliable prediction of boundary-layer development, flow separation,

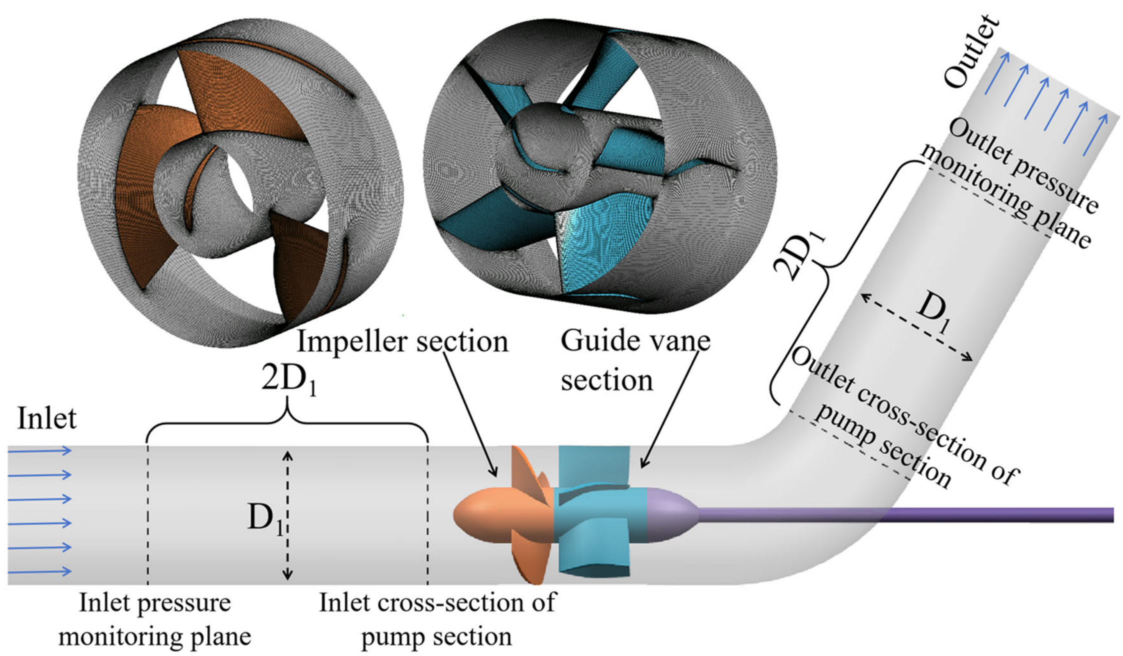

Figure 8. Geometry of the axial-flow pump used for numerical simulation.

and adverse pressure-gradient effects. When the same turbulence model is used consistently throughout the simulations, the numerical error in hydraulic efficiency is very small and practically negligible. The mean y^+ of the impeller wall is below 15, which lies within the applicable range for accurate wall-function treatment in the SST k-ω model. Owing to these advantages, the SST k-ω model is widely adopted in the numerical simulation of rotating machinery, including axial-flow pumps, where accurate resolution of near-wall flow structures is essential for performance evaluation (Li, 2025; Menter, 2002). The pressure-velocity coupling is calculated using the Semi-Implicit Method for Pressure-Linked Equations-Consistent (SIMPLEC) algorithm, and the residual convergence criterion is set to 10^{-5}.

3.3. Mesh independence study

In this study, a fully automated platform built with Python is established to perform high-quality modelling and mesh generation of the axial-flow pump via scripting in ICEM and TurboGrid, thereby eliminating variations in mesh topology. To ensure that the simulation results are not affected by the mesh resolution, the mesh independence study is performed. The Grid Convergence Index (GCI) method proposed by Roache (Roache, 1997; Stern et al., 2001; Tao & Wang, 2021) can assess the convergence and stability of the computational results by analyzing the numerical solution errors under different meshing conditions. Its defining formula is as follows:

$$\mathrm{GCI} = F_s \cdot \frac{e_a^{21}}{\lambda^p - 1} \tag{27}$$

Where F_s is the safety factor, indicating the degree of conservatism of the analysis. In the two-group grid calculations $F_s = 3$, while in the three-group grid calculations $F_s = 1.25$, reflecting higher convergence accuracy. The parameter p denotes the convergence accuracy, which is usually obtained by actual calculations; if it cannot be determined, then it is recommended to set it to an empirical value. In this work, $p = 1.97$.

The expression for the relative error e_a^{21} is:

$$e_a^{21} = \left| \frac{\varphi_1 - \varphi_2}{\varphi_1} \right| \tag{28}$$

Where φ_1 and φ_2 denote the numerical solution results obtained for different number of grids, respectively. The mesh refinement factor λ is calculated by using the following equation:

$$\lambda = \sqrt[3]{\frac{N_1}{N_2}} \tag{29}$$

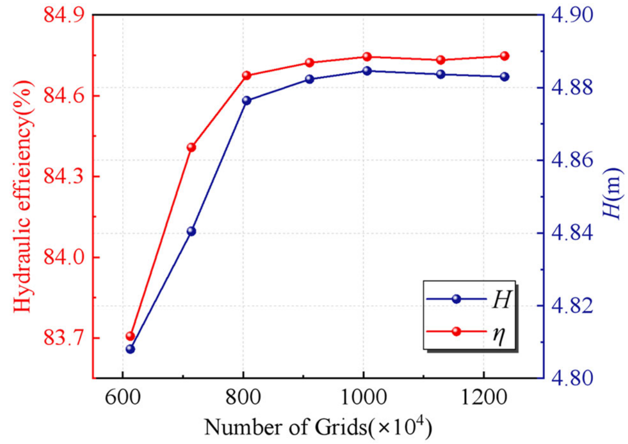

Figure 9. Numerical simulation results of the axial-flow pump under investigation.

Table 2. Grid Convergence Index (GCI) verification results.

Number of grids/10^4	φ	e_a^{21}	F_s	p	GCI/%
802	84.59	0.000945	1.25	1.97	1.39
908	84.67				
1003	84.69				
802	4.884	0.000614	1.25	1.97	0.90
908	4.887				
1003	4.889				

Where N_1 and N_2 are the numbers of nodes in the coarse and fine meshes, respectively. Figure 9 shows the numerical simulation results of seven numerical simulation the of axial-flow pump under investigation with different grid numbers. Design specifications listed in Table 1 are used in all the seven numerical simulations.

Three representative grid schemes are selected, whose grid numbers are 8.02, 9.08, and 10.03 million, respectively. When the grid number reaches eight million, the hydraulic efficiency and head of the axial-flow pump have stabilised, with the hydraulic efficiency differing by no more than 0.1% compared to the finer grid. Combined with the analysis in the previous subsection, the total error arising from boundary–condition variations (±0.1%) and mesh discretization (±0.1%) is conservatively estimated to be less than 0.2%. Moreover, in this study, a fully automated platform was established to perform high-quality modelling and mesh generation of the axial-flow pump via scripting, thereby eliminating variations in mesh topology. The head and hydraulic efficiency are used for the grid error analysis, and the results are shown in Table 2. The GCI is less than 3%, indicating that the numerical simulation results of the second grid scheme are converged. Considering the numerical accuracy and computational cost, the selected total number of grids is 9.08 million.

4. CTGAN-Based optimization framework for energy performance prediction of axial-flow pumps

This section presents the construction of the energy performance prediction model for the axial-flow pump based on a CTGAN, which is employed to augment the limited CFD dataset and enhance the representation of the nonlinear relationships between design parameters and performance indicators. On this basis, a machine-learning-based surrogate model is developed and trained to predict pump energy performance efficiently, providing the foundation for subsequent optimisation.

4.1. Energy performance prediction model for the axial-flow pump

A high-precision performance prediction model for axial-flow pumps constitutes a pivotal component for the intelligent optimisation method; thus, it is necessary to identify and select appropriate neural network algorithms and design an effective network structure to build up an efficient prediction model. The back propagation neural network (BPNN) is a classical feed-forward neural network model, whose fundamental premise is to optimise the network parameters through error back propagation. The BPNN typically comprises an input layer, a hidden layer, and an output layer. The input layer is responsible for receiving external data, while the output layer generates predictions. The hidden layer plays a crucial role in transforming the data nonlinearly through the utilisation of weights, biases, and activation functions. During the training process, the data undergoes a series of transformations. Initially, they are transferred from the input layer to the output layer. In each layer, the input data undergoes a weighted summation and is processed by an activation function. In essence, the resultant output is the culmination of these processes. Meanwhile, the network calculates the error between the predicted value and the actual value, reverses the error signal from the output layer to the input layer, and gradually adjusts the weights and biases of each layer to minimise the error. This process is realised by the gradient descent method until the network performance reaches the expectation. In forward propagation, the operation of each layer is specified as follows:

$$z_j = \sum_{i=1}^{n} w_{ij} x_{ij} + b_j \tag{30}$$

$$X_j = f(z_j) \tag{31}$$

Where i represents the number of input nodes, j represents the number of layers, w represents the weight value, b represents the bias value, X_j represents the input matrix for each layer, and f represents the activation function.

The goal of the backward propagation is to compute the gradient of the loss function with respect to each weight and adjust the weights according to the gradient to minimise the error. Assuming that the result of the output layer is y_j, the error function is defined as:

$$E(w, b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - d_i)^2 \tag{32}$$

Where $E(w,b)$ represents the error function with respect to the bias and weight values, d represents the real value of the output, which can be used to calculate the error.

The BPNN uses a continuous adjustment of the weight and bias values to minimise the error. The gradient descent method is usually used to optimise the error function until it reaches a minimum value. In the gradient descent method, the learning rate (η_{loss}) is a pivotal hyper parameter that controls the step size of each update of weights and bias. In practice, the selection of an appropriate learning rate and number of training rounds (epochs) often relies on experience, or it is determined by iterative trials. Given that different tasks, datasets, and model structures have different convergence characteristics, manually adjusting the learning rate is time-consuming. Moreover, extensive hyper parameter tuning can impose a substantial computational burden on complex models or large-scale datasets. In addition, the conventional fixed learning rate is typically incapable of adapting to the dynamic changes that occur during the training process. As training progresses and the model approaches the optimal solution, it is imperative to gradually reduce the parameter update step size, otherwise the optimal solution may be missed, or oscillations may occur due to large step sizes. This makes the process of manually adjusting the learning rate complex and rigid.

