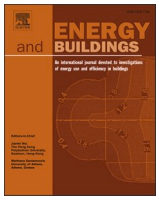
Energy & Buildings 348 (2025) 116352

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb



Data-driven optimisation of residential air-to-water heat pump performance using IoT and machine learning

Rasheed B. Ayoola [a,c], Olusegun M. Ilori [a,c,*], Noel Perera [a,c], Monica Mateo-Garcia [b,c], Kabir Akinyemi [a], David Boyd [b,c], Mike Leonard [b,c]

[a] Department of Engineering, School of Architecture, Built Environment, Computing and Engineering, Birmingham City University, UK
[b] Department of Architecture and Built Environment, School of Architecture, Built Environment, Computing and Engineering, Birmingham City University, UK
[c] Centre for Future Homes, Birmingham City University, Birmingham, UK

ARTICLE INFO

Keywords:
Air-to-water heat pump
Field data
Machine learning
Grid search
Hyperparameter tuning
Feature engineering

ABSTRACT

Residential heating accounts for about 27 % of the UK's energy consumption. While residential heat pumps (RHPs) are central to the transition toward sustainable energy, optimising their real-world performance requires robust experimental monitoring and predictive modelling. This study presents a data-driven approach for evaluating and optimising the performance of residential air-to-water heat pumps (A2WHPs) using real-time data and machine learning (ML). A full-scale experimental setup was deployed in a UK-based end-terrace building, incorporating IoT-enabled sensors to capture 275 days of operational data that was processed into a 6,600-hour dataset. Key thermal, electrical, and environmental parameters were measured at high temporal resolution and used to develop predictive models for the system's coefficient of performance (COP). Several ML models, including Random Forest, Support Vector Regression (SVR), eXtreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Long Short-Term Memory (LSTM), were evaluated using rigorous preprocessing, principal component analysis, and GridSearchCV hyperparameter tuning. LSTM, XGBoost, and ANN achieved the highest prediction accuracy with low error values across MAE, MSE, RMSE, CVRMSE, and NMBE. Diagnostic plots and residual analysis further confirmed the generalisability of the models and their sensitivity to non-linear operational behaviours. The findings demonstrate that integrating ML with real-world data can provide a robust predictive framework for operational diagnostics, performance evaluation, and efficiency improvement in residential heat pumps. This approach supports scalable, data-driven energy management and contributes to decarbonising the built environment.

Abbreviations: SPF, Season Performance Factor; COP, Coefficient of performance; ASHP, Air Source Heat Pumps; A2WHP, Air-to-water heat pump; GSHP, Ground source heat pump; DHW, Domestic hot water; SH, Space heating; CH, Central heating; RHPs, Residential Heat Pumps; UFH, Under-floor heating; HP, Heat pump; SEFF, System efficiency; ML, Machine learning; MID, Measuring Instrument Directive; SVR, Support Vector Regressor; XGBoost, Extreme Gradient Boosting; RF, Random Forest; LR, Linear Regression; PR, Polynomial Regression; MLR, Multiple Linear Regressor; ANN, Artificial Neural Network; RNN, Recurrent Neural Network; LSTM, Long Short-Term Memory; LCA, Life Cycle Assessment; KNN, K-Nearest Neighbour; KDE, Kernel Density Estimation; MCS, Microgeneration Certification Scheme; MIR, Mutual Information Regression; PCA, Principal component analysis; flowT, Flow temperature; returnT, Return temperature; outsideT, Outside temperature; heat_kW, heat output power; heat_kWh, heat output energy; elec_kWh, cumulative electrical energy consumed; Rhum, Relative humidity; roomT, Room temperature; U_value ($W/m^2.K$), overall heat loss coefficient; R^2, Coefficient of determination; MAE, Mean Absolute Error; MSE, Mean Squared Error; RMSE, Root Mean Squared Error; NMBE, Normalised Mean Bias Error; CVRMSE, Coefficient of Variation of Root Mean Squared Error; RFE, Recursive Feature Elimination; IoT, Internet of Things; IEF, Integrated Evaluation Framework; M & V, Measurement and Validation; HVAC, Heating, Ventilation, and Air Conditioning; BEM, Building energy management; PINN, Physics-informed neural network; PI-LSTM, Physics-informed long short-term memory; GridSearchCV, Grid search cross-validation; AFDD, Automated fault detection and diagnosis; EEPOM, Energy Efficiency Prediction and Optimisation Model; EN1434, European standard for heat meters; m^3/hr, metre cube per hour.

* Corresponding author.
E-mail address: Olusegun.ilori@bcu.ac.uk (O.M. Ilori).

https://doi.org/10.1016/j.enbuild.2025.116352
Received 26 April 2025; Received in revised form 19 August 2025; Accepted 21 August 2025
Available online 24 August 2025

1. Introduction

The installation of heat pumps and their market penetration across European countries have been steadily increasing, driven by the need for sustainable energy solutions and decarbonisation efforts. Recently, approximately 25.5 million heat pumps were installed across 19 European countries in residential, small office, and retail spaces, helping to avoid 21 billion cubic metres of gas importation, and saving about 45 megatonnes of CO_2 emissions per year in Europe [1]. The sharp rise in sales in countries such as Sweden and France underscores their potential

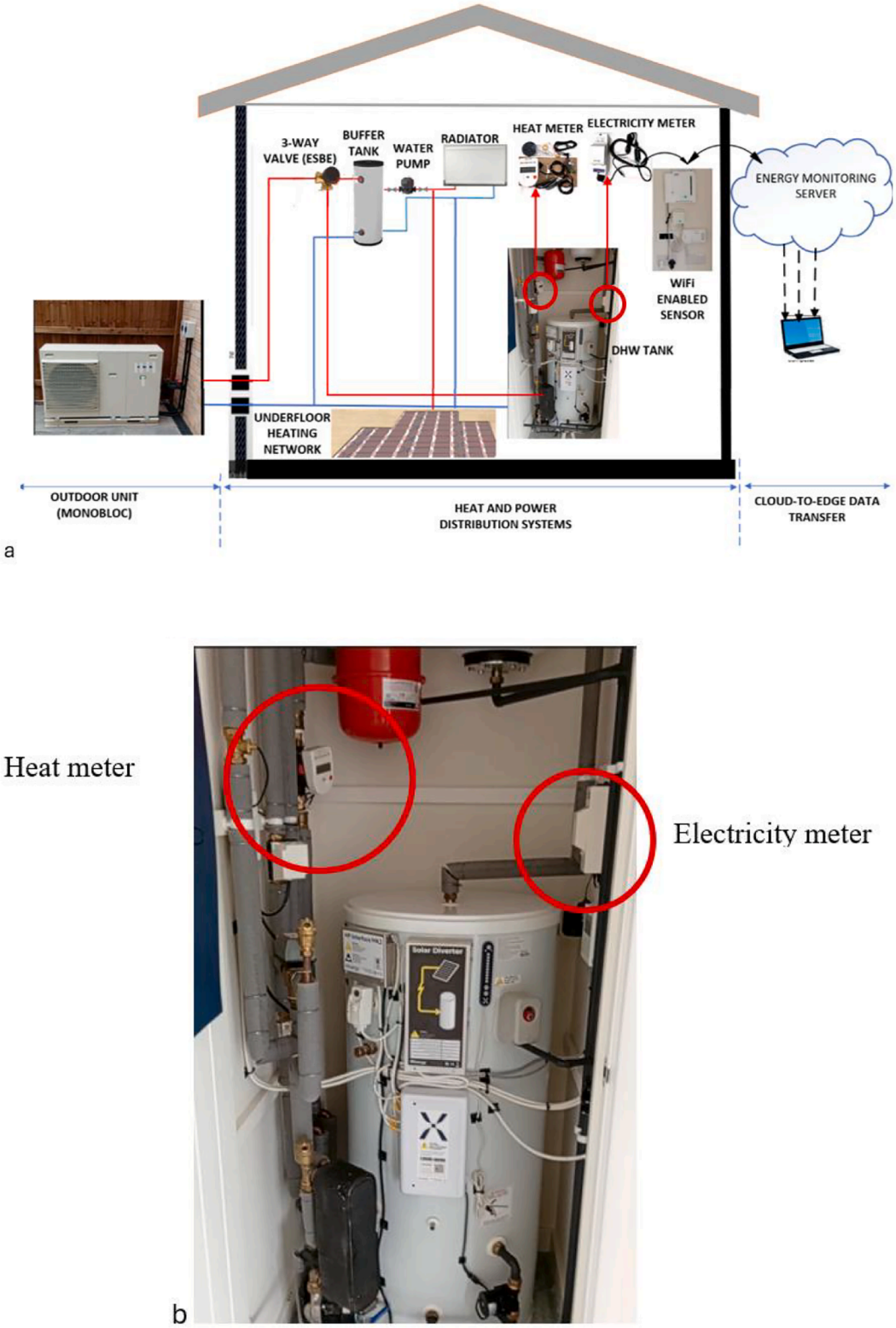


Fig. 1. Schematic diagram of the case-study building (a) A2WHP system (b) The position of heat and electricity meters within the pipe network of the A2WHP system.

Table 1
System-level Summary.

Component	Characteristic	
Heat pump type	Air source: air-to-water (A2W)	The heat pump extracts heat from the ambient air
Heat pump capacity	6 kW	The manufacturer's datasheet capacity for the heat pump system
Refrigerant	R32	Working fluid for the heat pump system
Design flow temperature	45 °C	The temperature of water-glycol that is extracted from the heat pump condenser before entering indoor heat exchangers, such as radiators, underfloor pipe networks, and DHW tanks.
Manufacturer-defined outside design temperature.	−2.2 °C	The minimum temperature at which the heat pump system can operate efficiently, as specified by the manufacturer's datasheet.
Building archetype	Semi-detached	Case study building
Building floor area	96 m^2	The case study building floor area.
Data source	Field measurement	Remotely transmitted data
Manufacturer model reference	Altherma 3	
Electricity meter	class 1 with an accuracy of ± 1 %	Measurement device specification
Heat meter	class 2 with an accuracy of ± 2 %	Measurement device specification
The combined uncertainty (COP, energy, flow)	~1–2 %	Measurement device specification
Manufacturer-defined air source heat pump operating regime (source/sink temperatures)	A2W35	Characterised by 2 °C outside source (air) temperature and a 35 °C Water-glycol sink temperature

to dominate the residential heating market. Life Cycle Assessment (LCA) studies on air source heat pumps (ASHPs) indicate significant environmental benefits, reducing heating-related emissions by 37 % in older dwellings and 54 % in new constructions across 18 European countries [2]. Heat pumps play a pivotal role in energy network decarbonisation, efficiently converting low-carbon or zero-carbon electricity into heat for space heating and domestic hot water supply. Unlike air conditioners, heat pumps can provide heating and cooling through a four-way diverter (reversing valve).

Heat pumps offer greater efficiency compared to alternative heating systems [3]. However, early field trials (e.g., Energy Trust, 2010; 2013) recorded lower-than-expected seasonal performance factors (SPF), primarily due to installation, commissioning, and control issues [4]. In contrast, more recent data indicate an average SPF for heat pumps across hybrid systems is clustered between 2.5 and 3.5, depending on system boundary definitions [5]. This performance gap has led researchers to explore the external factors influencing real-world efficiency, questioning whether broader lessons can enhance heat pump performance. External environmental variables significantly impact heating systems, introducing uncertainties within building-integrated heat pump operations [6]. Addressing these inefficiencies requires modelling, calibration frameworks, and parameter optimisation, which can enhance energy efficiency [7]. An integrated evaluation framework (IEF) combined field experiments, operational monitoring, and modelling to provide a robust tool for performance assessment and optimisation [8]. Energy and exergy analysis remain fundamental in identifying heat pump efficiency improvements through modelling and optimisation strategies [9]. Bridging the gap between predicted and real-world heat pump performance requires adopting IEF, enabling stakeholders to make informed decisions. Yu et al. [10] emphasise the importance of design-oriented modelling and uncertainty analysis in evaluating complex systems under varying conditions. System modelling approaches generally fall into analytical, empirical, and simulation-based methods [11]. Subramanian et al. [12] categorised modelling contributions into two domains – modelling approach and field applications.

Recent developments in the use of Internet of Things (IoT) sensors in management systems, such as real-time monitoring, predictive control, and online fault detection and diagnosis (OFDD), have significantly enhanced operational efficiency and energy optimisation in buildings. IoT-enabled Heating, Ventilation, and Air Conditioning (HVAC) systems facilitate uninterrupted communication between devices, enabling real-time data exchange on operational performance and environmental conditions [13]. Rapid attention is shifting towards IoT integration in building HVAC systems, to allow real-time prediction and intervention, while strengthening adaptive and responsive control strategies [14,15]. Several authors have proposed different methods to integrate building performance evaluation with IoT. Gao and Shardt [16] proposed a learning-based and iterative IoT system for energy management in connected buildings, and validated the system through simulation using real-world building data. Deep reinforcement learning was utilised for HVAC control actions, and the iterative optimisation algorithm was employed to integrate physics-based and learning-based models. The proposed IoT-based system achieves zero-energy building management leveraging real-time energy efficiency optimisation, renewable energy integration, and adaptive control strategies. Yaïci et al. [15] leveraged an IoT-based system for monitoring and controlling heating and cooling in residential buildings, utilising two control strategies based on outside air temperature and time-oriented temperature. The first control strategy serves as a baseline strategy that is regulated based on external temperature variations, allowing the control system to alternate between standby, active, and inactive modes. The second control strategy, referred to as the optimised baseline strategy after undergoing feedback and refinement, introduces a time-oriented method for temperature adaptation when the system is in heating or cooling mode. This approach relied on time-of-day-dependent thermal dynamics commonly observed in residential buildings. According to these authors, the IoT system reduced energy consumption by 21 % during heating scenarios. This is further validated in Abdelwahed et al. [17] and Corrà et al. [18]. Prioritising fault detection and diagnosis in heat pumps is crucial to safe plant operations, maintaining projected energy consumption, and sustaining the set indoor thermal comfort. Chew & Yan [19] developed a three-layer diagnostic Bayesian network, integrating maintainability rules to improve intelligent and automated fault detection and diagnosis (IAFDD) for HVAC systems utilising data from IoT sensors.

Machine learning (ML) models, often referred to as "black-box" methods, exhibit high flexibility in handling dynamic systems but require modifications to their input for interpretability [20,21]. These models use historical data, statistical patterns, and similarity analysis to generate accurate predictions. ML algorithms are broadly classified into supervised and unsupervised learning, with empirical methodologies crucial in Air source heat pump (ASHP) performance prediction. Prediction models generally use single or ensemble learning approaches [22]. Standard algorithms include decision trees, support vector regression (SVR), artificial neural networks (ANN), random forest (RF), XGBoost, and Long Short-Term Memory (LSTM), all of which have demonstrated success in building performance prediction [23]. These models learn from historical performance data, gaining insights from interactions within the overall building system. LSTM, in particular, excels in handling time-series data, capturing long-term dependencies and complex temporal patterns [24].