To address these challenges, numerous optimisation algorithms have been developed, including the RIME optimisation algorithm (RIME), which incorporates an adaptive learning rate mechanism. The RIME algorithm makes flexible adjustments at different training stages by adaptively adjusting the learning rate to the performance of the current parameters in each iteration. At the beginning of training, a larger learning rate facilitates the acceleration of the parameter updates. It has been demonstrated that the learning rate undergoes a reduction when approaching the optimal solution; then, this adjustment is implemented to ensure convergence to the global optimal solution. This mechanism circumvents the inefficiency and instability that may ensue from the manual selection of the learning rate, thereby improving the training speed and circumventing the issues

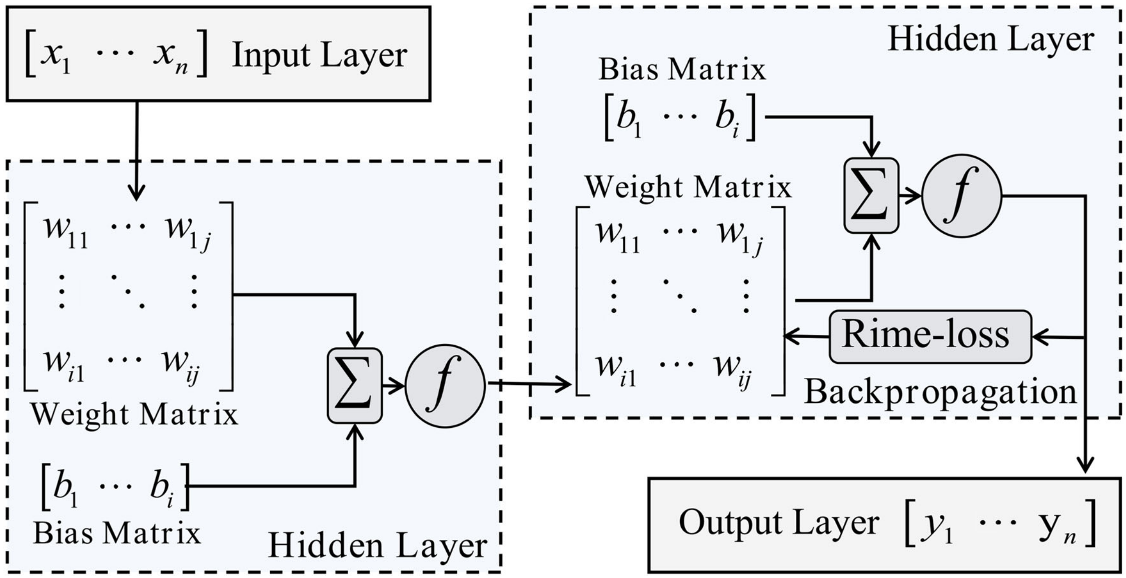

Figure 10. Architecture of the RIME-BPNN performance prediction model.

caused by manual adjustment. The RIME algorithm was selected for this adaptive tuning task over other common optimisers (e.g. PSO) due to its demonstrated efficiency in navigating complex, non-convex parameter spaces and its robust balance between exploration and exploitation, which is particularly beneficial for fine-tuning the hyperparameters of the nonlinear BPNN model. The configuration of the BPNN that incorporates the RIME algorithm (RIME-BPNN) is shown in Figure 10. The adaptive parameter tuning RIME-BPNN axial-flow pump performance prediction model is henceforth designated as the RIME-BPNN performance prediction model.

4.2. CTGAN-Based data augmentation for axial-flow pump performance

To address the limited availability and high computational cost of CFD-generated datasets, this subsection introduces a CTGAN-based data augmentation module for axial-flow pump performance prediction. The CTGAN is trained on the original simulation data to learn the joint distribution of design parameters and performance indicators, enabling the generation of high-quality synthetic samples that preserve the underlying statistical characteristics. The augmented dataset improves model generalisation and provides a more robust foundation for training the performance prediction model.

4.2.1. Fundamentals of generative adversarial networks (GANs)

The generative adversarial network (GAN) is a deep learning model comprising a Generator and a Discriminator. The generator aims to produce synthetic data indistinguishable from the real ones, while the discriminator learns to discriminate between real and generated samples. The core idea of GAN is that, through the adversarial training, both the generator and discriminator constantly improve their own capabilities. The generator improves its strategy by attempting to fool the discriminator, while the discriminator enhances its capacity to accurately classify real and generated samples. Eventually, the generator produces high-quality and realistic data. The optimisation formula is the following:

$$\min_G \max_D V(G, D) = E_{x\sim P_{\mathrm{data}}(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z)))] \tag{33}$$

Where G is the generator and D is the discriminator; $V(G,D)$ represents the loss function for the generator and the discriminator. The generator attempts to minimise this function, while the discriminator attempts to maximise it. P_{data} is the probability distribution of the original data. P_z is the noise distribution input to the generator. $G(z)$ represents synthetic samples generated by the generator. $D(\cdot)$ is the discriminant function. The specific workflow is as follows:

1. **Input of random noise:** The generator is fed with random noise sampled from the distribution P_z.

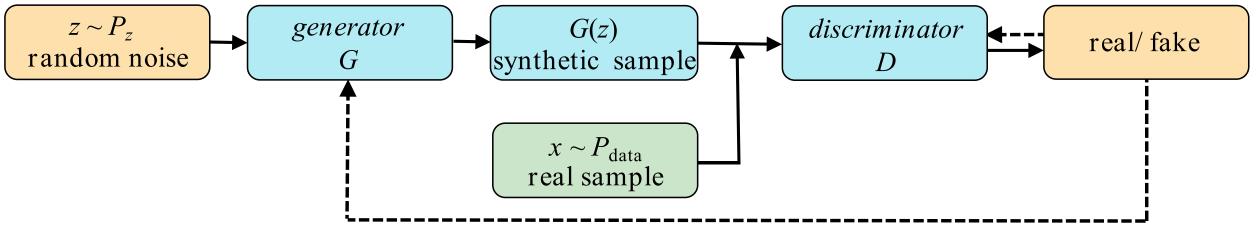

Figure 11. Schematic architecture of a GAN.

2. Generation of synthetic samples: The generator maps the input noise to the data space, producing a synthetic sample $G(z)$.

3. Discriminator evaluation: The discriminator D receives a sample (either a real sample x or a generated sample $G(z)$) and determines its authenticity.

4. Optimisation of generator and discriminator: The generator minimises the loss function to produce sufficiently realistic samples that challenge the discriminator's ability to distinguish them from real data. Meanwhile, the discriminator optimises its parameters to enhance its discriminative power, striving to accurately differentiate between real and generated samples.

5. Iterative training: The generator and discriminator engage in a recurrent adversarial cycle. In each training iteration, the generator attempts to produce more convincing samples, while the discriminator continually refines its ability to identify synthetic data. During the training process, the generator's output increasingly approximates the real data distribution and the discriminator's judgment becomes more accurate.

6. Convergence: The system reaches an equilibrium point where the generator produces samples indistinguishable from real data, and the discriminator can no longer reliably differentiate between them. At this stage, the GAN training process is complete, and the model exhibits optimal generative and discriminative capabilities.

The specific workflow of the model is illustrated in Figure 11.

4.2.2. Basic introduction to CTGAN and the data augmentation module

Given the need to augment the limited, high-dimensional tabular data derived from CFD simulations, the Conditional Tabular Generative Adversarial Network (CTGAN) was selected over other generative models (e.g. standard GANs, VAEs, or SMOTE). This choice is driven by CTGAN's specific design for handling tabular data, which often contains mixed continuous and discrete variables with multi-modal distributions. CTGAN addresses key limitations such as mode collapse in traditional GANs and provides mechanisms like mode-specific normalisation to faithfully capture the complex joint distributions of pump design and performance parameters.

Despite the notable advancements made by GAN in the field of probabilistic modelling and sample generation, its fundamental architecture remains encumbered by inherent limitations, such as mode collapse and unstable gradients. To enhance the model performance, distinct variants are proposed based on the improvements of network structures and loss functions, including PacGAN that introduces the probabilistic constraint. However, when the tabular data containing mixed features (such as numerical parameters and discrete labels) are applied, current generative models exhibit insufficient capability in modelling the distribution of discrete features. This discrepancy often leads to statistical divergence between the generated samples and the real situation. The CTGAN is a generative adversarial network designed for tabular data, which is a variant of the traditional GAN (Zhang et al., 2018). The primary objective of this work is to construct structured tabular data that conforms to authentic statistical characteristics by deeply mining the intrinsic distribution patterns of the data.

In comparison with the traditional generative model, the CTGAN one uses a distinctive integration of the feature-aware mechanism with the distribution reconstruction technology to formulate the modal adaptive normalisation method for the continuous-type parameters in the axial-flow pump tabular data. This method maintains the multi-peak distribution characteristics of the data. The specific steps are as follows:

1. The number of modes for each continuous variable is identified by the variational gaussian mixture model (VGMM) by fitting a Gaussian mixture model. Figure 12(a) shows the three modes identified by the VGMM, denoted as e_1, e_2, and e_3, and the corresponding Gaussian mixture model is denoted as:

$$P_{Ci}(c_{i,j}) = \sum_{k=1}^{3} \mu_k N(c_{i,j}; e_k, \phi_k) \tag{34}$$

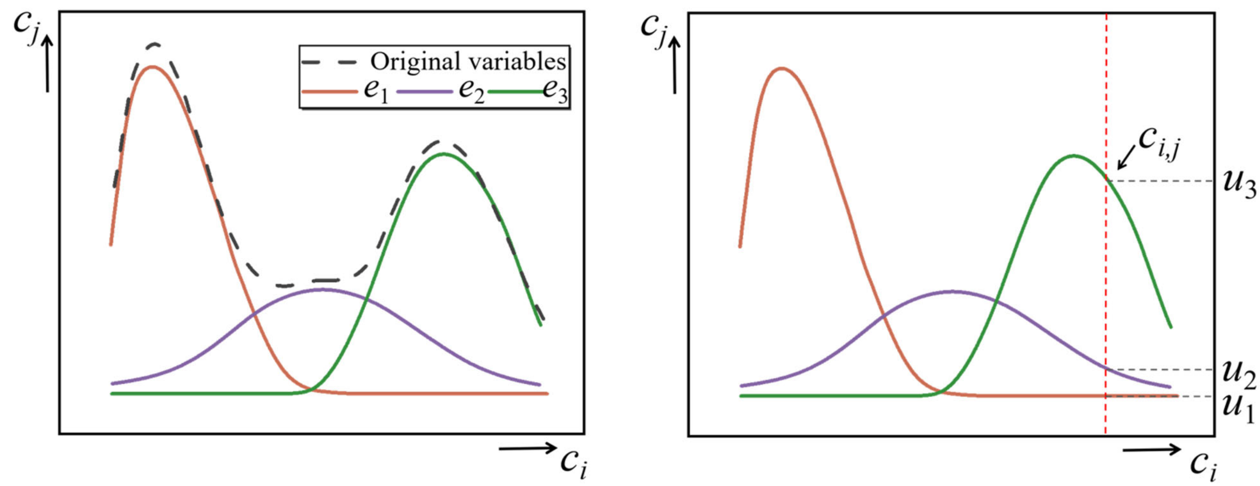

Figure 12. Variational Gaussian mixture model (VGMM)-based mode identification.

where C_i is the continuous column; $c_{i,j}$ is the continuous column sample value; μ_k is the weight of the pattern; and Φ_k is the pattern standard deviation.

2. The probability that each value belongs to a different mode in a continuous column is calculated. As shown in Figure 12(b), the probability densities of the three modes are denoted by u1, u2, and u3 as follows:

$$u_k = \mu_k N(c_{i,j}; e_k, \phi_k) \tag{35}$$

3. One mode is chosen randomly from multiple modes and used to normalise the values. Considering Figure 12 as an example, assuming that the 3rd mode u_3 is selected based on the probabilities of u_1, u_2 and u_3, the values $c_{i,j}$ in the consecutive columns C_i are represented as one-hot encoding $\beta_{i,j} = [0,0,1]$, which points to the 3rd mode, and 1 scalar $\alpha_{i,j} = \frac{c_{i,j} - e_3}{4\Phi_3}$ used to represent the values within the modes as a final normalisation result.

The axial-flow pump frequently requires a significant amount of time to generate a complete set of data, including design parameters, hydraulic efficiency, and head due to the long calculation time. Based on this background, the CTGAN demonstrates notable advantages in the field of axial-flow pump data generation, which can quickly generate a substantial volume of high-fidelity synthetic data by aligning with the distribution characteristics of axial-flow pump factors (e.g. cascade solidity, vortex coefficient, head, etc.). The use of these synthetic data demonstrated to be an effective method for expanding the training dataset, thereby providing rich samples for the machine learning model, and accelerate the intelligent design and energy performance prediction of axial-flow pumps.

4.2.3. Development of the CTGAN-based performance prediction model

The CTGAN-based energy performance prediction model of axial-flow pumps proposed in this work mainly comprises the two primary components mentioned above: the RIME-BPNN energy performance prediction model and the CTGAN data enhancement module. Due to the time-consuming nature of numerical simulations, it requires 2 h (average) to complete each individual case based on a 9-million grid, using a 36-core CPU FLUENT for the CFD computation. After the training, the CTGAN network can generate thousands of high-quality data samples within seconds. As the RIME-BPNN energy performance prediction model training typically requires thousands of data sets, the incorporation of the CTGAN data augmentation module has effectively mitigated this issue by accelerating the long data preparation time in the preliminary phase. However, the data generated by the CTGAN may contain outliers, which can affect the prediction accuracy of the performance model. For this reason, this work adopts the isolation forest algorithm to detect and clean the outliers to ensure higher data quality. The isolation forest algorithm recursively divides the data set by constructing multiple isolation trees, randomly selecting features and cut points. The outliers have a shorter path through the tree, and they are identified by calculating the outlier scores. After enhancing the data through the principal component analysis for dimensionality reduction, the anomaly score of each sample point is calculated and the outlier points are removed to ensure the data quality. The algorithm steps are specified as follows:

1. **Data selection:** Selecting the feature data to be detected.

Constructing the isolation tree: Constructing the isolation tree by randomly selecting features and cutting points by repeating this process multiple times to build multiple isolation trees and form a forest.

2. **Calculating the anomaly score:** Each data point is isolated by multiple isolation trees, and the path length (e.g. isolation depth) of each data point is calculated. The anomaly score is calculated based on the average value of the path depth. The anomaly score is calculated as follows:

$$score_m(x) = 2^{-\frac{h(x)}{c(n)}} \tag{36}$$

Where $h(x)$ is the average path length of sample x, and $c(n)$ is a normalisation constant, related to the total number of samples n.

$$h(x) = \frac{1}{N}\sum_{i=1}^{N} h_i(x) \tag{37}$$

$$c(n) = 2\left(\frac{\log(n-1)}{n}\right) \tag{38}$$

Where N is the number of trees, and $h_i(x)$ is the path length of sample x in the i_{th} tree.

3. **Judging anomalies:** According to the anomaly score, a threshold value is set, marking the points with anomaly scores higher than the threshold value as outliers.

4.3. Intelligent optimization framework

The trained performance prediction model needs to be optimised by an intelligent optimisation algorithm. A comparison of the standard genetic algorithm (SGA) with other intelligent optimisation algorithms reveals its notable advantages in the optimisation of BPNN. The SGA's advantages primarily manifest in its capacity for extensive global search and its ability to exhibit high adaptability. Traditional optimisation methods, including the gradient descent, are prone to falling into local optima, especially when facing complex, nonlinear, and multi-peaked objective functions. Conversely, the SGA can effectively avoid the problem of local optimisation by conducting a global search through the diversity of the population. The SGA does not rely on the continuity or the smoothness of the objective function, rendering it more advantageous in dealing with nonlinear problems in BP neural networks. In comparison with algorithms such as the PSO and the SA, the SGA has been demonstrated to exhibit superior capabilities in terms of exploratory ability, particularly in the context of optimizing a discrete space or complex hyper parameters. The SGA has also been demonstrated to be more effective in identifying a globally optimal solution. Therefore, the SGA can better optimise the BP neural network, thereby improving the performance and generalisation ability of the energy performance prediction model. Figure 13 shows the flowchart of the entire framework.

5. CTGAN-Based model training and optimization framework for axial-flow pump energy performance prediction

This section describes the training process and optimisation framework of the proposed CTGAN-based model for axial-flow pump energy performance prediction. The CTGAN-enhanced dataset is used to improve model robustness, while an intelligent optimisation search is applied to efficiently explore the design space and identify optimal performance configurations.

5.1. Definition of design parameters and sample generation

The optimal parameter configuration range has been obtained by two-stage orthogonal optimisation test, and the subsequent analysis does not take into account the variations in the camber line control coefficient. Eight key parameters, selected for the purpose of optimisation, comprise the impeller projection coefficient P_r, the impeller blade setting angle coefficient C_{rsa}, the guide vane blade setting angle coefficient C_{ssa}, the impeller blade thickness coefficient T_r, the guide vane blade thickness coefficient T_s, the impeller blade cascade solidity σ_r, the guide vane cascade solidity σ_s, and the vortex coefficient a. The specific values are listed in Table 3.

During the sub-dataset construction process, a random sampling strategy is adopted, as the axial-flow pump design parameters form a continuous parameter space without distinct classification boundaries. This results in a 2000-group data scheme. Both the hydraulic efficiency and head of the axial-flow pump are obtained through numerical simulation. To constrain the head and, thereby, obtain the axial-flow pump with the optimal energy performance, the result ζ of Eq.(39) is used as a comprehensive optimisation index of the axial pump. Since there are head losses during the operation of the axial-flow pump, the head target is set to 4.9 m.

$$\xi = \eta - |H - 4.9| \times co \tag{39}$$

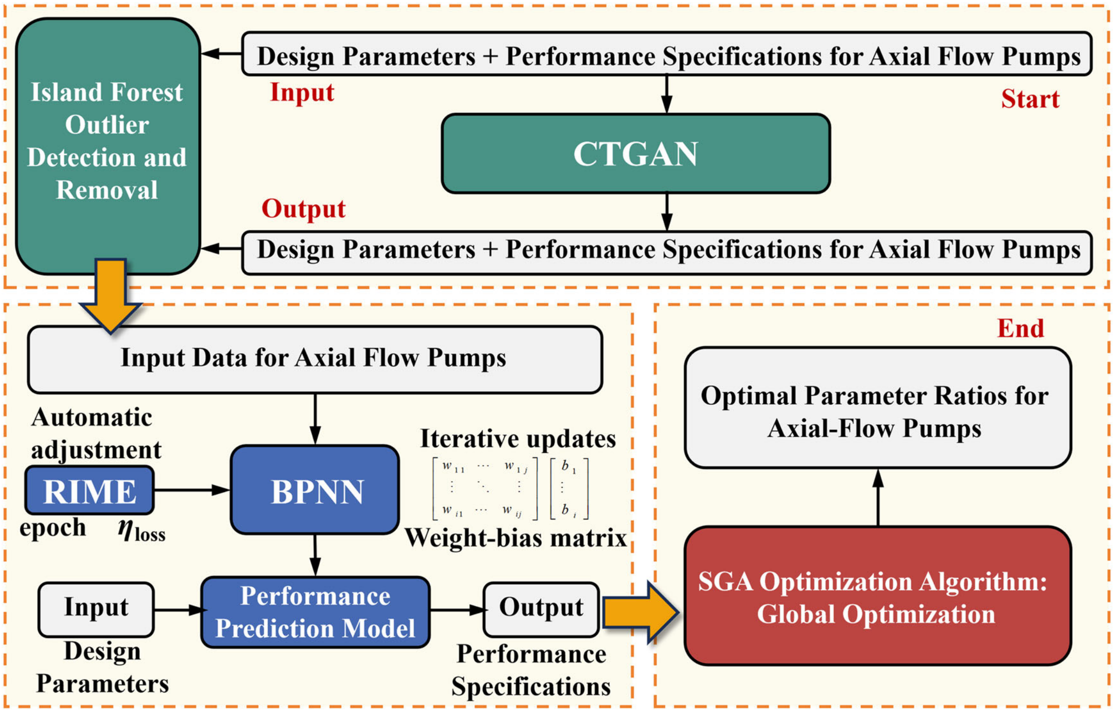

Figure 13. CTGAN-based optimisation framework for energy performance prediction of axial-flow pumps.

Table 3. Design parameter ranges adopted for sample generation.

Number	Parameter	Value	Number	Parameter	Value
1	P_r	$0 \sim 1$	5	T_s	$0.04 \sim 0.09$
2	C_{rsa}	$0 \sim 1$	6	σ_r	$0.70 \sim 1.20$
3	C_{ssa}	$0 \sim 1$	7	σ_s	$0.70 \sim 1.20$
4	T_r	$0.04 \sim 0.09$	8	a	$-1 \sim 1$

Table 4. Optimal design parameter values of the axial-flow pump model.

Number	Parameter	Value	Number	Parameter	Value
1	P_r	0.5517	5	T_s	0.0472
2	C_{rsa}	0.6113	6	σ_r	0.7023
3	C_{ssa}	0.4989	7	σ_s	0.8865
4	T_r	0.0446	8	a	0.7689

Where η is the hydraulic efficiency, H is the axial-flow pump head, and co denotes the penalty factor, and $co = 0.1$.

5.2. Training and optimization of energy performance prediction models for axial-flow pumps

This subsection details the training procedure and optimisation strategy of the energy performance prediction models for axial-flow pumps. The CTGAN-augmented dataset is used to enhance model accuracy and generalisation, while an intelligent optimisation algorithm is employed to identify optimal parameter combinations within the design space.

5.2.1. Training and optimization of the RIME-BPNN performance prediction model

To evaluate the improvement from the application of the CTGAN data enhancement module on the RIME-BPNN energy performance prediction model, the RIME-BPNN energy performance prediction model is initially trained with the original dataset of 4000 samples without the CTGAN data enhancement module. Since this work aims to solve the multi-objective optimisation problem, it is needed to simultaneously optimise two objectives such as the hydraulic efficiency and head. To simplify the optimisation process, Eq.(39) is used as the optimisation objective function to achieve this target. The trained RIME-BPNN energy performance prediction model is optimised using the SGA algorithm, and the optimal parameters of the optimisation results are listed in Table 4. The optimum axial-flow pump model is predicted to achieve a hydraulic efficiency of 87.81% and a head of 4.92 m. The hydraulic performance of the optimised model is verified using the numerical simulation. The hydraulic efficiency is 86.99% (the relative percentage error is 0.82%), the head is 4.87 m (the relative percentage error is 0.05 m), and the specific speed is 1024.16. The model performance is assessed by comparing the predicted results with the baseline values, as quantified by the previously defined relative percentage errors.

5.2.2. Training and optimization of the CTGAN–RIME-BPNN performance prediction model

The CTGAN network model is used for data augmentation training. Detailed hyperparameter settings, including network architecture, learning rate, batch size, and training epochs, are provided in Section 5.2.3 to ensure reproducibility. A small learning rate (8×10^{-5}) is adopted to prevent training instability arising from imbalanced learning speeds, allowing the generator and discriminator to be optimised synergistically with comparable update steps.