The Coefficient of Performance (COP), as defined in Equation (1), is the standard metric for evaluating heat pump efficiency under steady-state conditions; however, it does not always accurately reflect real-world performance [25]. The COP of ASHP is highly influenced by ambient conditions, including temperature and humidity [26], but

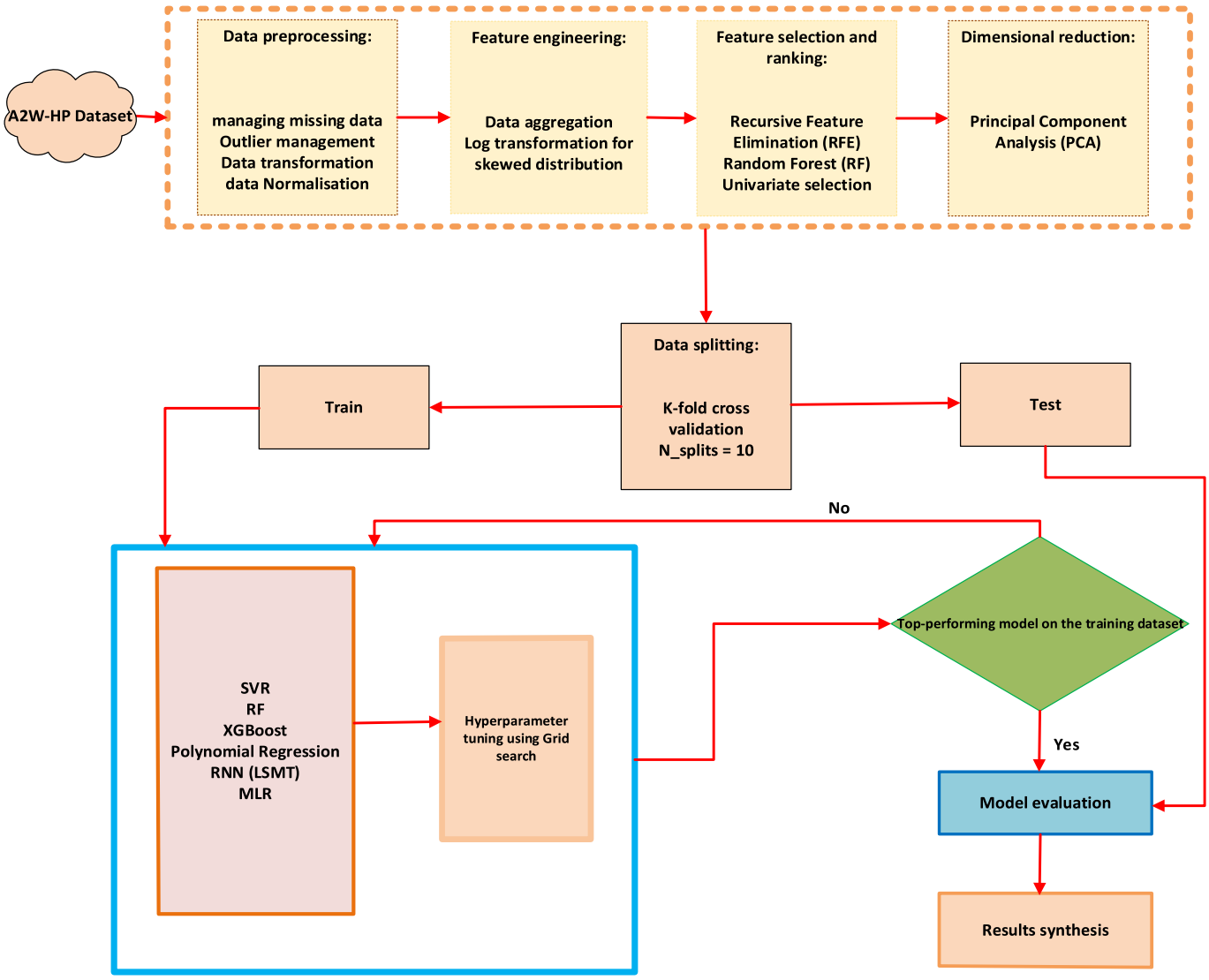


Fig. 2. Sequential workflow diagram of this study.

non-weather-related factors also play a significant role [23]. Operational monitoring and predictive modelling are essential for COP estimation, optimisation, energy savings assessment, and fault detection. Incorporating real-time data, advanced modelling, and uncertainty quantification helps align predicted and real-world performance outcomes [27]. Achieving accurate COP predictions within a multi-dimensional feature space requires optimal model selection and parameter tuning [28]. Advanced ML models have demonstrated strong potential for improving the prediction accuracy of heat pump COP, heat output, and power consumption, as well as facilitating system control optimisation [28–30].

$$COP = \frac{Heat\ Energy\ generated\ (Q_{HP})}{Input\ electrical\ energy\ (E_{HP})} \quad (1)$$

Existing studies have developed various ML models and optimisation algorithms for heat pump performance and hydronic modelling across different applications. Chesser et al. [23] applied ML algorithms to predict ASHP in situ performance, demonstrating that ensemble methods like RF outperform advanced statistical and single predictive algorithms. Tangwe and Simon [31] employed multiple linear regression (MLR) to model the performance of ASHP water heaters based on flow and return water temperatures, water quantity, relative humidity, and ambient temperature. Cho et al. [28] employed TRNSYS 18 to model a residential ground source heat pump (GSHP) and trained ML models for GSHP COP prediction. Many studies predicting the coefficient of performance (COP) of heat pumps rely on previously conducted field trials or simulated data from models. However, data sparsity across time and space can introduce confirmation and historical bias. Moreover, the increasing adoption of residential heat pumps necessitates data-driven energy efficiency assessments using IoT-enabled energy meters to reduce costs and enhance power grid resilience [27,32].

This study employs an independent performance monitoring approach for real-time data collection and analysis, evaluating the effectiveness of eight ML algorithms – linear regression (LR), MLR, polynomial regression (PR), RF, SVR, XGBoost, ANN, and LSTM, in predicting COP for monitored air-to-water heat pumps (A2WHPs) in a residential building. These algorithms were selected based on the multi-dimensional nature of heat pump datasets and the non-linear interactions between features. The proposed Energy Efficiency Prediction and Optimisation Model (EEPOM) integrates prognostic methodologies to enhance model robustness and generalisability, given the complex relationships in ASHP performance data. The key contributions of this study are:

Table 2
Description of features of the trained dataset.

Features/Variables	Notation	Class	Description
Timestamp (Hour)	t	Independent/ feature	Driving force for other variables
Flow temperature (^{0}C)	T_flowT	Independent/ feature	Driving force for heat transfer in the hydronic distribution circuit
Return Temperature (^{0}C)	T_returnT	Independent/ feature	Used to calculate delta T (ΔT)
Volumetric Flowrate (m3/hr)	FR_flowrate	Independent/ feature	Used to determine heat transfer capacity
Outside temperature (^{0}C)	T_outsideT	Independent/ feature	It affects efficiency and performance dynamically
Coefficient of performance	COP	Dependent/ Target	Measure of heat pump efficiency (to be predicted and optimised)
Delta T (ΔT ^{0}C)	ΔT	Derived	Directly reflects performance efficiency.
Electricity consumed (kW)	P_elec	Intermediate	Input to COP/SPF calculation
Heat output (kW)	Q_heat	intermediate	Input to COP/SPF (Also dynamically controlled to achieve the desired indoor temperature {DV})
Relative humidity	Rhum	Independent/ feature	It affects efficiency and performance dynamically
Usage profile (e.g. Temperature setpoint)		Independent/ feature	It affects efficiency and performance dynamically
Building features (e.g. heat loss or gain from wall, floor, roof, and furniture) (W/m^2. K)	U_value	Independent/ feature	Little impact on season efficiency of heat pump (Burns et al., 2021)
Indoor temperature (^{0}C)	T_roomT	Dependent/ Target	It affects efficiency and performance dynamically

- Real-Time Data Collection – Unlike most studies, which rely on simulated or past field-trial data, this work deploys IoT-enabled energy heat meters to collect fresh, real-world data.
- Comprehensive ML Analysis – This study evaluates a diverse set of state-of-the-art ML algorithms, including ensemble, kernel-based, and deep learning models, whereas many studies consider only a few. Furthermore, it incorporates an unsupervised learning approach for dimensionality reduction, complementing supervised feature selection for more reliable predictions.
- Performance Optimisation through GridsearchCV Hyperparameter Tuning – Assesses manual versus optimised hyperparameter configurations to identify potential performance improvements across all ML models.

Therefore, this work is novel in its integration of IoT-based real-time monitoring, comprehensive machine learning evaluation, and model diagnostics to optimise residential A2WHPs performance and provide a scalable, field-tested approach for energy efficiency and smart heating system control.

2. Methodology

The methodology of this study consists of two phases, involving measurement and validation (M&V) processes that utilise experimental and data curation techniques. The first phase includes installing and commissioning an IoT-enabled energy monitoring device (heat meter) on operational A2WHPs (6 kW capacity, R32 refrigerant, and design flow and outside temperatures of 45 and – 2.2 °C, respectively) with a hot water tank (Mixergy MX012661, 180 L capacity with legionella target temperature of 55 °C and equipped with immersion heater of 3.0 kW) to capture real-time performance data from a single-family semi-detarched dwelling (case study building) with a total floor area of 96 m^2, located in the United Kingdom. The heat meter – Sontex-Superstatic-789, with a capacity of up to 7 kW, features a measurement accuracy of 1–2 %, Pt1000 Temperature Sensors, continuous flow rates of 2.5 m^3/hr, and is glycol tolerant. It is class 2 MID approved to EN1434, ensuring compliance with industry norms. Sensor networks were used to measure various parameters, including instantaneous indoor and ambient temperatures, heat energy input/output, flow temperatures, and the contributions of heat energy toward central heating and domestic hot water. The schematic diagram of the installed A2WHP and associated measuring system setup is shown in Fig. 1a. Fig. 1b. Specifically, the measurement devices used for data collection possess the following specifications: The Electricity meter is class 1 with an accuracy of ± 1 %, the Heat meter is of class 2 with an accuracy of ± 2 %, and the combined uncertainty (COP, energy, flow) is ~ 1–2 % [33]. Table 1 gives a summary of System-level parameters in the study.

2.1. Air-to-water heat pump (A2WHP)

The A2WHP provides space heating (SH) and domestic hot water (DHW), and the monobloc outdoor unit (installed at the rear of the building to comply with MC regulation) is integrated with an indoor hot tank system to maximise heat recovery (Fig. 1). It incorporates three energy controls, a solar diverter, an immersion heater, and a heat exchanger, to manage the building's dynamic load profile. The solar diverter supplies power to the immersion heater as needed and redirects excess energy to the grid. The secondary circuit, comprising 85 % water and 15 % glycol (Propylene), circulates fluid through radiators, underfloor heating systems, and the DHW tank. A 3-way Honeywell's esbe' valve is installed to enable the A2WHP unit to provide SH and DHW effectively. The monobloc A2WHP generates hot water with distinctive properties, which is then delivered to each secondary circuit unit through pipes and pumps. This study focuses on the thermal and hydraulic behaviour of the system at the secondary circuit level, analysing water flow rates, heat capacity, and flow/return temperatures at different operating conditions. However, the thermodynamic behaviour of the refrigerated circuit (standalone A2WHP) is not considered here.

2.2. Challenges and mitigations in field-based monitoring

To achieve seamless and reliable data collection from the site, prioritising high monitoring quality is vital. One of the early challenges encountered was incorporating external environmental variables, such as ambient temperature, into the monitoring scheme, which proved to be challenging due to the extended period required for data integration. However, the Met Office weather data was successfully integrated with other operational features of the ASHP systems, facilitating the collection of comprehensive data blocks. Furthermore, securing constant approval from the occupier of the case study site (test bed) for necessary interventions requires several days, which delays the proposed commencement of the data collection exercise. However, this was subsequently resolved through the coordinated efforts of the field monitoring team and the site manager.

2.3. A2WHP monitoring, data acquisition and processing

The second phase of the study focused on data curation and preparation. This included preprocessing, imputation of missing values using the K-nearest neighbour (KNN) method, normalisation, and dimensionality reduction through Principal Component Analysis (PCA). These steps were conducted using Python 3.11 (64-bit) in an Anaconda environment, with a system powered by a 2.10 GHz AMD Ryzen 5 PRO 4650U processor. The overall research workflow is summarised in Fig. 2,

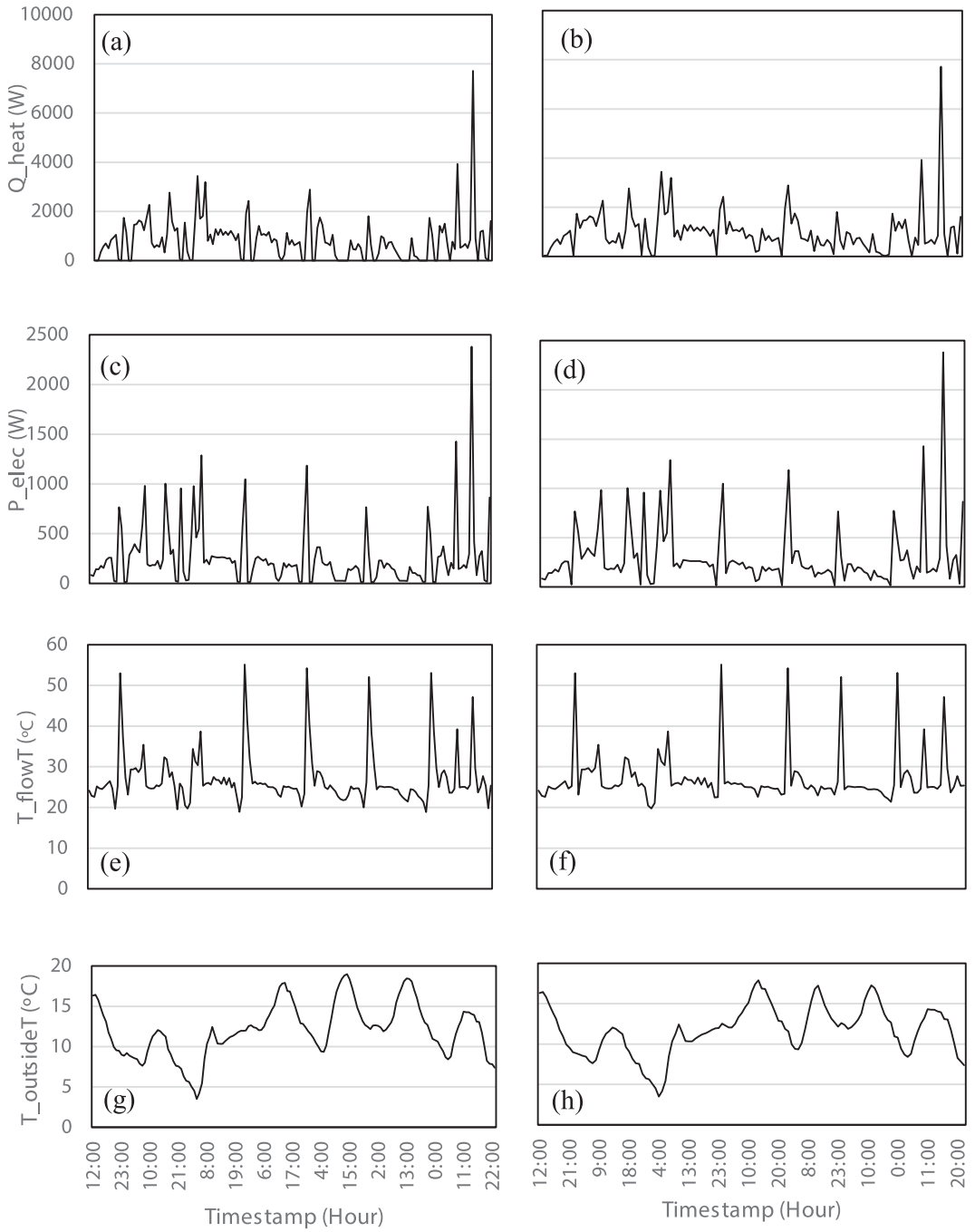


Fig. 3. Time series profiles of key operational features from the heat pump system over a week, with subplots (a, c, e, g) showing the raw data before preprocessing and subplots (b, d, f, h) showing the data after preprocessing.