In this work, the original dataset is augmented with data using a ratio of 2:1, i.e. 2 000 sets of samples are generated with the trained CTGAN network (The original dataset consists of 4000 samples). Figure 14(a-h) presents the normalised cumulative distribution plots for axial-flow pump design parameters, and Figure 14(i) for the score from Eq. (36). There is no clear separation between the real data and the data augmented by the CTGAN in terms of distribution, and their cumulative trends are highly consistent, indicating that the augmented data effectively preserves the distribution characteristics of the original data. Kolmogorov–Smirnov tests further confirm that all eight design parameters retain their original distributions ($p > 0.05$). In addition, the correlation matrices of the original and augmented datasets were compared. The average absolute difference in Pearson correlation coefficients between the two datasets was found to be less than 0.05, indicating that the intefeature relationships are also well preserved.

Since the data generated by CTGAN may contain outliers, which may interfere with the prediction results if used directly for subsequent prediction model training, the enhanced data needs to be cleaned. Given the high dimensionality of the original data, the principal component analysis (PCA) is carried out to downscale the data to retain the main features and reduce the noise and redundant information, enhancing the effect of anomaly detection. Figure 15 illustrates the distribution of the original data along with the anomaly detection results, where Figure 15(a) shows the original data distribution and Figure 15(b) shows the anomaly detection results. The outlier score of each sample point is calculated with Eq.(36): the shorter the path, the higher the score of the sample point. This indicates that the sample point is more likely to be an outlier, as it is easier to be isolated. In this work, we use the Isolation Forest algorithm with an anomaly score threshold of 0.6, meaning samples with $score(x) > 0.6$ are considered outliers and removed. The anomaly score threshold of 0.6 was empirically determined by analyzing the distribution of anomaly scores for true samples in the original training set. Specifically, we first trained the Isolation Forest model using the training set (excluding any synthetic samples) and calculated the anomaly scores for each real sample. We found that only about 1.8% of the real samples had scores higher than 0.6. Therefore, setting the threshold to 0.6 effectively removes outliers in the synthetic data that deviate from the true distribution while retaining the vast majority of real samples. As a cross-validation test, we tried thresholds of 0.5 and 0.7: at 0.5, the false rejection rate was too high (approximately 6% of real samples were misclassified), while at 0.7, the number of suspicious synthetic samples missed increased. The value of 0.6 performed best in terms of statistical distribution and the final model's MSE, and was therefore adopted. The algorithm is configured with 100 isolation trees, each trained on a random subsample of 256 data points. This threshold was empirically determined by examining the score distribution of the original training data, where less than 2% of real samples exceeded this value. After generating 2000 synthetic samples with the trained CTGAN, 88 samples were flagged as outliers and removed, leaving 1912 high-quality samples. The distribution of outlier points in the augmented datum is shown in Figure 16.

The 1,912 enhanced data sets (2000 sets are generated by the CTGAN and 88 sets are removed by the outlier detection) and the 4,000 original data sets are used for training the energy performance prediction model. The mean square error (MSE) expression for the performance prediction model is:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\xi_{\mathrm{pre}} - \xi_{\mathrm{act}})^2 \tag{40}$$

Where n is the total number of samples, ζ_{pre} is the predicted score, ζ_{act} is the actual calculated score.

Figure 17 displays the MSE comparison of the RIME-BPNN performance prediction model for the test set with and without the CTGAN data enhancement. With data enhancement, the MSE of the RIME-BPNN energy performance prediction model is reduced from 1.015×10^{-4} (the relative percentage error is 1.007%) to 7.88×10^{-5} (the relative percentage error is 0.887%). The experimental results show that, under the same conditions, the incorporation of the CTGAN data enhancement module can remarkably improve the accuracy and generalisation capability of the axial-flow pump energy performance prediction model.

Moreover, the CTGAN data enhancement module used in this work boasts the core advantage of 'train once, deploy repeatedly'. Although the initial training of the CTGAN data enhancement module requires a certain computational investment, typically completed within tens of minutes, this constitutes a one-time cost. Once the model is trained, subsequent data augmentation tasks

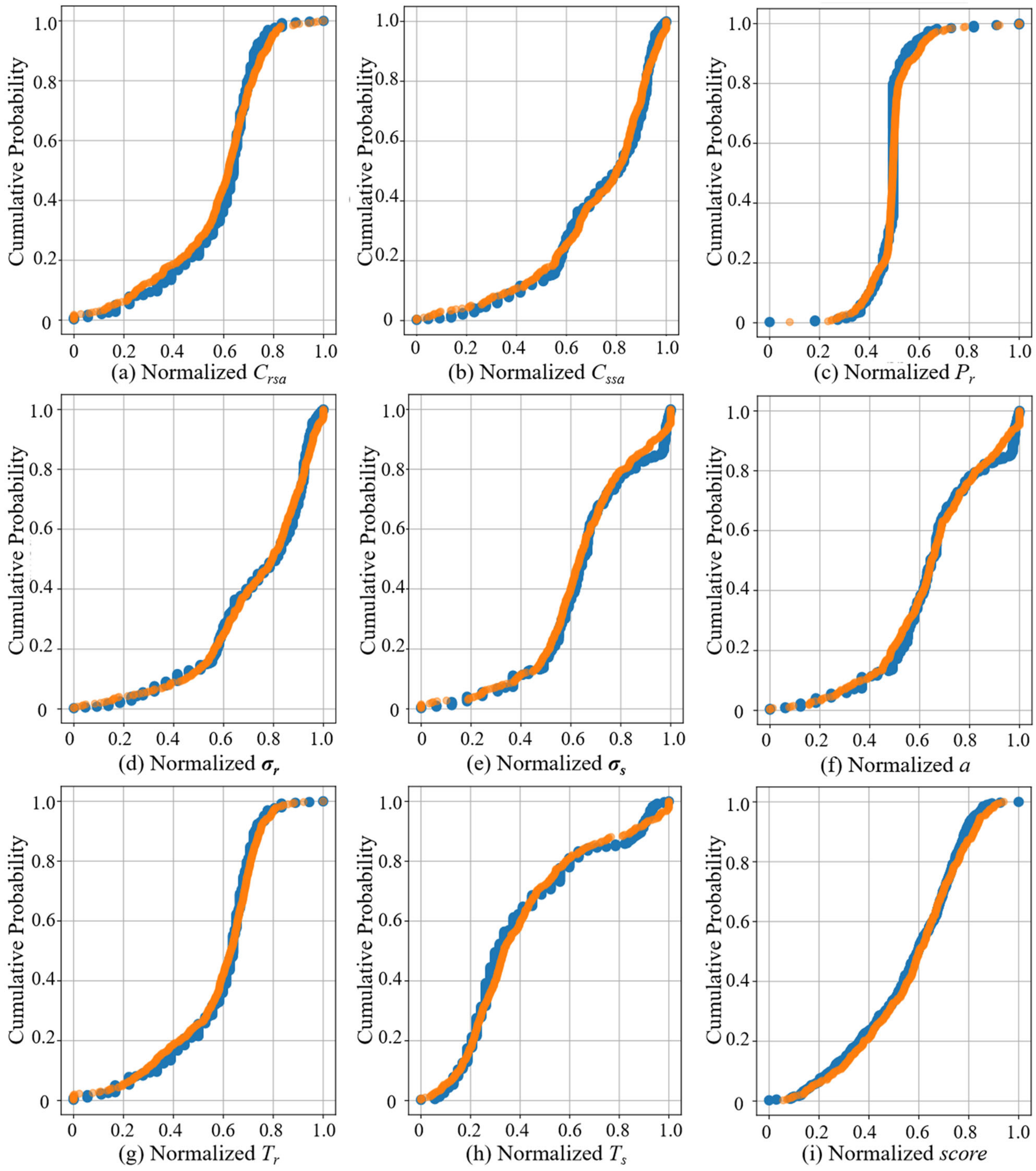

Figure 14. Normalised cumulative distribution functions of selected features.

are accomplished within seconds, enabling near real-time data generation. Coupled with PCA and Isolation Forest, the entire process of anomalous data scrubbing is completed within tens of seconds. In this work, the application of the CTGAN data enhancement module yields a total of 1912 usable samples. Compared to CFD simulations, this approach saves approximately 3824 h of computational time (based on a 9-million grid, using a 36-core CPU with ANSYS FLUENT for the CFD computation, the average time consumed for each group of samples is about 2 h). The time required for individual stages is illustrated in Figure 18. By reducing the traditional CFD computation cycle from hour-level to second-level, the proposed methodology demonstrates significant efficiency advantages in long-term, multi-iteration engineering workflows.

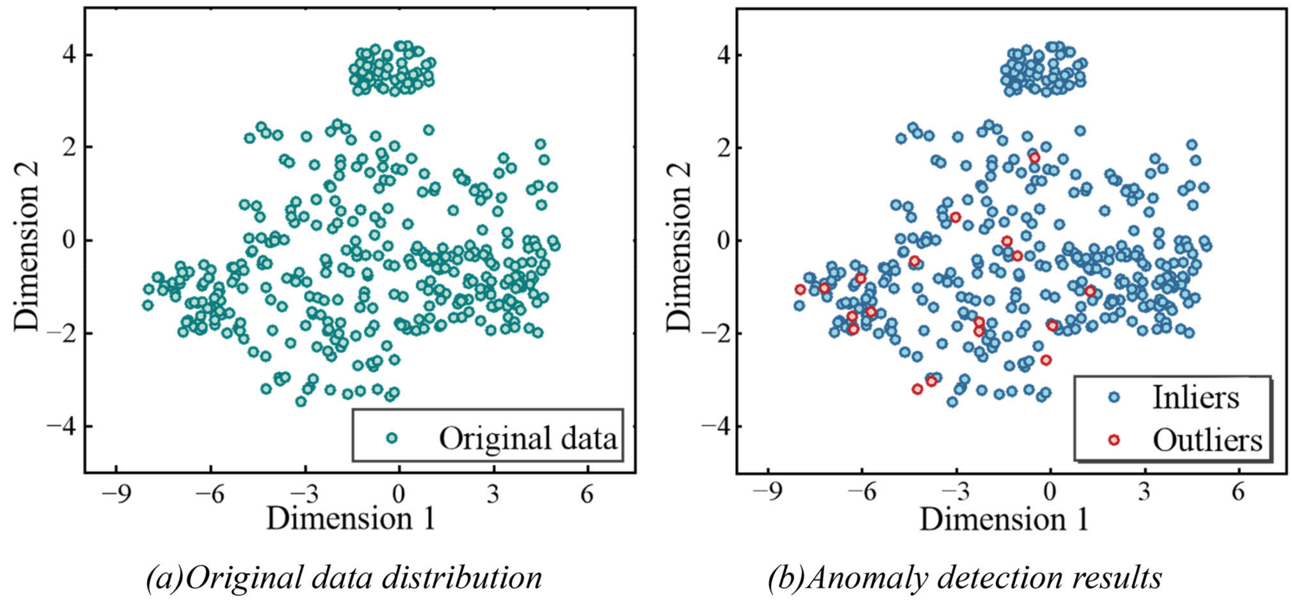

Figure 15. Outlier detection in feature distributions after dimensionality reduction.

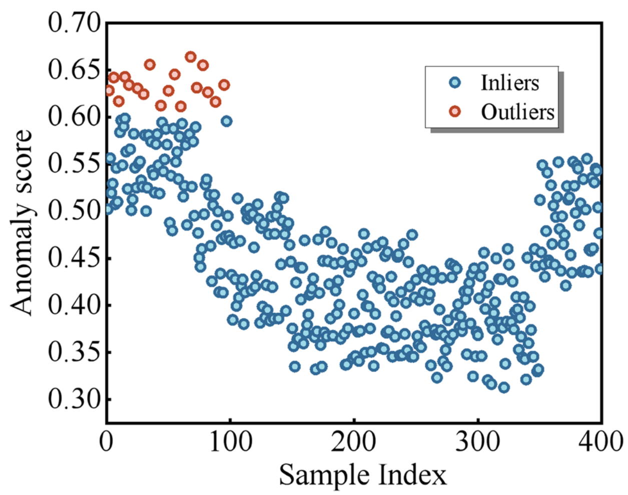

Figure 16. Distribution of detected outlier points.

The trained RIME-BPNN energy performance prediction model is optimised using the SGA algorithm, and the optimal parameters of the optimisation results are shown in Table 5. The optimised axial-flow pump model achieves a hydraulic efficiency of 88.49% and a head of 4.92 m. The energy performance of the optimised model is verified using the numerical simulation. The hydraulic efficiency is 87.87% (the relative percentage error is 0.71%), the head is 4.95 m (the relative percentage error is 0.03 m), and the specific speed is 1011.71.

It is noteworthy that the optimal vortex coefficient $a = 0.7721$ is close to the free vortex assumption ($a = 1$), indicating that the optimisation tends toward a more uniform circumferential velocity distribution at the impeller outlet, which helps reduce tip leakage vortices and secondary flow losses. Meanwhile, the optimal cascade solidity $\sigma_r = 0.7011$ (close to its lower bound) implies a lower blade solidity, which can reduce friction losses while still balancing the required work capacity. These prove that the machine learning optimisation results are physically interpretable.

In addition, it should be noted that the synthetic data generated by CTGAN, while statistically consistent with the original CFD dataset, may not fully represent all rare or edge-case scenarios present in real-world operations. The current framework's predictions are therefore most reliable within the design space covered by the original samples. Future work should incorporate physics-informed constraints and targeted validation of extreme parameter combinations to further enhance the generalizability of the synthetic-data-augmented approach.

5.2.3. Detailed configuration of machine learning models and optimization algorithms

This section provides the specific parameter settings for the CTGAN data augmentation module, the BPNN surrogate model, the RIME hyperparameter optimisation, and the SGA global optimisation, ensuring the reproducibility of the proposed framework. Additionally, the computational environment for this study was configured as follows: CFD simulations were performed using ANSYS FLUENT on a 36–core Intel Xeon Gold 6240 CPU cluster at Hohai University's High–Performance Computing Platform; machine learning models were implemented in Python 3.9. Additionally, regarding the parameter settings for each module, the number of layers

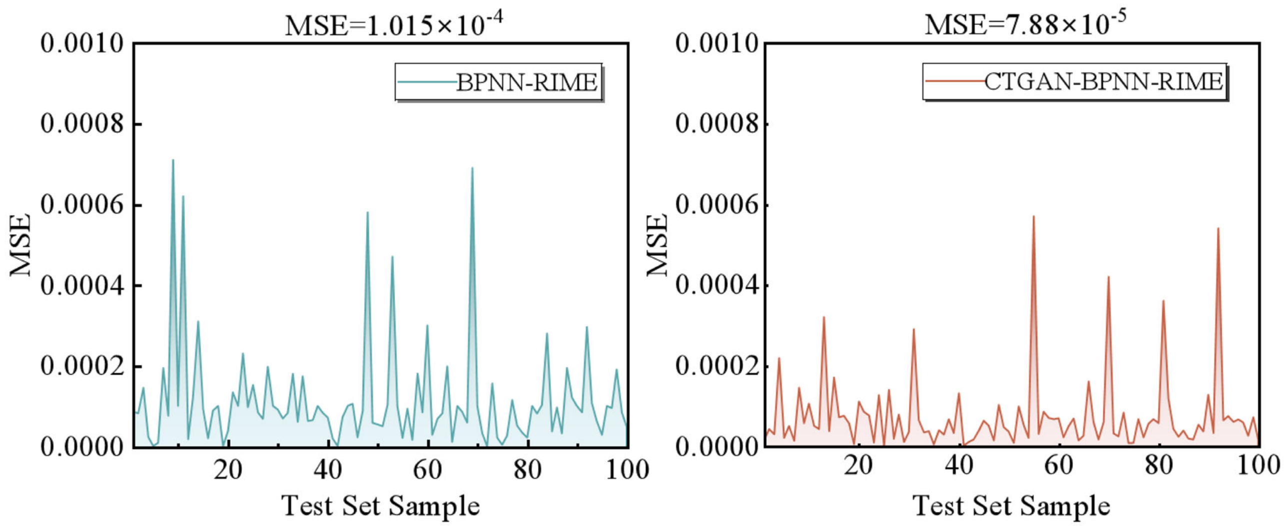

Figure 17. Comparison of test-set mean squared error (MSE) with and without data enhancement.

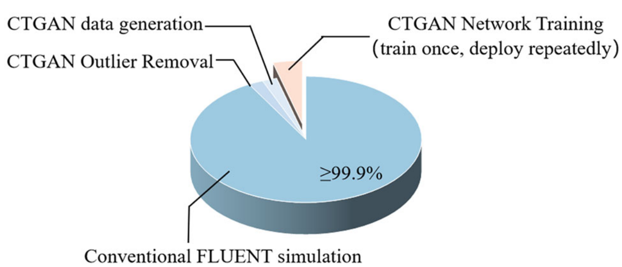

Figure 18. Computational time required for individual stages of the proposed framework.

Table 5. Optimised design parameter values for the axial-flow pump model.

Number	Parameter	Value	Number	Parameter	Value
1	P_r	0.5623	5	T_s	0.0475
2	C_{rsa}	0.6125	6	σ_r	0.7011
3	C_{ssa}	0.5001	7	σ_s	0.8752
4	T_r	0.0438	8	a	0.7721

and neurons in the CTGAN' s generator and discriminator were based on the recommended configurations from Xu et al. (2019) and Habibi et al. (2023) for tabular data of similar scale; Learning rates and batch sizes were determined through preliminary grid searches to maintain a balanced training process for the generator and discriminator; the population size and number of iterations for RIME were set based on preliminary convergence tests (where the loss function stabilised after 40 generations); SGA parameters adopted empirical values commonly used in the field of genetic algorithms (Deb et al., 2002). The specific settings for each module are outlined below.

The original CFD dataset consists of 4000 samples generated by random sampling within the parameter ranges specified in Table 3. To prevent data leakage and artifacts, the dataset is first randomly split into training, validation, and test sets using a 70:15:15 ratio, with stratification based on efficiency deciles to maintain distribution consistency across subsets. All subsequent data processing is then guided solely by the training set: feature normalisation parameters are computed from the training set and applied to all three subsets; a CTGAN model is trained exclusively on the training set to generate synthetic samples, which are used only to augment the training data; and the generated samples are validated against the original training set to ensure statistical consistency. This ensures that the validation and test sets remain untouched until the final evaluation.

The CTGAN architecture is configured as follows (Habibi et al., 2023; Xu et al., 2019): the generator consists of two hidden layers with 64 neurons each (ReLU activation), while the discriminator comprises five hidden layers with 1024, 512, 256, 168, and 64 neurons respectively (LeakyReLU activation with a negative slope of 0.2). The embedding dimension for categorical features is set to 128. Both networks are trained using the Adam optimiser with a learning rate of 8×10^{-5}, $\beta_1 = 0.5$, $\beta_2 = 0.999$, a batch size of 500, and 1000 epochs. After training, 2000 synthetic samples are generated, and outliers are removed using the Isolation Forest algorithm with an anomaly score threshold of 0.6, resulting in 1912 high-quality samples.

The BPNN surrogate model is implemented using the scikit-learn MLPRegressor with a single hidden layer. The size of this hidden layer is optimised by the RIME algorithm (see below). The hidden layer employs the ReLU activation function, and the network is trained using the Adam solver with default parameters. The output layer consists of two neurons corresponding to the

Table 6. Hyperparameter configuration of modules.

Module	Parameter	Value / Range
CTGAN	Generator hidden layers	2 layers, 64 neurons each, ReLU
	Discriminator hidden layers	5 layers: 1024, 512, 256, 168, 64; LeakyReLU (negative slope of 0.2)
	Embedding dimension for categorical features	128
	Learning rate / β_1 / β_2	8×10^{-5} / 0.5 / 0.999
	Batch size / epochs	500 / 1000
	Isolation Forest trees / subsample size / anomaly threshold	100 / 256 / 0.6
BPNN	Number of hidden layers	1 (size optimised by RIME)
	Hidden layer size range (RIME search space)	1–3000 (integer)
	Activation function / solver	ReLU / Adam
	Max iterations / early stopping	200 / based on validation loss
RIME	Optimised variables	hidden layer size (integer, 1–3000); initial learning rate (continuous, 5×10^{-5} – 0.01)
	Population size / iterations / early stopping	10 / 50 / 20 generations no improvement
	Soft–rime (β) / hard–rime (α)	0.6 / 0.2
SGA	Population size / generations	100 / 200
	SBX crossover: distribution index (η) / probability	20 / 0.8
	Polynomial mutation: η / probability	20 / 0.1
	Selection / elitism	Tournament (size = 3) / top 5%
	Termination condition	max generations or 30 generations no improvement

predicted hydraulic efficiency and head. The loss function is the mean squared error (MSE), and training is performed for a maximum of 200 iterations with early stopping based on validation loss to prevent overfitting.

Hyperparameter optimisation is conducted using the RIME algorithm from the mealpy library. RIME is a metaheuristic algorithm that explores the search space through soft-rime and hard-rime mechanisms, not via grid or random search. The objective function is the negative mean squared error obtained from 5-fold cross-validation on the training set. Two decision variables are optimised: (i) the hidden layer size, an integer ranging from 1 to 3000, and (ii) the initial learning rate, a continuous variable ranging from 5×10^{-5} to 0.01. The RIME algorithm is configured with a population size of 10 and 50 iterations. The soft-rime parameter $\beta = 0.6$ controls the exploration phase, while the hard-rime parameter $\alpha = 0.2$ governs the exploitation phase. The optimisation process terminates if no improvement in the best fitness is observed for 20 consecutive iterations.

The global optimisation of the eight design parameters listed in Table 3 is performed using a standard genetic algorithm (SGA). The SGA is configured with a population size of 100 and evolves for 200 generations. Simulated binary crossover (SBX) with a distribution index $\eta = 20$ is applied with a probability of 0.8, and polynomial mutation with $\eta = 20$ is applied with a probability of 0.1. Tournament selection with a tournament size of 3 is used to select parents for reproduction, and an elitism strategy preserves the top 5% of individuals in each generation. The objective function to be maximised is the comprehensive optimisation index ζ defined in Eq. (39), which balances hydraulic efficiency and head deviation from the target value of 4.9 m. The optimisation process terminates when the maximum number of generations is reached or when the best fitness does not improve for 30 consecutive generations. The specific configurations of the CTGAN, BPNN, RIME and SGA modules are listed in Table 6.

RIME optimises the BPNN's hyperparameters (hidden layer size and learning rate), while SGA optimises the eight pump design parameters. The improved prediction accuracy from RIME-BPNN directly benefits SGA's search for optimal designs.

5.3. Comparison of axial-flow pump performance and flow regimes

The key performance indicators (KPIs) of baseline and optimised models in design condition are listed in Table 7, and the curves of the external characteristics of baseline and optimised models in full-load are shown in Figure 19. Compared to baseline model, the hydraulic efficiency of optimised model is improved by 0.88% in design operating condition, but slightly decreases in partial-load ones. In full-load, the hydraulic efficiencies of the baseline and optimised models are almost the same. These results indicate that the CTGAN data enhancement module remarkably capture the optimal parameter combinations and significantly improve the overall energy performance of the axial-flow pump in normal operating range conditions.