and an overview of the dataset features is provided in Table 2. The dataset consisted of 275 days of real-time data collected from the case study building between January 12, 2023, and September 1, 2024. Data collection was enabled through sensor-integrated heat meters from OpenEnergyMonitor [34], providing detailed readings on flow and return temperatures, electric power consumption, heat output, flow rate, indoor and ambient temperatures, and corresponding hourly COP values. Measurements were initially recorded at 15-minute intervals and later aggregated to hourly resolution, resulting in a dataset of 6,600 hours. In real-world scenarios, high-frequency data are often accompanied by noise and transient effects, particularly during compressor startup and ramp-down. This can temporarily raise or reduce COP values due to a lag between the electrical power consumed and heat delivery at these instances. Several studies have documented this phenomenon as transient cycles that skew performance metrics [35]. To minimise this temporal fluctuation and focus on steady-state trends, the dataset was aggregated to hourly intervals, as recommended in prior studies [23, 36]. This smoothing decreases signal noise and enhances model stability, which is essential for our targeted control-oriented modelling and MPC implementation. This is particularly crucial due to the factors related to heat pump cycling behaviour [37]. The dataset was split into training and test sets. K-fold cross-validation was performed on the training set alone to select and tune models, and the final model was evaluated on the test set. GridSearchCV was employed to enhance prediction accuracy for hyperparameter tuning. The predictive performance of each ML algorithm was assessed using actual vs. predicted

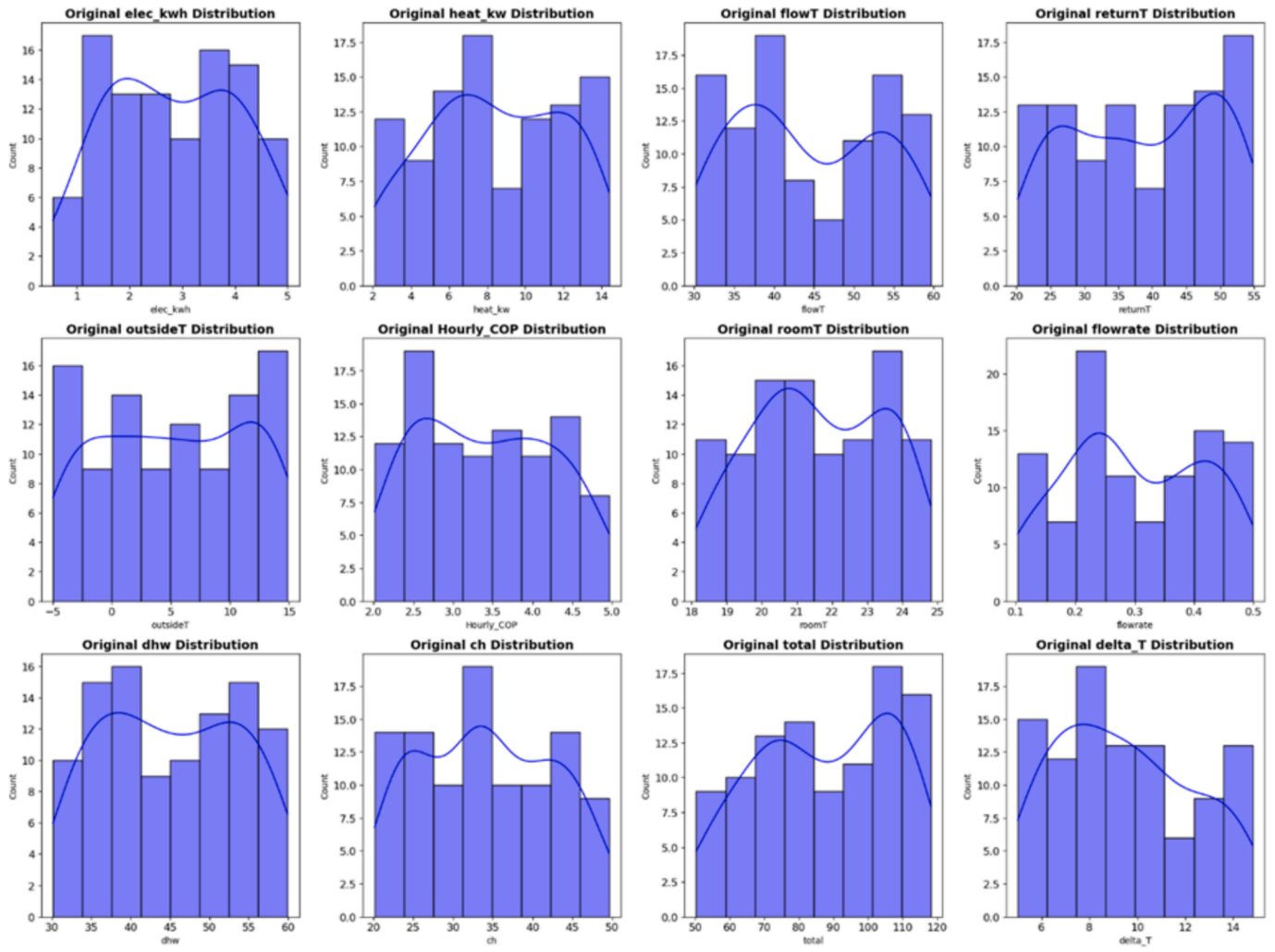


Fig. 4. Dataset distributions before quantile transformation.

outputs, error analysis, and R^2 scores, both before and after tuning. This robust combination of real-time monitoring and advanced data-driven methods helped identify performance inefficiencies and optimise the A2WHP.

2.3.1. *Data preprocessing*

Data preprocessing is essential for preparing raw data for ML models, involving cleaning, organising, and standardising data to ensure it is suitable for modelling and to enhance model reliability. Due to variations in data sources, collection techniques, and sensor accuracy, raw datasets often contain noise, missing values, and inconsistencies [38]. ML algorithms are susceptible to data quality, and preprocessing can directly influence their accuracy and predictive performance [39]. In this study, the 'date' feature was removed as it offered limited predictive value for some models. A quartile transformation technique was applied to normalise feature distribution, ensuring a rank-based adjustment [40]. This transformation redistributes modal values, enhancing data uniformity while reducing the impact of outliers (Zuang et al., 2020). Furthermore, this technique fosters model robustness, mitigating the influence of near-outliers and improving generalisation [41]. Fig. 3 displays the time series profiles of observed features, including power consumption, heat output, flow temperature, and outside temperature. Subplots (a, c, e, g) present the raw data before preprocessing, while subplots (b, d, f, h) show the data after preprocessing. The comparison demonstrates the effects of initial conditioning, which involved linear interpolation of missing values, smoothing, and outlier filtering on the time-series data [42,43].

2.3.2. *Data normalisation and transmission*

Data normalisation ensures data integrity by structuring and organising information, reducing redundancy (duplicate data), and minimising skewness in statistical analysis [44]. Given its role in improving feature distribution, normalisation is critical for ensuring ML models operate efficiently. Studies have demonstrated its importance in enhancing model performance and preventing feature dominance [45].

Regression models and other statistical approaches often assume homoscedasticity (constant variance in errors). However, violations of this assumption will lead to heteroscedasticity, where variance fluctuates across different predictor levels [46]. Detecting and addressing variance inconsistencies is crucial for ensuring reliable model predictions [47]. Data transformation techniques are widely used to mitigate heteroscedasticity [48]. This study applies quantile transformation, redistributing data based on principles of uniform or normal distribution. This approach spreads out frequent observations, effectively reducing the influence of outliers and making it a robust preprocessing strategy. A quantile transformation was performed on the training dataset, ensuring that all features were scaled between 0 and 1. Fig. 4 depicts the original variable distributions, revealing a multimodal and skewed structure indicative of heteroscedasticity. After applying a quartile-based data transformation, Fig. 5 demonstrates a stabilised variance, approximating a normal distribution and improving model consistency.

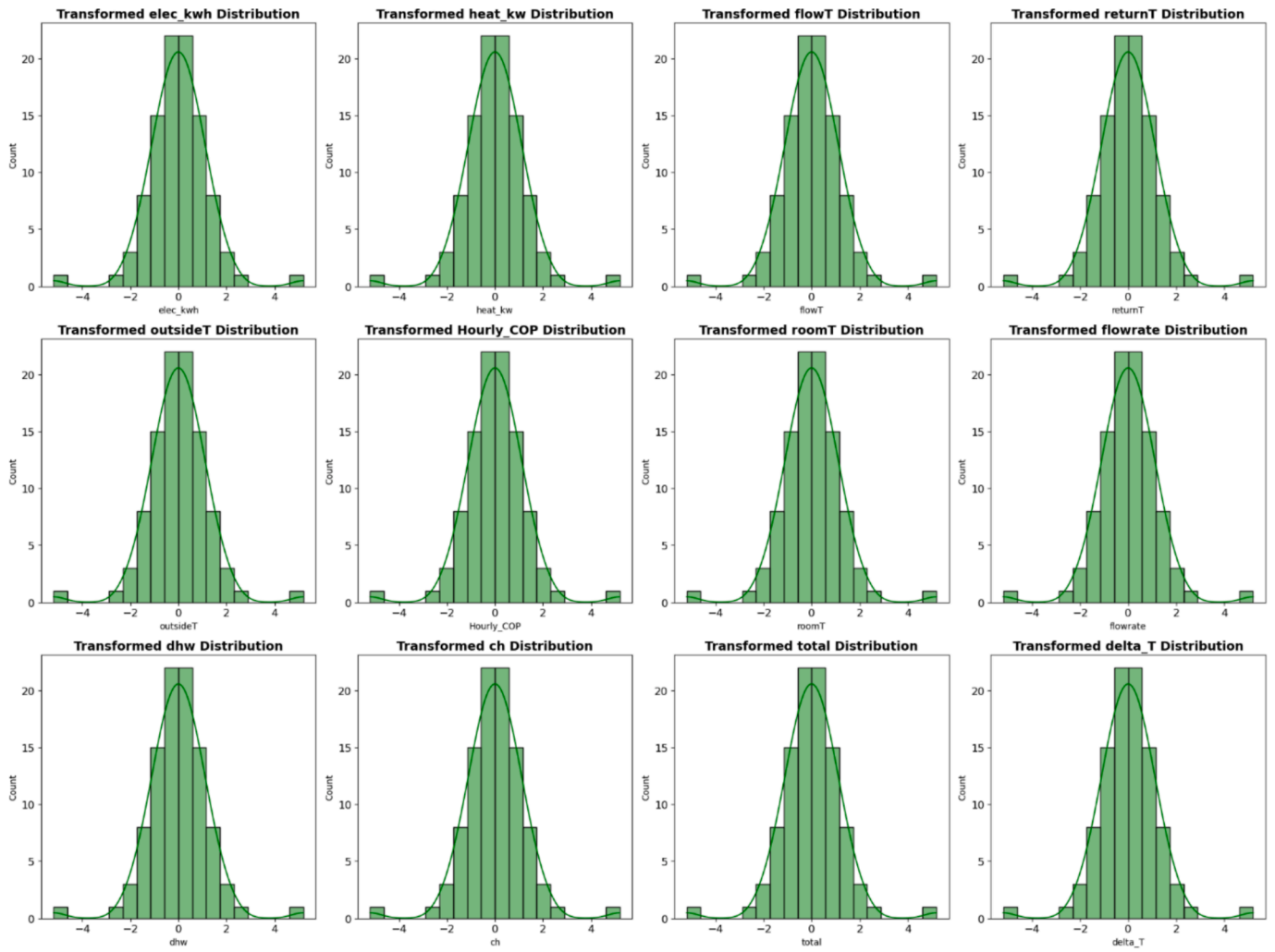


Fig. 5. Dataset distribution after quantile transformation.

2.3.3. *Feature engineering*

To account for seasonal variations in estimating COP, seasonal decomposition techniques were employed [49], [50], [51]. This process isolates seasonal components, allowing a deeper understanding of the relationship between key features independent of cyclical variations. Therefore, trend analysis of the target feature (COP), seasonal decomposition for heat output, and anomaly detection in daily electricity consumption were carried out. Furthermore, seasonal components were removed using STL (Seasonal-Trend Decomposition using Loess (Locally Estimated Scatterplot Smoothing)) to ensure data stationarity for further modelling. This technique enhances predictive accuracy by eliminating cyclical distortions.