To analyze the flow mechanisms for the energy performance improvement in design conditions of the optimised model (Kan et al., 2025c), Figure 20 shows the comparison of the flow distribution between baseline

Table 7. Comparison of baseline and optimised axial-flow pump model parameters.

Parameter	H (%)	H (m)	n_s
Baseline model	86.99	4.87	1024.16
optimised model	87.87	4.95	1011.71

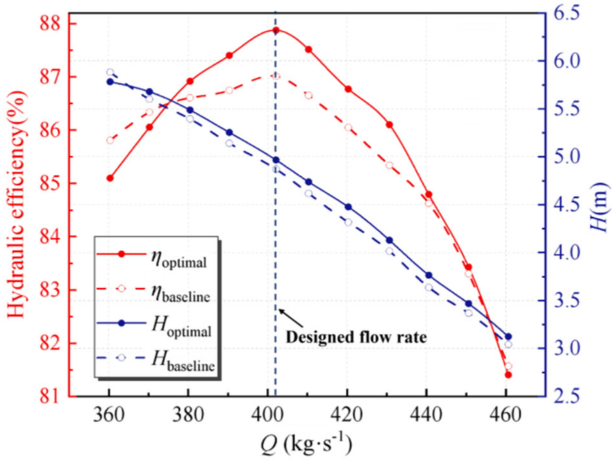

Figure 19. External performance characteristics of the baseline and optimised axial-flow pump models in full-load conditions.

and optimised models at the near-wall location. At the rim wall (Span = 0.95), no significant undesirable flow regimes are observed in the impeller region for either the baseline or the optimised model. No pronounced large-scale vortical structures are observed near the guide vane trailing edge in the optimised model, whereas small-scale vortices are present in the baseline model. At the hub wall (Span = 0.05), no significant differences are observed between the baseline and optimised models in the guide vane region. In the impeller region, however, the baseline model exhibits pronounced flow separation at the blade leading edge, whereas this separation is substantially suppressed in the optimised model, with the flow adhering more closely to the blade surface at the inlet. Overall, the optimised model achieves a more uniform flow distribution and smoother internal flow conditions within the pump passage.

Figure 21 shows the vorticity distribution for the baseline and optimised models at near-wall locations in design conditions. At the rim wall (Span = 0.95), the baseline model exhibits pronounced vortical structures concentrated at the interface between the impeller and guide vane regions (yellow markers) and near the guide vane trailing edge (black markers). In contrast, the corresponding vorticity in the optimised model is markedly weakened. At the hub wall (Span = 0.05), the optimised model shows a significant reduction in vorticity at the guide vane outlet compared with the baseline model. Overall, while the vorticity levels in the impeller region are comparable for both models, the optimised model demonstrates substantially reduced vorticity and improved flow structures in the guide vane region, indicating a clear enhancement in internal flow quality.

Figure 22 illustrates the turbulence kinetic energy (TKE) distributions of the baseline and optimised models at near-wall locations under design conditions. In conjunction with Figure 21, a strong correspondence is observed between the vorticity and TKE distributions. At the rim wall (Span = 0.95), the baseline model exhibits elevated TKE levels within the guide vane passage and

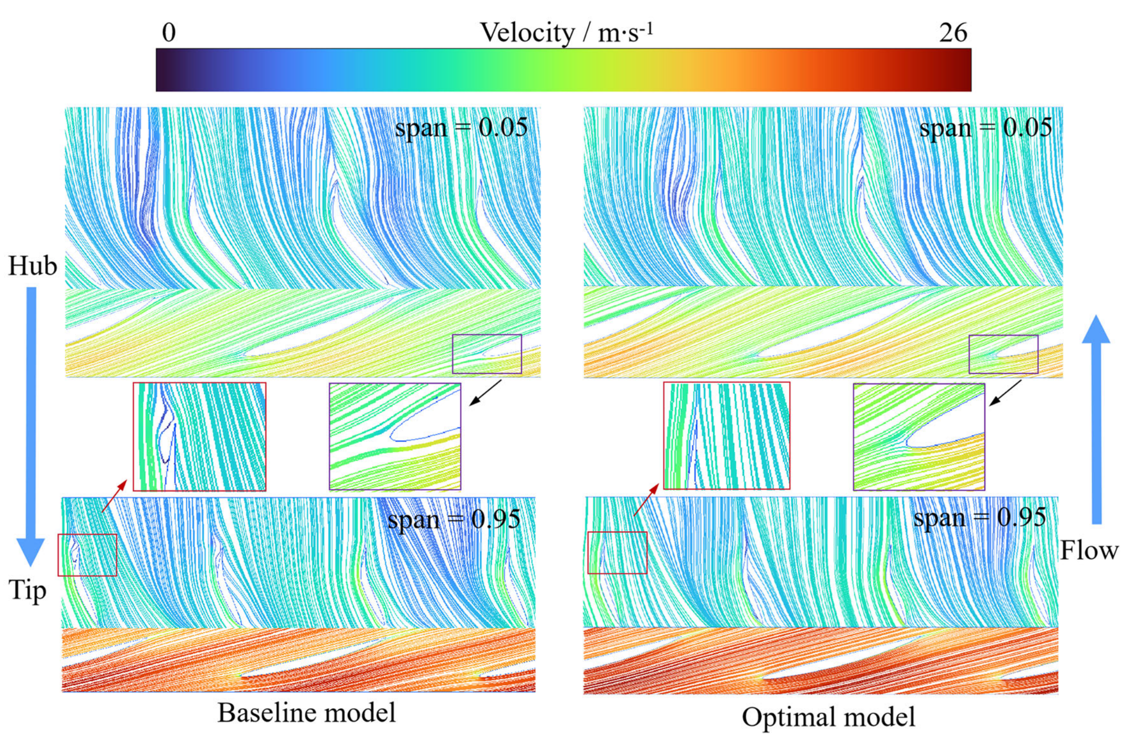

Figure 20. Near-wall streamline distributions of the baseline and optimised models in design conditions.

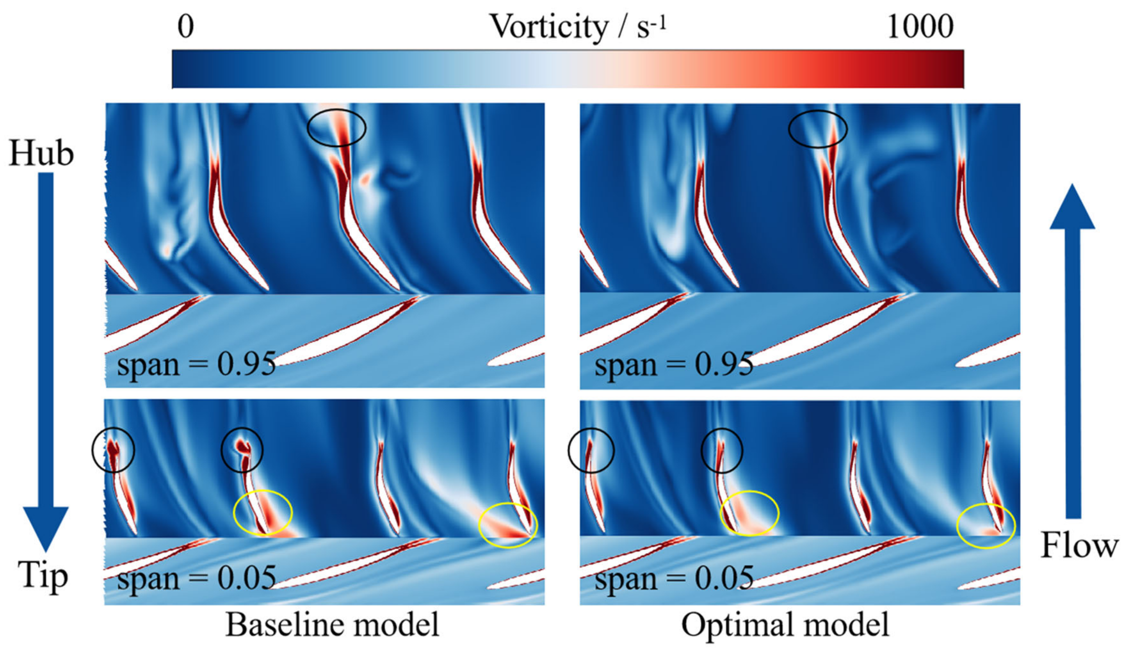

Figure 21. Near-wall vorticity distributions of the baseline and optimised models in design conditions.

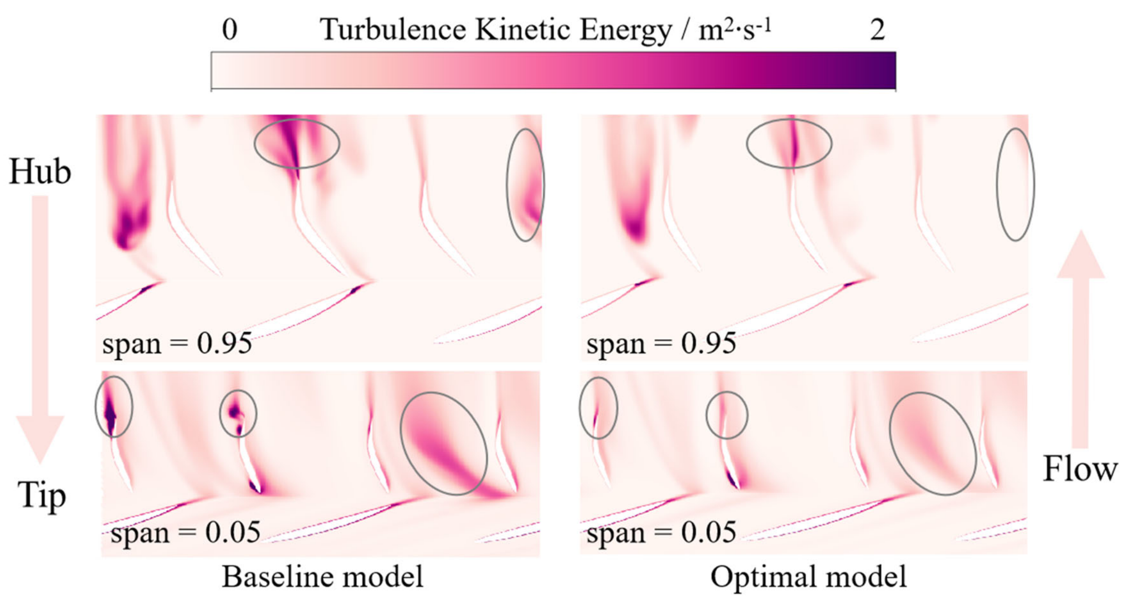

Figure 22. Near-wall turbulence kinetic energy distributions of the baseline and optimised models in design conditions.

near the guide vane trailing edge, whereas these levels are substantially reduced in the optimised model. At the hub wall (Span = 0.05), as highlighted by the black markers, the baseline model shows a broader spatial extent and higher intensity of TKE, while the optimised model presents a markedly reduced distribution range and lower magnitude at the same locations. Overall, the combined analysis of Figures 21 and 22 indicates that vorticity and TKE in the guide vane region are effectively suppressed in the optimised model, leading to improved flow regimes, reduced flow losses, and enhanced energy performance of the axial-flow pump.

5.4. Comprehensive analysis and validation of the optimized performance

The results presented in Section 5.4 demonstrate that the proposed CTGAN-RIME-BPNN framework achieves a 0.88% improvement in hydraulic efficiency compared to the baseline RIME-optimised model. While this gain may appear modest numerically, its practical impact is substantial for large-scale water transfer projects. The Eastern Route of the South-to-North Water Diversion Project alone consumes over 2 TWh annually; a 0.88% efficiency improvement would save approximately 17.6

GWh per year, equivalent to the electricity consumption of 6,000 households and a CO_2 reduction of 14,000 tons. At China's industrial electricity rate of ¥0.8/kWh, this translates to annual cost savings of ¥14 million ($\sim$\$2 million). Moreover, the gain exceeds typical manufacturing tolerances (±0.3%) and measurement uncertainties (±0.2%), ensuring that it is practically realizable in physical prototypes.

The above analysis shows that the 0.88% gain has some engineering value. For comparison, Nguyen et al. (2023) gained 0.974% with impeller only optimisation, and Sun et al. (2022) gained 1.31% for a high specific speed impeller. However, our baseline efficiency is already relatively high (86.99%), and starting from such a level, further improvement is inherently limited. Our framework also saves over 3800 CPU hours, a benefit less emphasised before. Therefore, despite the modest absolute gain, the improvement still offers some positive significance given the higher baseline, broader optimisation scope, and considerable cost savings.

To further validate the reliability of this improvement, the following subsections provide a multi-faceted analysis encompassing ablation study, statistical robustness and limitations.

5.4.1. Ablation study

To quantify the individual contributions of CTGAN augmentation and RIME tuning, we conducted an ablation study with four configurations. It should be noted that the baseline for comparison is the RIME-optimised BPNN model without CTGAN augmentation, which achieves a CFD-validated efficiency of 86.99% as reported in Section 5.2.1. This represents the best performance attainable without synthetic data enhancement.

The four configurations are:

(1) Original BPNN (no RIME, no CTGAN) – a standard BPNN without any optimisation
(2) Baseline (RIME-BPNN, no CTGAN) – the model after RIME hyperparameter optimisation
(3) BPNN + CTGAN (no RIME) – model trained on augmented data without RIME tuning
(4) Full model (CTGAN + RIME-BPNN) – the complete proposed framework

Each configuration was trained and evaluated on the same test set (2000 original samples). Table 8 summarises the results, averaged over 10 random seeds.

The results reveal several insights. First, RIME tuning alone (comparing Original BPNN to Baseline) reduces MSE by 17.7% (from 12.34×10^{-5} to 10.15×10^{-5}) and improves CFD-validated efficiency from 86.71% to 86.99%, a gain of 0.28%. Second, applying CTGAN augmentation without RIME tuning improves efficiency from 86.71% to 87.41%, a gain of 0.70% relative to the original BPNN, and achieves a 0.42% improvement over the Baseline (87.41% vs. 86.99%). Most importantly, compared to the Baseline (86.99%), the Full model achieves a further 0.88% efficiency improvement, reaching 87.87%. This confirms that the 0.88% gain reported as the main result of this work is precisely the contribution of CTGAN augmentation on top of the already RIME-optimised model. The full model's 36.1% MSE reduction (from 12.34×10^{-5} to 7.88×10^{-5}) relative to the original BPNN further demonstrates the synergistic effect of combining both techniques.

It should be noted that SGA was not isolated in the ablation study, because SGA is a global optimiser whose performance depends on the prediction accuracy of the prediction model, which in turn depends on CTGAN and RIME. Moreover, all compared models uniformly employ SGA, making isolation infeasible. The contribution of SGA is therefore implicitly reflected in the final comparison between the full model and the baseline (0.88% improvement).

5.4.2. Statistical robustness analysis

To ensure that the observed improvements are not due to random chance, we repeated the entire experiment with 10 different random seeds (0–9). For each seed, the CTGAN was retrained, the RIME optimisation was rerun, and the final model was evaluated. Table 9 reports the mean ± standard deviation of key metrics, along with paired t-test results comparing the full model against the baseline.

All metrics passed the Shapiro–Wilk normality test ($p > 0.05$). Cohen's d effect sizes exceed 0.8 for all comparisons, indicating large practical significance. These results confirm that the improvements are statistically robust and not artifacts of a particular random seed.

Table 8. Comparison of axial-flow pump hydraulic efficiency in the ablation study.

Configuration	Test MSE ($\times 10^{-5}$)	Predicted η (%)	CFD-validated η (%)	$\Delta\eta$ vs. Baseline
Original BPNN	12.34 ± 1.21	87.21	86.71	−0.28%
Baseline (RIME-BPNN)	10.15 ± 0.89	87.81	86.99	-
BPNN + CTGAN	9.42 ± 0.76	88.02	87.41	+0.42%
Full (CTGAN + RIME-BPNN)	7.88 ± 0.52	88.49	87.87	+0.88%

Table 9. Performance metrics and statistical significance test results.

Metric	Without CTGAN	With CTGAN	Improvement	*p*-value (paired t-test)
Test MSE ($\times 10^{-5}$)	10.15 ± 0.89	7.88 ± 0.52	−22.4%	< 0.001
Test MAE ($\times 10^{-3}$)	8.23 ± 0.71	6.45 ± 0.43	−21.6%	< 0.001
Predicted η (%)	87.81 ± 0.23	88.49 ± 0.15	+0.68%	< 0.001
CFD-validated η (%)	86.99 ± 0.19	87.87 ± 0.12	+0.88%	< 0.001

In addition to evaluating variability across different random seeds, we assessed the sensitivity of the results to different splits of the training data. With a fixed random seed (seed = 0), we performed five independent 70/15/15 train/validation/test splits of the original dataset, each using a distinct random partition. The CTGAN model was retrained on the training portion of each split, followed by the standard RIME–BPNN training and SGA optimisation. The standard deviation of the resulting CFD–validated hydraulic efficiencies across the five splits was found to be comparable to that reported in Table 9 for the 10 different random seeds (0.12 percentage points for the full model). This close agreement indicates that the performance of the proposed framework is not sensitive to the particular choice of data partition. These results further confirm that the reported improvements are statistically robust and not artifacts of a specific training split.

5.4.3. Limitations regarding tip clearance

In the present study, the optimisation framework is established on an idealised geometric model in which the tip clearance is not taken into account. It should be noted that in practical axial–flow pumps, the existence of tip clearance inevitably generates leakage vortices and thereby influences the local flow structure and energy dissipation in the blade tip region. The present work focuses on the main blade geometric parameters that govern the blade profile and flow conditions, which are considered to have a more dominant influence on the overall energy performance. Tip clearance is therefore excluded from the current optimisation. For the rated operating condition of the specific pump under consideration ($n_s = 1030$), the contribution of tip clearance to the total hydraulic loss is relatively minor.

In practical applications where a typical engineering clearance (e.g. 0.5–1.0 mm) is present, the absolute value of this efficiency gain may exhibit a certain degree of variation. However, because tip clearance contributes similar additional losses to both the baseline and the optimised pump configurations, the relative efficiency improvement delivered by the optimisation remains robust. In other words, even after accounting for tip clearance effects, the optimised design is expected to retain a meaningful and valid efficiency gain. When applying the optimised design to actual pump units, it is advisable to perform case–specific CFD validation or to employ empirical loss models for correction based on the actual tip clearance value.

6. Conclusions

In this work, a CTGAN-enhanced energy performance prediction and optimisation framework for axial-flow pumps is proposed by integrating a BP neural network, the RIME algorithm, and an SGA. The method significantly improves the accuracy and efficiency of rapid energy performance prediction and optimisation. The optimised pump achieves a hydraulic efficiency of 87.87%, representing an improvement of 0.88% compared to the baseline model without CTGAN enhancement.

The main conclusions are summarised as follows:

(1) Eight key geometric design parameters are identified to construct a parameterised axial-flow pump model. An initial dataset of 4000 samples is generated via random sampling and expanded to 6000 samples using CTGAN data enhancement. After outlier detection with the isolation forest algorithm, 5912 high-quality samples are retained. The CTGAN-based enhancement substantially shortens the sample generation cycle, saving approximately 3824 h of CFD computation time, while effectively alleviating sample scarcity and quality limitations by learning the intrinsic data distribution.

(2) A RIME-optimised BPNN (RIME-BPNN) energy performance prediction model is developed, in which the RIME algorithm adaptively tunes the learning rate and iteration number, avoiding empirical parameter tuning. With CTGAN-enhanced training data, the MSE decreases from 1.015×10^{-4} to 7.88×10^{-5}, and the relative percentage error is reduced from 1.007% to 0.887%, demonstrating improved prediction accuracy and generalisation capability.

(3) The trained CTGAN-RIME-BPNN model is further optimised using SGA to obtain the optimal design. In design conditions, the optimised pump achieves a hydraulic efficiency of 87.87%, a head of 4.95 m, and a specific speed of 1011.71, with prediction errors

below 1%. Compared to the baseline model, flow field analysis shows suppressed vorticity at the guide vane trailing edge and reduced flow separation at the impeller inlet, leading to lower turbulence kinetic energy, reduced flow losses, and improved overall energy performance.

Although the effectiveness of CTGAN-enhanced learning for axial-flow pump optimisation is validated, the present study focuses on the data-driven relationship between design parameters and performance indicators without explicitly incorporating the physical laws governing internal flow. Future work will address this limitation by integrating physics-informed neural networks (PINNs) that embed the Navier–Stokes equations into the learning process, thereby enhancing physical consistency and interpretability. Additionally, more refined geometric parameters and multi-objective optimisation strategies should be incorporated to simultaneously improve efficiency, head, and cavitation performance. Extending the framework to full operating conditions will further enhance pump stability and efficiency under variable flow rates and rotational speeds. It should also be noted that although the proposed framework was developed for a specific axial-flow pump, its methodology is transferable to other pump types. The framework is not a pure black–box data–driven method – the design parameters are physically based, and the training data come from high–fidelity CFD simulations. Therefore, the CTGAN model learns from physically meaningful inputs and outputs. The same framework can be applied to other hydraulic machinery (e.g. centrifugal pumps, turbines) given a similar parameterisation and an initial CFD dataset. Future work will validate its generalizability across a wider range of designs.

Acknowledgement s

CRediT: **Kan Kan:** Conceptualisation, Funding acquisition, Methodology, Supervision, Validation, Writing – Original draft, Writing – review and editing. **Siqi Gao:** Conceptualisation, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Ye Li:** Conceptualisation, Methodology, Software, Writing – review and editing. **Zhaodan Fei:** Software, Supervision, Validation, Writing – review and editing, Funding acquisition. **Hui Xu:** Funding acquisition, Supervision. **Jinbo Chen:** Formal analysis, Validation. **Mosè Rossi:** Supervision, Writing – review and editing.