2.3.4. *Exploratory analysis of time series data*

Effective system optimisation and energy-saving strategies require a comprehensive analysis of performance trends, efficiency metrics, and seasonal influences, particularly in high-dimensional energy systems. New features were derived from the key variables influencing residential A2WHP performance. Seasonal analysis, day-of-the-week analysis, temperature differences over time, patterns in energy consumption, and comparative analysis of evaluation metrics across seasons and day-of-the-week are visualised using techniques such as lines, bar charts, and box plots. Fig. 6 (a-f) shows the patterns that are derived from the seasonal decomposition of the time series data, including trend analysis of heat output, seasonal components, cumulative energy consumption, heat-to-electricity ratios, seasonal heat demand, and average energy usage patterns by day of the week. Fig. 6a illustrates the visible jump in daily aggregated heat demand in January, likely aligning with a rise in heat demand due to colder outdoor conditions. Following that period, the heat demand remains relatively stable, with mild fluctuations, through to November. The cumulative slope observed suggests the presence of temporal dependencies (the influence of historical and current states on future operations) in the operation of the heat pump system. Fig. 6b shows the seasonal component of heat output usage, highlighting the recurring variation in heat demand. For instance, the high peaks observed in January, March, July, and September indicate above-average seasonal heating demand during those periods. Meanwhile, February, April, August, and October witness below-average seasonal demand. A large magnitude of fluctuation reflects season variation, mainly attributed to either weather, occupancy, or system efficiency. Fig. 6c reveals that there is a stable but cyclic heat-to-electricity ratio, alternating between 3.6 and 3.8 throughout the period. Based on the observed regular crests and troughs, we can identify predictable seasonal and operational patterns, likely attributed to prevailing external temperature and usage profile. Fig. 6(d) presents the distribution of daily heat output across four seasonal periods, revealing an asymmetric distribution with a median range of 140,000–150,000 kWh, highlighting variability across seasons [52,53]. However, winter and fall, characterised by consistent heating demand, exhibit higher variability in daily heat demand with a wider interquartile range. Overall, heat demand remains stable but elevated year-round, likely due to system configuration. Fig. 6e compares the average daily heat energy

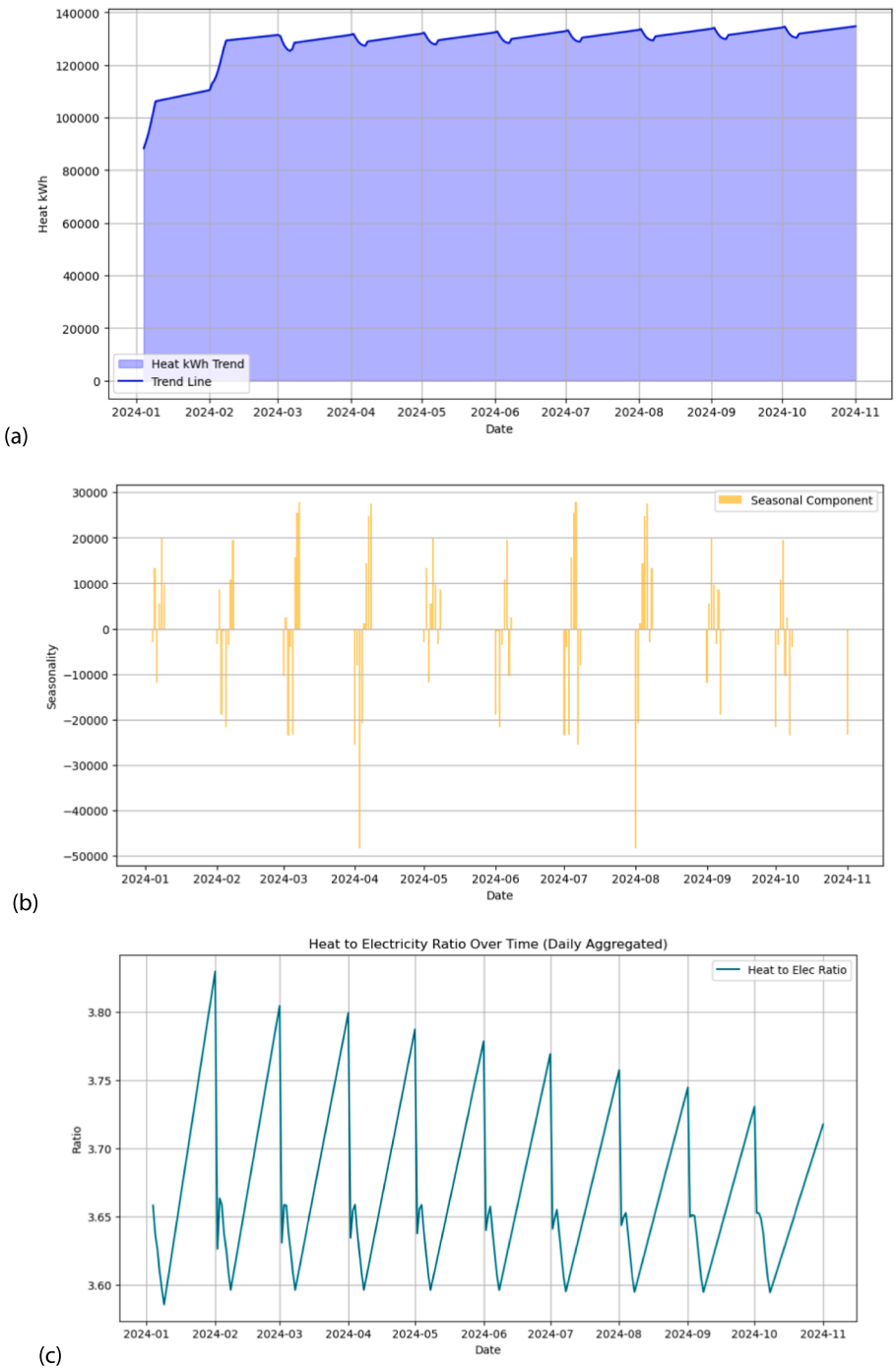
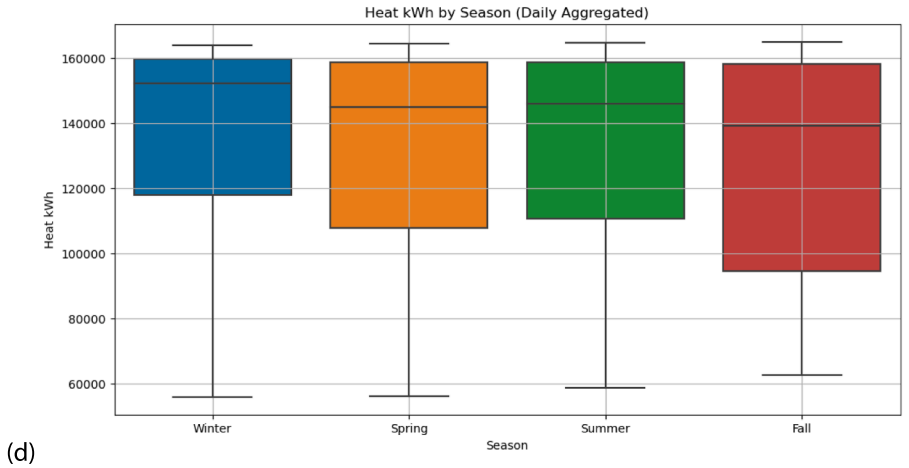


Fig. 6. Characteristics of the time-dependent sequential dataset, (a) Trend analysis of heat output, (b) season component of heat output, (c) Heat to electricity ratio over time, (d) Heat output for four typical UK seasons, (e) Average daily energy usage pattern by day of the week.

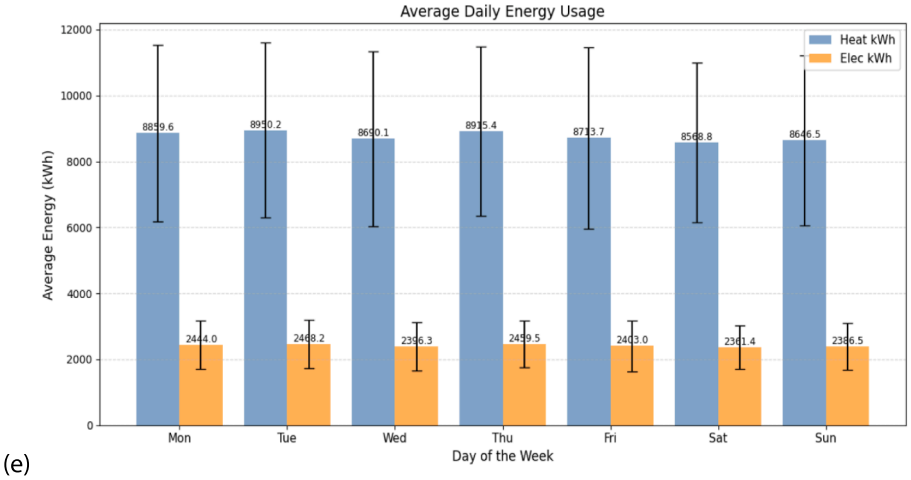
demand pattern to electrical consumption by day of the week. The plot shows consistent heat demand across the weekdays with a slight dip on Saturday, suggesting reduced occupancy [54]. Meanwhile, the electrical energy mirrors the heat energy profile, raising the possibility that the system maintains a consistent energy balance across the week. The embedded error bars reflect the natural variability in heat demand and energy consumption across the week. This observed spread conformed with the 1–2 % uncertainty range of the class 1 (electric meter) and class 2 (heat meter) used for our measurements.

2.4. Machine-learning-based models

It has been established that no single ML model excels across all problem domains, as performance varies by application [55,56]). For instance, decision trees provide interpretability for smaller datasets, while neural networks capture complex patterns in high-dimensional data [55]. This study uses SVR, RF, LSTM, PR, and XGBoost based on their effectiveness in handling complex datasets and their proven effectiveness across various predictive tasks [57–59]. After identifying the model, its hyperparameters are systematically tuned to enhance the prediction effectiveness of the model on the given dataset, avoid



(d)



(e)

Fig. 6. (*continued*).

overfitting and enhance the model's generalisability [,60,61]. Various techniques were employed, including exploratory data analysis, statistical tests, feature engineering, and visualisation methods to assess the non-linearity within the case study dataset. Fig. 7(a) presents a three-dimensional (3D) scatter plot to provide an intuitive insight into variable relationships through clusters, trends, and outliers [62,63]. However, scatter plots are limited by noise and data sparsity, restricting their ability to capture complex interactions between variables fully. To address these limitations, Fig. 7(b) introduces a 3D surface plot, which provides smooth approximations of local topology, effectively highlighting non-linear relationships and intricate feature interactions in the dataset [64], [65]. These visualisations collectively confirm that the dataset has high dimensionality and complexity, necessitating advanced modelling approaches. Furthermore, the high variability in the measured COP over time, as shown in the 3D plot, suggests fluctuations beyond ambient and secondary circuit flow temperatures, implying multiple underlying factors influencing system performance [66].

Following the characterisation of the dataset as dynamic and highly dimensional, the following five ML algorithms, including RF [67], SVR [68], XGBoost [69], LSTM [70], and ANN were selected for predictions based on their ability to model the inherent non-linearity observed between the features in the dataset and target variables. These algorithms' hyperparameters are fine-tuned before retraining on the training dataset and the unseen case study A2WHP dataset (i.e., test dataset). A grid search was employed, given the limited and discrete characteristics of the hyperparameter space for the selected ML and the simplicity of the parameter architecture. This systematic exploration approach examines all possible hyperparameter combinations within a specified grid to optimise model performance [71], [72]. Table 3 presents various ML algorithms and their effectiveness according to the characteristics of the dataset, which are used as model suitability criteria in the proposed ML models.

2.4.1. *Random forest (RF)*

The RF is an advanced ensemble learning approach widely used for classification, regression, and non-standard processes, such as clustering, ranking, feature selection, and outlier detection [73]. By combining multiple decision trees, RF enhances prediction accuracy and mitigates overfitting. This method effectively captures complex patterns and reduces errors associated with individual tree models. RF is particularly well-suited for high-dimensional datasets, making it applicable across diverse domains [74]. It employs bootstrap aggregation (bagging), where individual decision trees are trained on random subsets of the dataset, and the mean prediction from all trees determines the final output in regression tasks [73]. This aggregation process controls errors, improves model robustness, and enhances predictive reliability [67]. Additionally, due to its tree-based framework, RF excels in modelling non-linear relationships and capturing complex variable interactions.

2.4.2. *Support vector regression (SVR)*

The SVR is a supervised learning algorithm based on support vector machine (SVM) principles. It predicts continuous values by mapping data into a high-dimensional feature space and identifying a best-fitting function within a defined tolerance margin (ε). This approach allows controlled error tolerance, making SVR particularly effective in

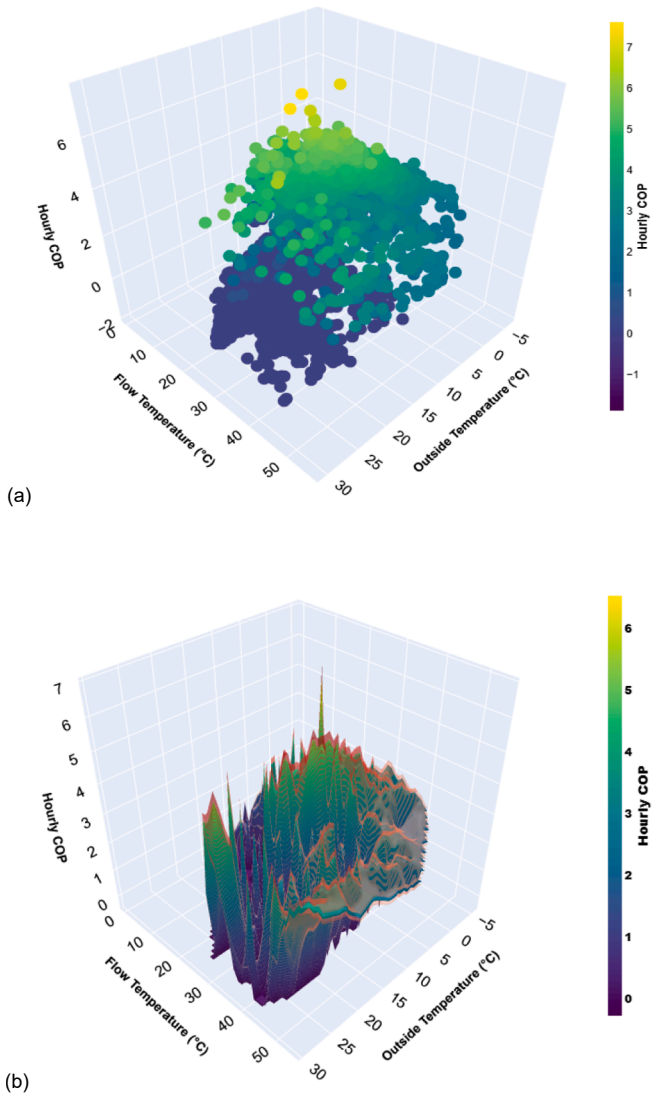


Fig. 7. Three-dimensional (3D) representation of heat pump field data (a) Scatter plot, (b) The surface plot showing the non-linear interaction of features of the dataset.

modelling non-linear interactions among variables. Fig. 8 shows the structure of SVR.

where $f(x)$ represents predicted values for input x, $k(x_i, x)$ denotes the kernel function defining the shared features between training samples x_i and the input x. The support coefficients a_i^* represent the non-zero weights of data points outside the ε tube influencing model predictions. y_i denotes the target outputs in the training data and b^* represents the bias term, ensuring the model accurately fits the data while keeping the deviation within ε boundaries. SVR uses loss functions that show that data points within the ε tolerance limits do not affect the error, while those outside the margin influence the prediction error. Mathematically, Equation (2) shows the ε–Intensive loss function is expressed as:

$$L_\varepsilon(y_1, f(x)) = \begin{cases} O, |f|y_1 - f(x_i)| \le \varepsilon \\ |y_1 - f(x_i)| - \varepsilon, otherwise \end{cases} \quad (2)$$

2.4.3. Long short-term memory (LSTM)

The LSTM is a specialised recurrent neural network (RNN) architecture that retains past inputs using memory cells and a gating mechanism. These features enable effective handling of sequential data, making LSTM well-suited for time series predictions [75]. Unlike conventional RNNs, LSTM mitigates vanishing and exploding gradient issues through its gated memory structure. By preserving long-term dependencies, LSTM captures complex temporal patterns in time series data, including trends, seasonality, cyclic fluctuations, and residual (random) noise [76,77]. Fig. 9 shows the LSTM architecture, depicting the flow of information and memory control through its gating mechanism.

Table 3
Features of machine learning algorithms for this study.

S/N	Machine learning algorithms	Characteristics of the dataset as suitability criteria for the ML models
1	SVR	Effective for small datasets with non-linear patterns [68]. Small-to-moderate datasets, linear or slightly non-linear relationships, high-dimensional data, and noisy datasets.
2	XGBoost	Handles non-linear relationships, is robust to outliers, and performs well on structured data [69]. Tabular datasets, large datasets, non-linear relationships, imbalanced data, need for computational efficiency.
3	MLR	Baseline model and easy to interpret
4	RF	Captures non-linear interactions and reduces overfitting [67]. Large datasets, non-linear relationships, mixed data types, outliers, feature importance insights.
5	PR	Model non-linear relationships. Simple datasets with clear polynomial trends, low-dimensional data, and small datasets.
6	LSTM	Handles complex and high-dimensional relationships and addresses vanishing gradients with gating mechanisms. [70]. Sequential or time-series data, long-term dependencies, large datasets, and non-stationary data.

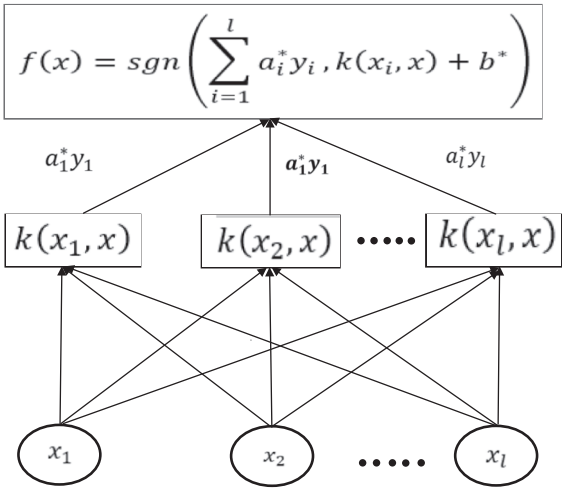


Fig. 8. Structure of support vector regression.

2.4.4. Artificial neural network (ANN)

The ANN regressor is a deep learning model designed to predict labelled outcomes by learning patterns and feature interactions within a dataset rather than relying on explicitly programmed rules. Inspired by the processing system of the human brain, ANNs generalise from training data to accurately predict unseen inputs [78]. The primary benefits of ANNs include their ability to model non-linear relationships, fast runtime, and robust performance under varying conditions. ANNs optimise learning through backpropagation and gradient descent, adjusting internal parameters to minimise prediction error [79]. The ANN architecture is shown in Fig. 10.

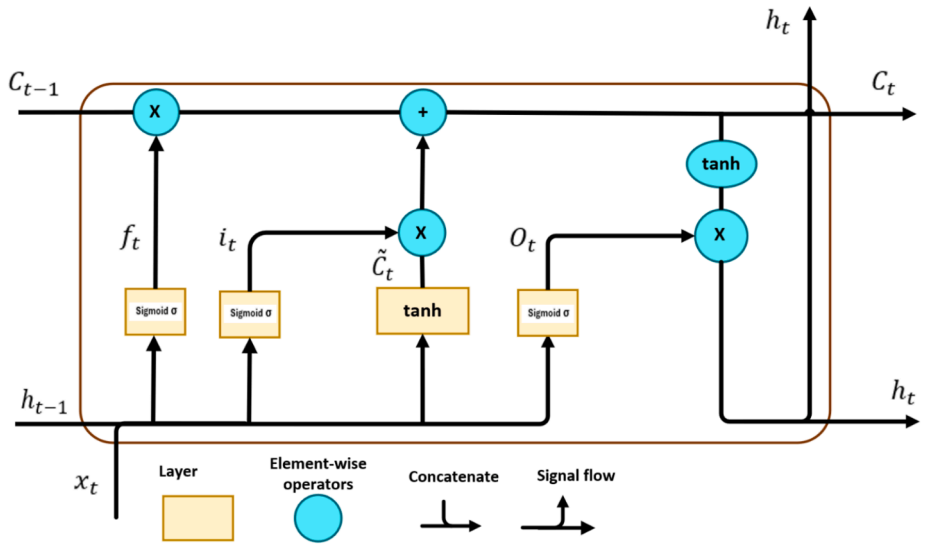


Fig. 9. Schematic diagram of LSTM network.

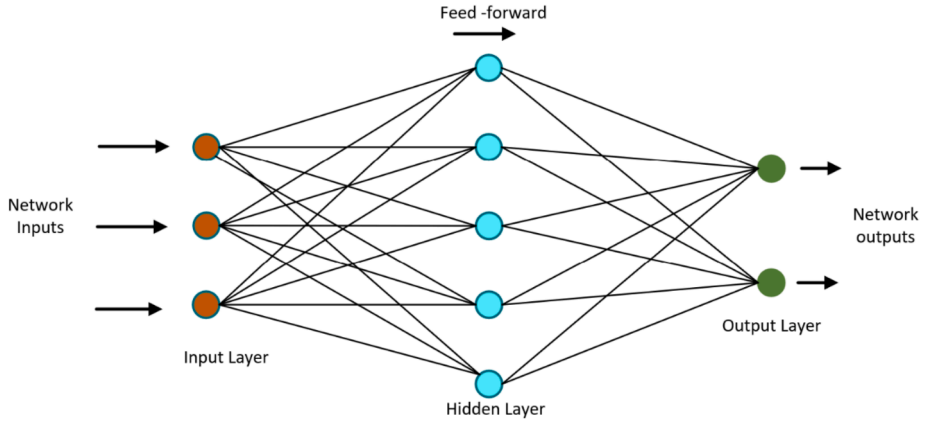


Fig. 10. Schematic Diagram of Artificial Neural Network Architecture.

Table 4
The hyperparameter tuning architecture for the ML algorithms.

Algorithm	Hyperparameters	combinations evaluated	unique combinations	Total fits (5-fold CV)
ANN	hidden_layer; number of units; activation functions ('relu', 'tanh'); optimiser ('adam', 'rmsprop')	2*3*2*2 = 24; each 5 times → 120	24	120
LSTM	hidden_layer; number of units; activation functions ('relu', 'tanh'); optimiser ('adam', 'rmsprop')	2*3*2*2 = 24; each 5 times → 120	24	120
RF	max_depth; n_estimator	3*2 = 6; each 5 times → 30	6	30
XGBoost	max_depth; n_estimator	3*2 = 6; each 5 times → 30	6	30
SVR	C'; 'gamma'; 'kernel'	4*4*3 = 48; each 5 times → 240	48	240

2.5. Evaluation metrics

Evaluating the predictive performance of machine learning models is critical to ensuring their robustness and reliability. This study employed a multi-metric evaluation approach, recognising that no single metric can fully capture the complexity of model behaviour. Six metrics are used, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root MSE (RMSE), Coefficient of Variation of RMSE (CVRMSE), Normalised Mean Bias Error (NMBE) and R-squared (R^2). Each metric provides insight into model accuracy, variance, and bias [80]. As shown in Equations (3)–(8), these error-based metrics form the foundation for assessing the goodness-of-fit and generalisability of the models across different scenarios. Following an initial screening of the eight ML models, five top-performing algorithms were selected based on these metrics for further analysis. In the context of measurement and verification (M&V), combining multiple metrics enhances interpretability and ensures a more nuanced understanding of predictive capability. Residual patterns and quantitative error scores provide a comprehensive basis for comparing model performance and identifying potential limitations.

$$R^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad (3)$$

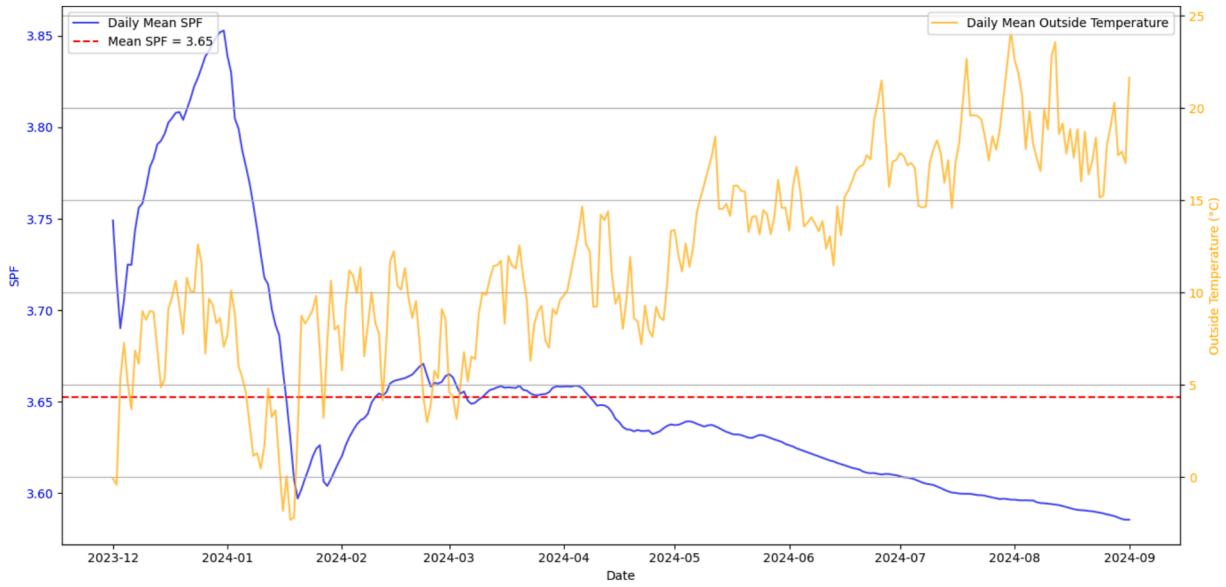


Fig. 11. Daily aggregated SPF variability and outside temperature.

Table 5
Descriptive statistics for A2WHP operational periods.

	Count	Mean	std	min	0.25	0.50	0.75	max
P_elec (kW)	3520.00	441.38	398.34	10.70	232.61	333.66	480.04	3000.02
Q_heat (kW)	3520.00	1630.74	1140.25	3.19	948.90	1456.95	1944.25	8013.73
T_flowT (°C)	3520.00	28.67	6.54	10.03	25.34	28.08	30.66	55.78
T_returnT (°C)	3520.00	26.53	5.29	12.97	24.63	25.85	27.24	50.73
FR_flowrate (m^3/hr)	3520.00	10.35	5.00	0.00	6.96	9.00	14.77	24.94
T_roomT (°C)	3520.00	20.28	1.50	6.43	19.97	20.66	21.04	25.10
T_outsideT (°C)	3520.00	8.73	4.78	−6.18	5.74	9.12	11.39	29.89
Hourly_COP	3520.00	3.86	1.04	0.10	3.33	4.11	4.58	5.99

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i| \tag{4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n}} \tag{6}$$

$$CVRMSE = \sqrt{\frac{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n}}{\overline{Y}}} \times 100\% \tag{7}$$

$$NMBE = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)}{\overline{Y}} X100\% \tag{8}$$

Where Y_i is the observed value of the heat demand in the building, $\widehat{Y}_i$ is the predicted value, $\overline{Y}$ is the mean of measured values, and n represents the number of data points.

2.6. Hyperparameter tuning

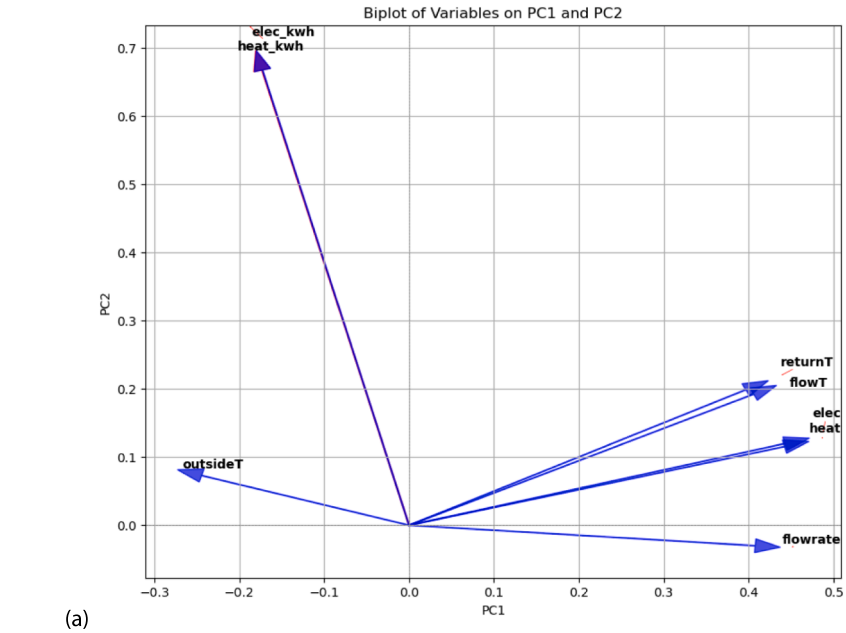
Optimising hyperparameter for ML algorithms involves balancing computational efficiency and accuracy. Various methods, including traditional and metaheuristic approaches, have been explored to achieve this balance. The choice of method can significantly impact the model's performance, as different techniques offer unique advantages in terms of convergence speed, flexibility, and robustness. There are four key characteristics of design parameter space, including dimensions, bounds, and size, and the goal is to identify the optimal combination of parameters for maximising the heat pump operational performance factor (Hourly COP). The following are the model-specific and sequential steps to perform a grid search for hyperparameter tuning. The grid search cross-validation method was implemented using GridSearchCV in the scikit-learn Pythonic library to perform an exhaustive search over the specified hyperparameter combinations for individual ML models [81]. Table 4 shows the optimal parameter search space, resulting in optimal hyperparameter values.