Author contributions

CRediT: **Kan Kan:** Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing; **Siqi Gao:** Conceptualization, Formal analysis, Methodology, Software, Validation, Writing – original draft; **Ye Li:** Conceptualization, Methodology, Software, Writing – review & editing; **Zhaodan Fei:** Funding acquisition, Software, Supervision, Validation, Writing – review & editing; **Hui Xu:** Funding acquisition, Supervision; **Jinbo Chen:** Formal analysis, Validation; **Mosè Rossi:** Supervision, Writing – review & editing

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China (52379086), the Natural Science Foundation of Jiangsu Province (BK20241520) and the Jiangsu Innovation Support Programme for International Science and Technology Cooperation (BZ2023047). The computational work was supported by the High-Performance Computing Platform at Hohai University. The support of Hohai University is also gratefully acknowledged.

Data availability statement

The data that support the findings of this study are available from the corresponding author, [K.K.], upon reasonable request.

References

Abate, A., Katoen, J. P., Lygeros, J., & Prandini, M. (2010). Approximate model checking of stochastic hybrid systems. *European Journal of Control*, *16*(6), 624–641. https://doi.org/10.3166/ejc.16.624-641

Al-Obaidi, A. R., & Alhamid, J. (2025). Analyses of the transient turbulence flow in a 3D impeller axial pump using novel vibration signals and inner dynamic simulation techniques. *Flow Measurement and Instrumentation*, *102*, 102779. https://doi.org/10.1016/j.flowmeasinst.2024.102779

Asr, M. T., Nezhad, E. Z., Mustapha, F., & Wiriadidjaja, S. (2016). Study on start-up characteristics of H-darrieus vertical axis wind turbines comprising NACA 4-digit series blade airfoils. *Energy*, *112*, 528–537. https://doi.org/10.1016/j.energy.2016.06.059

Bertsimas, D., & Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science*, *8*(1), 10–15. https://doi.org/10.1214/ss/1177011077

Bohang, L., Li, N., Yang, J., Alfarraj, O., Albelhai, F., Tolba, A., Shaikh, Z. A., Alizadehsani, R., Plawiak, P., & Yee, P. L. (2025). Image steganalysis using active learning and hyperparameter optimization. *Scientific Reports*, *15*(1), 7340. https://doi.org/10.1038/s41598-025-92082-w

Borges, J. E., Gato, L. M. C., & Pereira, R. M. R. J. (1996). Iterative use of a time-marching code for designing turbomachine blade rows. *Computers & Fluids*, *25*(2), 197–216. https://doi.org/10.1016/0045-7930(95)00038-0

Chang, H., Yang, J., Wang, Z., Peng, G., Lin, R., Lou, Y., Shi, W., & Zhou, L. (2025). Efficiency optimization of energy storage centrifugal pump by using energy balance equation and non-dominated sorting genetic algorithms-II[J]. *Journal of Energy Storage*, *114*, 115817. https://doi.org/10.1016/j.est.2025.115817

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

China Water Affairs Group Limited. (2025). Environmental, Social and Governance Report 2023. https://www.chinawatergroup.com/wp-content/uploads/2025/04/2023072700053.pdf.

Chitrakar, S., Solemslie, B. W., Neopane, H. P., & Dahlhaug, O. G. (2020). Review on numerical techniques applied in impulse hydro turbines. *Renewable Energy*, *159*, 843–859. https://doi.org/10.1016/j.renene.2020.06.058

Deb, K., Anand, A., & Joshi, D. (2002). A computationally efficient evolutionary algorithm for real-parameter optimization. *Evolutionary Computation*, *10*(4), 371–395. https://doi.org/10.1162/106365602760972767

Dixon, S. L., & Hall, C. (2013). *Fluid mechanics and thermodynamics of turbomachinery (seventh ed.)*. Butterworth-Heinemann.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, December 8–13, 2014. https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

Habibi, O., Chemmakha, M., & Lazaar, M. (2023). Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT botnet attacks detection. *Engineering Applications of Artificial Intelligence*, *118*, 105669. https://doi.org/10.1016/j.engappai.2022.105669

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Jia, Y., Zhang, L., Niu, J., & Berndtsson, R. (2025). China's south-to-north water diversion project: A review and reach beyond China's borders. *Water*, *17*(22), 3275. https://doi.org/10.3390/w17223275

Kan, K., Li, Y., Chen, H., Gao, S., & Chen, J. (2025a). A method for parameterized design and rapid optimization of axial-flow pumps. *Engineering Applications of Computational Fluid Mechanics*, *19*(1), 2506814. https://doi.org/10.1080/19942060.2025.2506814

Kan, K., Liu, K., Xu, Z., Li, Z., Rossi, M., & Chen, H. (2025c). Fluid deformation induced energy loss of pump-turbines based on the transport of mean kinetic energy. *Renewable Energy*, *248*, 122998. https://doi.org/10.1016/j.renene.2025.122998

Kan, K., Yu, Y., Zhao, F., Chen, J., Binama, M., & Chen, H. (2025b). Fish-friendly optimization of low-head axial-flow turbines. *Ocean Engineering*, *317*, 120070. https://doi.org/10.1016/j.oceaneng.2024.120070

Khuri, A. I., & Mukhopadhyay, S. (2010). Response surface methodology. *Wires Computational Statistics*, *2*(2), 128–149. https://doi.org/10.1002/wics.73

Kim, S., Choi, Y. S., Lee, K. Y., & Kim, J. H. (2012). Interaction of impeller and guide vane in a series-designed axial-flow pump. *IOP Conference Series: Earth and Environmental Science*, *15*, 032027. https://doi.org/10.1088/1755-1315/15/3/032027

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on *Learning Representations (ICLR 2014)*, April 14–16, 2014. https://hdl.handle.net/11245/1.434281

Kleijnen, J. P. (2009). Kriging metamodeling in simulation: A review. *European Journal of Operational Research*, *192*(3), 707–716. https://doi.org/10.1016/j.ejor.2007.10.013

Lambora, A., Gupta, K., & Chopra, K. (2019). Genetic algorithm—a literature review. *Proceedings of 2019 International Conference on Machine Learning, Big data, cloud and parallel computing*, February 14–16, 2019.

Li, H., Govindarajan, V., Ang, T. F., Shaikh, Z. A., Ksibi, A., Chen, Y. L., Ku, C. S., Leong, M. C., Shabaruddin, F. H., Zamaniah, W., Ishak, W., & Por, L. Y. (2025). MSPO: A machine learning hyperparameter optimization method for enhanced breast cancer image classification. *Digital Health*, *11*, 20552076251361603. https://doi.org/10.1177/20552076251361603

Li, W. (2025). Performance of vortex pump in CFD simulations with rough walls. *International Journal of Fluid Engineering*, *2*(1), 79–91. https://doi.org/10.1063/5.0237732

Menter, F. R. (2002). Two-equation eddy-viscosity turbulence models for engineering applications. *AIAA Journal*, *40*, 254–266. https://doi.org/10.2514/3.12149

Nabeel, S. M., Bazai, S. U., Alasbali, N., Liu, Y., Ghafoor, M. I., Khan, R., Ku, C. S., Yang, J., Shahab, S., & Por, L. Y. (2025). Optimizing lung cancer classification through hyperparameter tuning. *Digital Health*, *10*, 20552076241249661. https://doi.org/10.1177/20552076241249661

Nguyen, D. A., Ma, S. B., Kim, S., & Kim, J. H. (2023). Hydrodynamic optimization of the impeller and diffuser vane of an axial-flow pump. *Journal of Mechanical Science and Technology*, *37*(3), 1263–1278. https://doi.org/10.1007/s12206-022-1217-0

Roache, P. J. (1997). Quantification of uncertainty in computational fluid dynamics. *Annual Review of Fluid Mechanics*, *29*, 123–160. https://doi.org/10.1146/annurev.fluid.29.1.123

Rong, L., Böhle, M., & Yandong, G. (2024). Improving the hydraulic performance of a high-speed submersible axial-flow pump based on CFD technology. *International Journal of Fluid Engineering*, *1*(1), 91–108. https://doi.org/10.1063/5.0191683

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/10.1038/323533a0

Shen, J., Cheng, L., Jiao, W., Fan, H., Zhang, B., & Pavesi, G. (2025). Mechanism of energy dissipation in cavitation vortices of low head axial flow pump-as-turbine via reduced-order variational mode decomposition. *Engineering Applications of Computational Fluid Mechanics*, *19*(1), 2547254. https://doi.org/10.1080/19942060.2025.2547254

Stepanoff, A. J. (1957). *Centrifugal and axial-flow pumps: Theory, design, and application*. John Wiley & Sons.

Stern, F., Wilson, R. V., Coleman, H. W., & Paterson, E. G. (2001). Comprehensive approach to verification and validation of CFD simulations—part 1: Methodology and procedures. *Journal of Fluids Engineering*, *123*(4), 793–802. https://doi.org/10.1115/1.1412235

Sun, Z., Tang, F., Shi, L., & Liu, H. (2022). Multi-conditional optimization of a high-specific-speed axial flow pump impeller based on machine learning. *Machines*, *10*(11), 1037. https://doi.org/10.3390/machines10111037

Tan, C. S., Hawthorne, W. R., McCune, J. E., & Wang, C. (1984). Theory of blade design for large deflections: Part II—annular cascades. *Journal of Engineering for Gas Turbines and Power*, *106*(2), 354–365. https://doi.org/10.1115/1.3239572

Tao, R., & Wang, Z. (2021). Comparative numerical studies for the flow energy dissipation features in a pump-turbine in pump mode and turbine mode. *Journal of Energy Storage*, *41*, 102835. https://doi.org/10.1016/j.est.2021.102835

Thin, K. C., Khaing, M. M., & Aye, K. M. (2008). Design and performance analysis of centrifugal pump. *World Academy of Science, Engineering and Technology*, *46*, 422–429.

Uy, M., & Telford, J. K. (2009). Optimization by design of experiment techniques. *Proceedings of 2009 IEEE Aerospace Conference*, March 7– 14, 2009.

Wang, D., Tan, D., & Liu, L. (2018). Particle swarm optimization algorithm: An overview. *Soft Computing*, *22*, 387–408. https://doi.org/10.1007/s00500-016-2474-6

Wu, C. H. (1952). A general theory of three-dimensional flow in subsonic and supersonic turbomachines of axial, radial, and mixed-flow types. *Transactions of the American Society of Mechanical Engineers*, *74*(8), 1363–1380. https://doi.org/10.1115/1.4016114

Wu, Y. C., & Feng, J. W. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, *102*, 1645–1656. https://doi.org/10.1007/s11277-017-5224-x

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling Tabular data using Conditional GAN. Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. https://nips.cc/virtual/2019/poster/13822.

Yang, X. S. (2020). *Nature-Inspired optimization algorithms*. Academic Press.

Yu, L., Cheng, L., Sheng, W., Lei, S., Xu, W., & Shen, J. (2025). Research on the energy characteristics of the turning region in the cavitation performance curve of an axial flow pump based on energy flow density and entropy production. *Engineering Applications of Computational Fluid Mechanics*, *19*, 1. https://doi.org/10.1080/19942060.2025.2499132

Zaffar, A., Ibrahim, B., Sarwar, M. A., Chattha, J. A., & Asif, M. (2018). Optimisation of blade profiles of cross flow turbine. *International Journal of Power and Energy Conversion*, *9*(4), 311–326. https://doi.org/10.1504/IJPEC.2018.094952

Zhang, B., Cheng, L., Jiao, W., & Luo, C. (2025). Blade-vortex-interaction in an axial-flow pump sump: Impact on hydraulic performance and flow stability. *Energy*, *322*, 135610. https://doi.org/10.1016/j.energy.2025.135610

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, April 30–May 3, 2018.