Input: Machine learning model f_{ML}, Training dataset .DGenerate: A set of candidate hyperparameters .$H = \{\theta_1, \cdots, \theta_n\}$Partition: Split D into k-fold cross-validation subsets .$\{ (D_{train,1}, D_{val,1,} \}, \cdots, \{ (D_{train,k}, D_{val,k,} \}$For each hyperparameter configuration θ_1 in H: For each fold $D_{train,j}, D_{val,j}$.: Train f_{ML} using $D_{train,j,}$ with .θ_i Compute error .$E_{i,j} = Error(f_{ML} trained on D_{train,j} validated on D_{val,j})$ Compute the average error for θ_i: .$E_i = \left(\frac{1}{k}\right)\Sigma(from j = 1 to k)E_{i,j}$Select optimal hyperparameters:.$\theta^* = argmin(over \theta_i in H)$Output: θ^*.f_{ML} trained with .θ^*

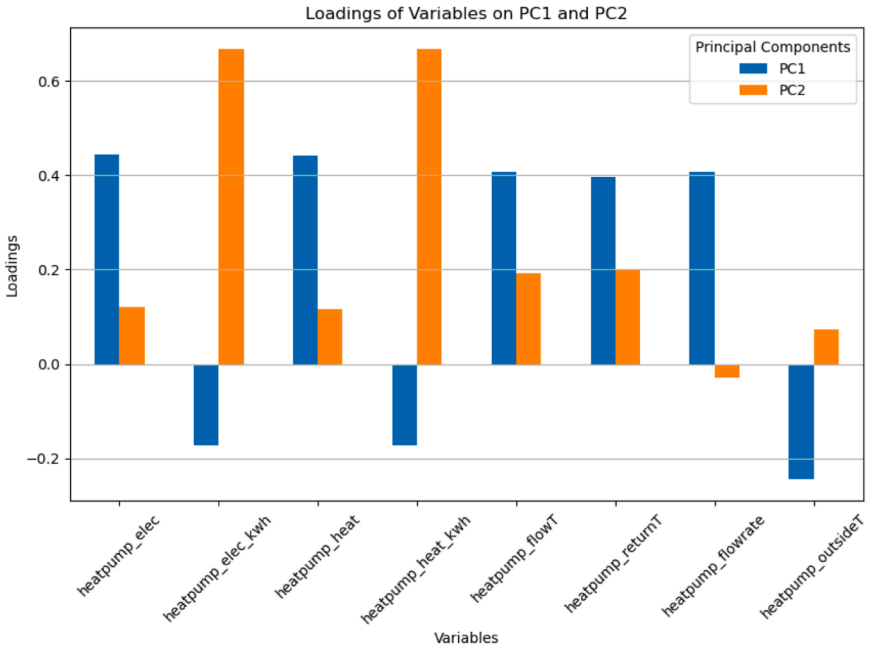
Input: Machine learning model \(f_{ML} \), Training dataset \ (D \).Generate: A set of candidates hyperparameters \ (H = \ {\theta_1, \dots, \theta_n\} \).Partition: Split \ (D \) into \ (k \)-fold cross-validation subsets \(\{(D_{train,1}, D_{val,1}), \dots, (D_{train, k}, D_{val, k}) \} \).for each hyperparameter configuration \ (\theta_i \in H \): for each fold \ ((D_{train, j}, D_{val, j}) \)

Train \(f_{ML} \) using \ (D_{train, j} \) with \ (\theta_i \).

Compute error \ (E_{i, j} = Error(f_{ML} \text {trained on} D_{train, j}, \text {validated on} D_{val, j}) \)



(a)



(b)

Fig. 12. Principal Component Analysis (a) Biplot (b) Variable loading plot.

Compute average error for \ (\theta_i \): \ (E_i = \frac{1}{k} \sum_{j = 1}^k E_{i, j} \)
Select optimal hyperparameters: \ (\theta^* = \text{argmin}_{\theta_i \in H} E_i \).
Output: \ (\theta^* \), \(f_{ML} \text {trained with} \theta^* \).

2.7. Season performance factor (SPF)

The SPF is the COP over time, often called the season coefficient of performance, and is defined in Equation (9). It is the main metric that heat pump manufacturers use while following regulatory specifications (e.g., MCS regulation for heat pump specifications). Fig. 11 shows the variation of daily aggregated SPF and outside temperature in this study.

$$SPF = \frac{Total\ Seasonal\ Heating\ Output}{Total\ Seasonal\ Electric\ Energy\ Input} = \frac{Q_{Hp} + Q_{Aux}}{E_{HP} + E_{aux} + E_{fan/wp} + E_{frost}} \quad (9)$$

Where Q_{Hp} is the heat generated by the heat pump, Q_{Aux} is the heat generated by the auxiliary heater, E_{HP} is the electrical energy consumed by the heat pump, E_{aux} is the electrical energy consumed by the auxiliary heater, $E_{fan/wp}$ is the electrical energy consumed by the fan or water pump, E_{frost} is the electrical energy consumed due to the defrosting operation

3. Results and analysis

The evaluation of various ML algorithms based on goodness of fit, error analysis, and hyperparameter optimisation has been

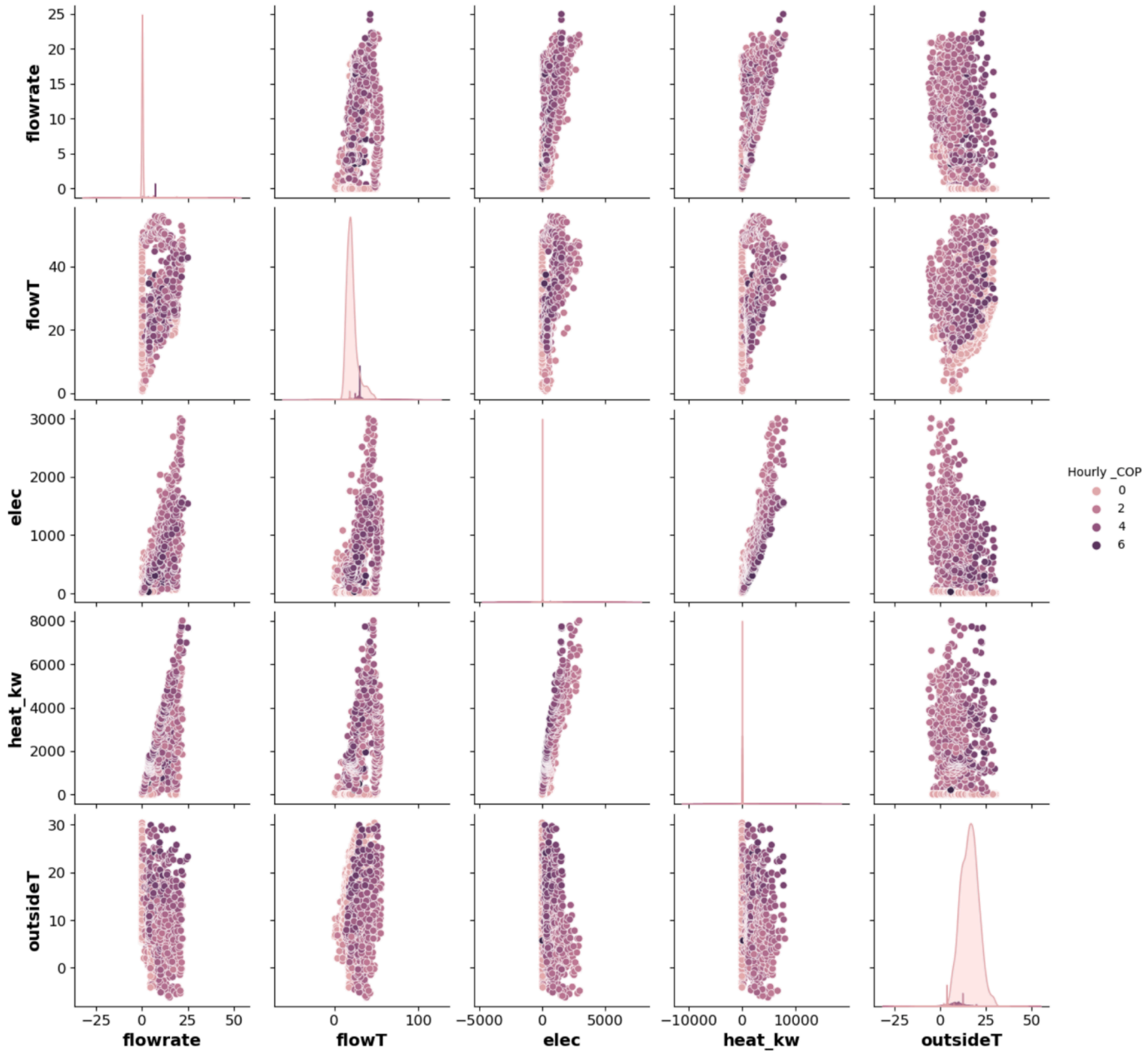


Fig. 13. Linear relationship assumed by the PCA.

systematically conducted using a real-world dataset from a monitored residential heat pump. The primary objective of this prediction is to compare the accuracy of ML algorithms in predicting heat pump efficiency in domestic settings by identifying the best-performing ML model, which can be utilised in a subsequent study and integrated into smart controls. This can be applied within adaptive control frameworks such as model predictive control (MPC) to enhance energy efficiency and thermal comfort in residential buildings.

3.1. Data descriptive statistics

Table 5 presents summary statistics that provide a quantitative snapshot of the characteristics of the dataset when the system was in active operation. It shows that the mean P-elec and Q_heat are 441.38 kW and 1630.74 kW, respectively. The average T_flowT for the case study when the system is operational is 28.67 °C. The maximum T_flowT is 55.78 °C, attributed to water heating and legionella prevention mode. The mean hourly T_outsideT is 8.73 °C when the A2WHP is in operation, whereas the minimum T_outsideT is −6.18 °C. In summary, the data demonstrate that the mean Hourly_COP during the operational period is 3.86, while the maximum Hourly_COP is 5.99 for the A2WHP active period. As part of the data diagnostic, Quick anomaly detection, such as a data acquisition (DAQ) error, idle periods, or performance drift, was spotted at a glance from the summary statistics before filtering the dataset of operational periods.

3.2. Principal component analysis (PCA)

RHPs often generate large volumes of high-dimensional, noisy data that require advanced processing and dimensionality reduction for efficient supervised learning [82]. Given the RHPs' complexity, incorporating unsupervised learning techniques such as PCA for dimension reduction is a justified approach to enhance predictive performance and reduce model complexity. PCA is a powerful dimensionality reduction

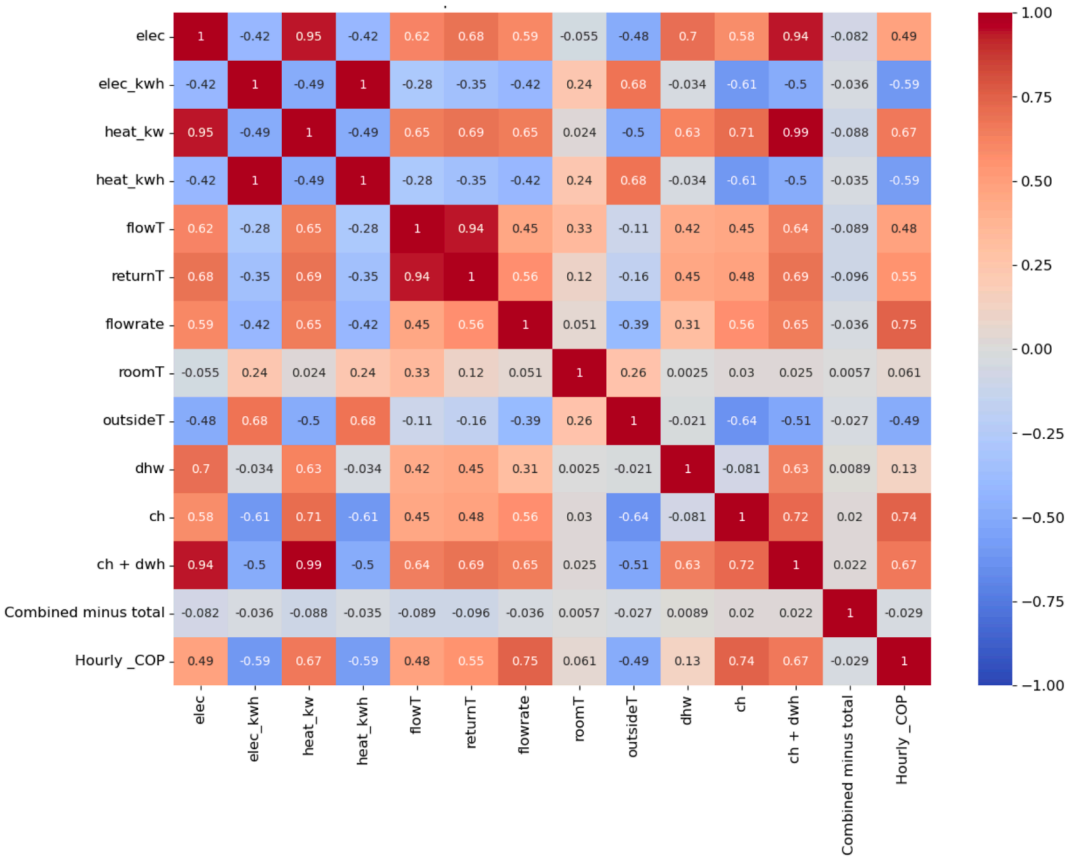


Fig. 14. Feature selection with a heatmap of correlation.

tool that eliminates redundancy by transforming the original feature space into a smaller set of principal components while preserving as much of the data's variance as possible [83]. This approach mitigates the effects of multicollinearity and overfitting, which are common issues in high-dimensional datasets, thereby contributing to the development of more robust and interpretable models. As shown in Fig. 13, this study used PCA and feature selection techniques to ensure alignment with the research objective and optimise model performance.

Fig. 12a presents the PCA biplot, which graphically shows both the scores (observations) and loadings (variable contributions) for the principal components, providing insights into the relationships between variables and observations [84]. For example, the arrow indicates that electrical power and heat energy contribute more significantly to PCA2, while the flow rate and outside temperature dominate PCA1. The angles between vectors in the biplot also reveal correlations among variables. For instance, flow and return temperatures are closely aligned, indicating a strong positive correlation, just like the electricity and heat output, which are given multivariate insights. Fig. 12b shows the variable loading plot, highlighting the magnitude and direction of each variable's contribution to the principal components. This plot helps identify the most influential features in explaining the data structure [85]. Although PCA is effective for capturing linear relationships, it may not adequately represent non-linear patterns within the data, as further described by the pair plot in Fig. 13, which shows complex and non-monotonic interactions among several variables.

3.3. Feature selection

Feature selection is critical in high-dimensional datasets, as it systematically identifies and retains the most informative variables while eliminating redundant or less relevant ones [69]. These techniques complement dimensionality reduction methods such as PCA, helping to improve model efficiency, reduce overfitting, and enhance interpretability. Feature selection methods are generally classified into supervised and unsupervised techniques. Under the supervised category, commonly used techniques include filter, wrapper, and embedded methods. In this study, four tools were applied, which include correlation-based filtering (i.e., correlation heatmap), univariate feature selection using mutual information regression (MIR), recursive feature elimination (RFE), and RF. The correlation analysis represents the linear relationships between variables using Pearson coefficients, where + 1 indicates a strong positive correlation, 0 is no correlation, and −1 is a strong negative correlation. It enables rapid identification of key feature interactions for model development. The heatmap uses colour gradients to help identify the strength and direction of these relationships across multiple variables. The MIR is a filter-based univariate method that excels at identifying linear and non-linear relationships between features and the target variable. It helps eliminate duplicated, redundant, and weakly informative features but does not address multicollinearity, which involves interdependencies among variables [86]. In contrast, RFE is a wrapper method that recursively eliminates less important features and provides an optimal subset for training. While RFE often yields higher predictive accuracy, it is computationally intensive [87,88]. The RF serves as an embedded feature selection method, combining the strengths of filter and wrapper approaches. It leverages multiple decision trees to rank features based on their importance, providing a balance of speed, accuracy, and the ability to handle feature interactions [89]. RF's built-in feature importance metric provides robust insights and improves model generalisability and resilience to overfitting.

Figs. 14-18 display the feature selection results obtained from applying four methods to the dataset, including correlation-based selection, MIR scores, features ranked via RFE, and importance rankings derived from RF. Fig. 14 shows the correlation analysis heatmap. The

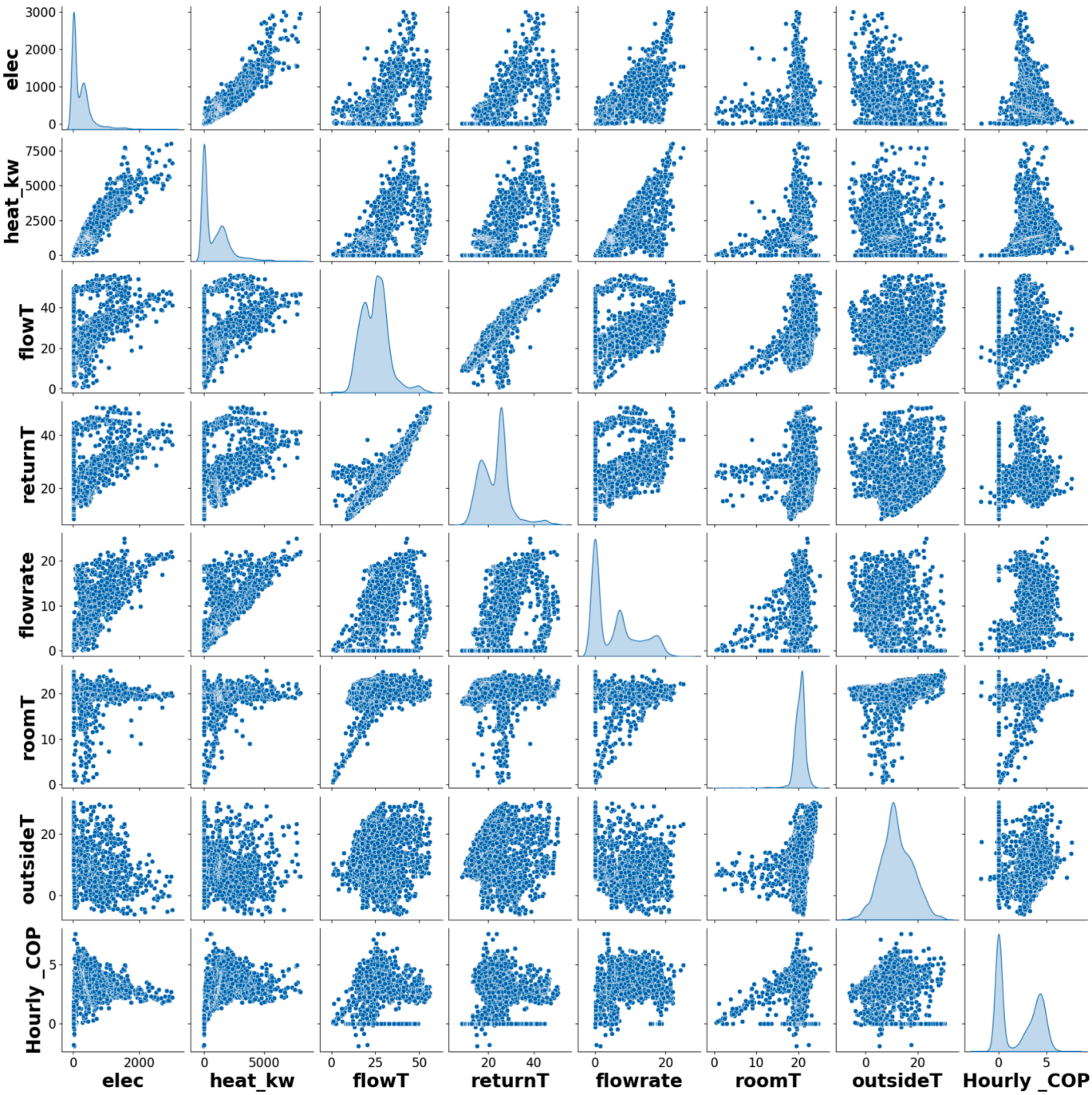


Fig. 15. Pair plot of independent variable and dependent variable (Hourly_COP).

analysis revealed several significant positive relationships among key variables in the dataset. Notably, a strong positive correlation exists between "heat_kwh" and "elec_kwh", suggesting that increases in heat output are closely aligned with rises in electrical energy consumption over time, reflecting a consistent relationship between heat generation and power usage. Similarly, "heat_kw" correlates strongly with both "elec" and "ch + dhw", indicating that higher heat demand, whether for central heating (CH) or domestic hot water (DHW), is met with increased electricity usage, underscoring the energy-intensive nature of combined heating loads. The correlation between "volumetric flow rate" and "Hourly_COP" suggests that higher flow rates are associated with better heat pump performance, reflecting their proportional influence on system efficiency. Additionally, the relationship between "dhw" and "elec" highlights the dependency of DHW on electricity consumption. As seen in the heatmap, these correlations uncover essential patterns and interdependencies between heat output and electricity-related variables, offering valuable insights for further modelling and system optimisation. The pair plot in Fig. 15 complements this analysis by showing relationships among variables through a matrix of scatter plots. Each diagonal element presents a density plot of variable distributions, while off-diagonal plots show pairwise relationships. The plot reveals that only the outside temperature follows a normal distribution and maintains a monotonic relationship with the target variable (Hourly_COP). In contrast, other variables exhibit non-linear and multivariate interactions, implying that multiple factors influence residential heat pumps (RHPs) in their performance. Also, the diagonal plots provide insight into the central tendency and dispersion of each variable, which reinforces the high dimensionality and non-linearity of the dataset and

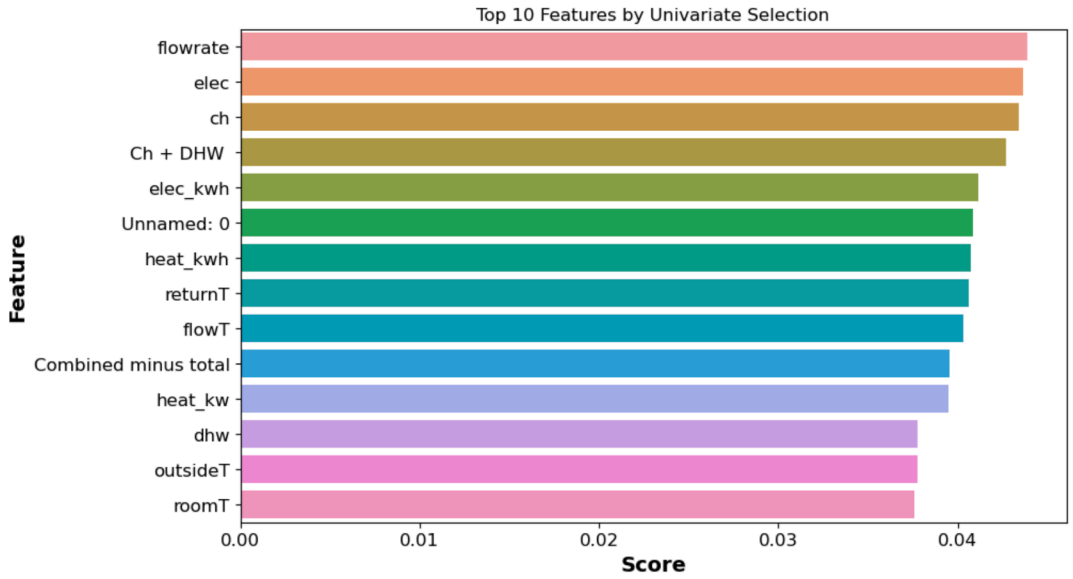


Fig. 16. Feature selection with the univariate method.

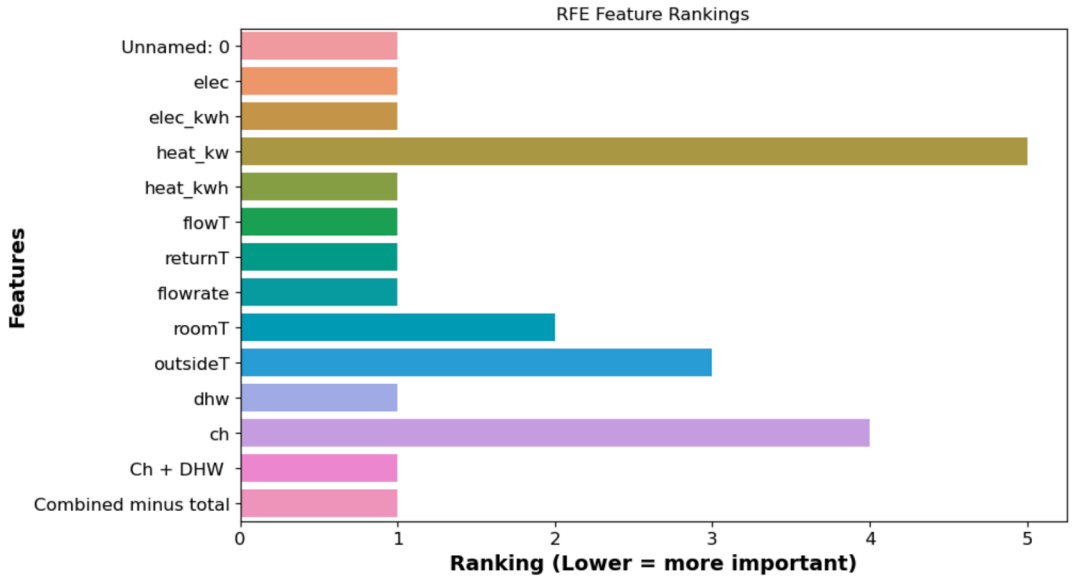


Fig. 17. Feature selection with the recursive feature elimination (RFE) method.

highlights the need for advanced modelling techniques to capture these dynamics effectively.

Fig. 16 shows insights that affirm that MIR effectively captures complex dependencies that are not evident through linear correlation alone [16,90]. The univariate ranking method positions all variables as equally important, implying that it is less sensitive to the interactions between variables. Although it provides insights, it does not fully explore the multi-dimensional interaction of the features in the dataset. Fig. 17 shows the ranking of input variables based on their predictive relevance, with flow rate, electric power consumption, and combined heating and hot water demand (CH + DHW) identified as the most influential features. Variables such as return temperature and timestamp (date) were found to contribute minimally and were excluded. Despite the low linear correlations observed in the heatmap (cf. Fig. 14), high mutual information scores for flow rate, electricity consumption (Elec), and combined heating demand (CH + DHW) in Fig. 18 reveal underlying non-linear relationships with the target variable, Hourly COP. These three features consistently emerge as the most influential across all selection methods, making them prime candidates for inclusion in the model. The RF method also highlighted flow and return temperatures as important predictors, reflecting non-linear interactions not captured by linear filters or univariate methods. While MIR did not prioritise the outside temperature, its selection by supervised methods, such as RFE and RF, suggests a monotonic relationship with Hourly_COP. Overall, RF demonstrated superior feature selection performance among the four methods, likely due to its integrated selection mechanism [69,91,92].

3.4. *Regression-Based model fitting*

To identify the most suitable ML model for Residential heat pumps (RHPs), eight ML algorithms were initially trained on the dataset to mitigate bias and improve the robustness of performance predictions. Hyperparameter tuning using grid search and K-fold cross-validation was performed to optimise model performance. This combined data- and algorithm-focused strategy ensured fair and reliable model comparison. Table 6 shows the baseline model performance before

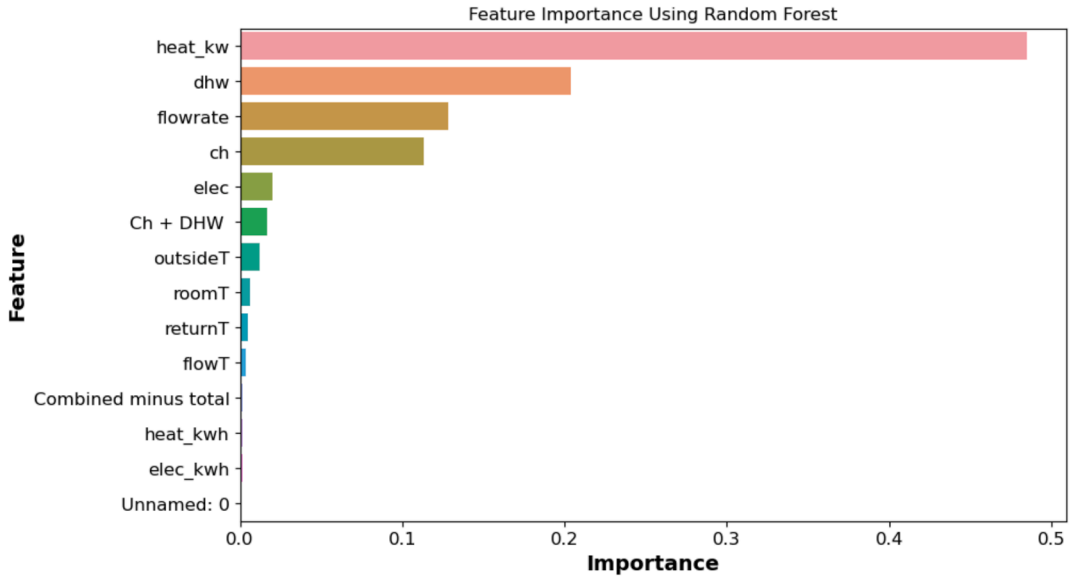


Fig. 18. Feature selection with random forest, including feature importance ranking (RF) method.

Table 6
Baseline model performance before optimisation.

ML Algorithm	R^2	MAE	MSE	RMSE	NMBE (%)	CVRMSE (%)
LR	0.681	0.759	0.774	0.880	6.191	38.049
RF	0.917	0.276	0.202	0.449	3.845	19.435
XGBoost	0.920	0.296	0.194	0.441	−0.867	19.070
PR	0.798	0.554	0.491	0.701	7.105	30.296
MLR	0.681	0.759	0.774	0.880	6.191	38.050
ANN	0.875	0.401	0.304	0.551	7.266	23.840
LSTM	0.900	0.358	0.243	0.493	6.640	21.323
SVR	0.806	0.464	0.470	0.685	12.505	29.643

Table 7
Refined model performance after optimisation.

ML Algorithm	R^2	MAE	MSE	RMSE	NMBE (%)	CVRMSE (%)
LSTM	0.998	0.046	0.008	0.088	1.460	4.170
XGBoost	0.996	0.063	0.017	0.131	−0.022	1.287
RF	0.994	0.069	0.025	0.159	0.025	7.544
ANN	0.997	0.071	0.011	0.107	−2.652	5.073
SVR	0.985	0.113	0.067	0.259	0.303	12.250
Ensemble	0.992	0.097	0.035	0.188	0.443	8.920

optimisation across the evaluation metrics. The baseline performance of all the models across the key performance indicators (KPIs) is generally low and likely due to their inability to capture the non-linear relationships within the dataset, as illustrated in the 3D surface plot (cf. Fig. 7b). Additionally, this highlights potential issues such as overfitting or model sensitivity to data quality [93,94]. However, among the evaluated models, the tree-based algorithms (RF and XGBoost) relatively outperformed others within the acceptable measurement device uncertainty bound. They achieved the highest R^2 (0.917 and 0.920) and lowest RMSE (0.449 and 0.441). Neural-based models (ANN, LSTM) offer comparable performance but with marginally higher error metrics. The kernel-based algorithm (SVR) also exhibits competitive performance metrics, with the error metrics slightly above measurement device resolution. The baseline model performance derived via default tuning was subsequently used to establish initial performance benchmarks, which were later optimised using the refinement pipeline. To develop ML models that are statistically coherent with the acceptable measurement device's threshold. For example, the ideal NMBE should fall within the measurement uncertainty range of ± 2 and ± 10 %, while ASHRAE guidelines 14 recommend a CVRMSE of less than 30 % for energy, temperature, and flow measurements [95]. It is then essential to implement a rigorous optimisation pipeline involving k-fold cross-validation, residual diagnostics, feature scaling, and grid-search hyperparameter tuning. Revilla-León et al. [96] emphasised the importance of comprehensive evaluation criteria and model selection strategies in developing a robust ML model for a complex system like a heat pump. Table 7 presents the summary of refined model performance after optimisation. Following the model tuning, LSTM emerged as the best-performing predictor with an R^2 of 0.998 and an RMSE of 0.088, outperforming other model architectures across all error metrics. XGBoost demonstrated significant generalisation with minimal bias (NMBE − 0.022 %) and lowest CVRMSE (1.287 %). All models demonstrate substantial improvements, such as NMBE values falling within ± 2 %, which satisfies both the measurement device uncertainty bound and the ASHRAE guideline 14 conservative threshold for hourly data across all models. However, the ANN algorithm yielded NMBE of − 2.652 %, which slightly above the tighter ± 2 % measurement device threshold, it satisfy the broader ± 10 % accuracy bounds recommended by ASHRAE guidelines 14 [95]. Meanwhile, SVR with a CVRMSE of 12.250 %, which marginally exceeds the ASHRAE guideline 14 conservative threshold (±10 %), remains well within the acceptable standard of ≤ 30 % for hourly resolution datasets. LSTM, ANN, RF, XGBoost, and ensemble algorithms achieve R^2 values above 0.99, aligning with DOE/FEMP best practice for hourly predictive modelling in building performance evaluation [97]. Overall, these metrics confirm that the optimised models are statistically coherent with the measurement device's resolution, aligning with ASHRAE guidelines 14 and 140, and adhere to DOE/FEMP best practice for predictive modelling of hourly resolution data in building applications, making them deployment-ready. Details of the measurement device's uncertainty bounds are presented in Table 1.

Fig. 19 compares the predicted and actual outcomes for each model. The linear regression (baseline) model, based on the aggregated daily_cop dataset, shows a significant deviation from the observed values. XGBoost exhibits near-perfect alignment with field data, followed by RF, while SVR, LR, and MLR significantly underperform. Fig. 20 illustrates the five shortlisted models selected for further evaluation based on their initial performance, as assessed using a mix of evaluation metrics under the default parameter configuration.

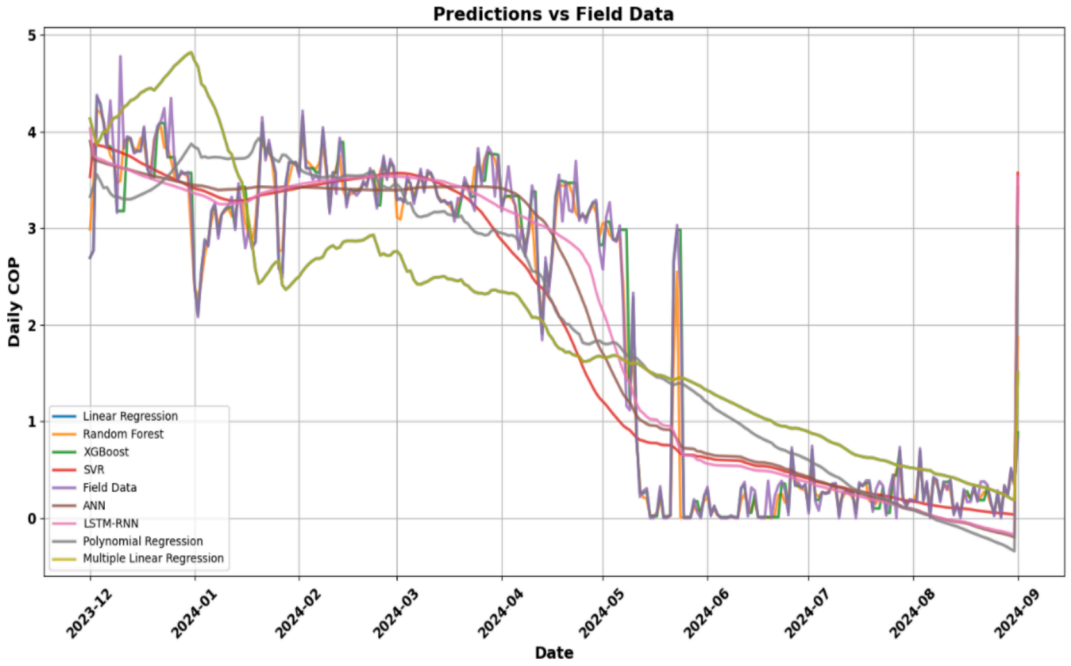


Fig. 19. Comparison of ML algorithm predictions to field data.

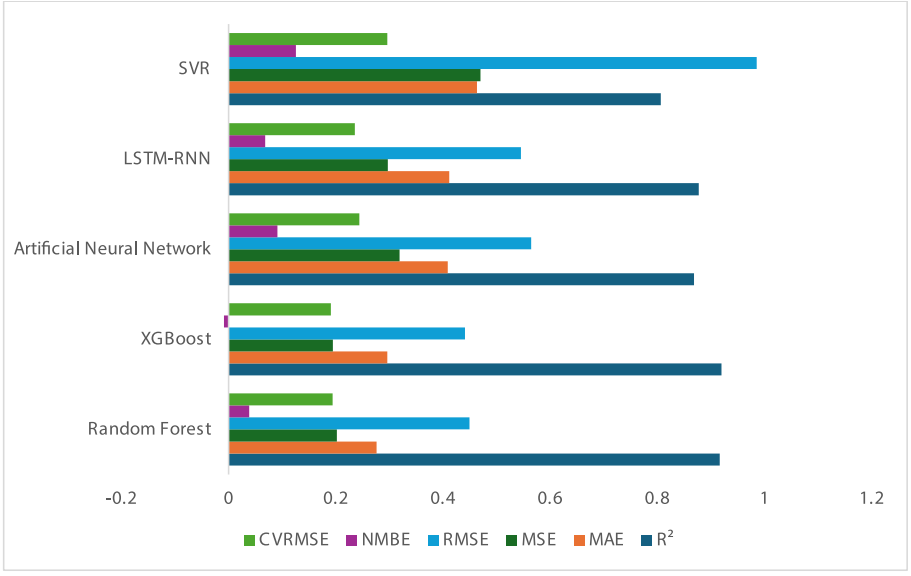


Fig. 20. Proportion of model performance across six evaluation metrics before hyperparameter optimisation.

3.5. Hourly COP prediction before hyperparameter tuning

In the predictions from different algorithms, the ideal fit line is a visual benchmark to compare how closely the model's predictions match the actual output. The reference line signifies optimal model performance and visually interprets the model prediction. As shown in Fig. 21 (a – e), the strengths and weaknesses of each algorithm can be understood from the data distribution around the ideal line under manual hyperparameter settings, despite the default internal optimisation attributes of certain ML models, such as XGBoost and LSTM, which are typically defined in popular frameworks like TensorFlow Keras. However, these default settings do not usually provide optimal performance, requiring further tuning approaches [98]. XGBoost, RF, and ANN predictions reasonably align with the ideal fit line under the standard hyperparameters. The near-accurate performance of these algorithms necessitates further hyperparameter optimisation to enhance predictive performance. The conceptual model of heat pump performance that has been computationally implemented requires verification and validation [99]. This assessment is partly supported by expressing the model in a plotted form after it has been automatically implemented on a computer. The five subplots in Fig. 21 represent the ML results of the target feature (Hourly_COP) on the training set for SVR, XGBoost, RF, ANN, and LSTM before the hyperparameter training. XGBoost, ANN and RF model predictions show moderate alignment with the field data as opposed to SVR and LSTM.

3.6. Residual analysis

The residual plots are employed to evaluate the presence of systematic error patterns, with ideally unbiased predictions appearing

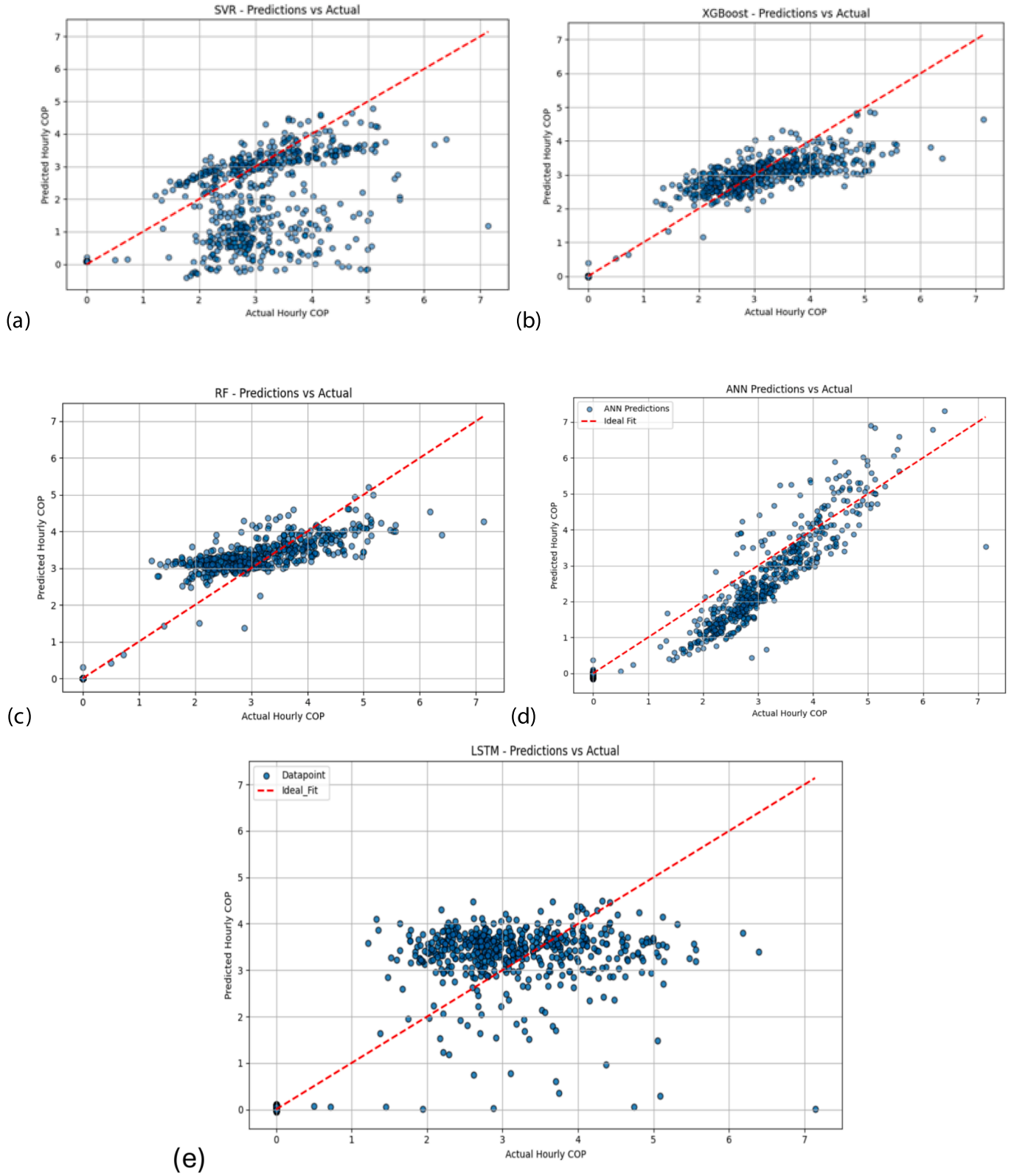


Fig. 21. ML model prediction before hyperparameter tuning vs. actual observations (a) SVR (b) XGBoost (c) RF (d) ANN (e) LSTM.

randomly distributed around zero. As defined in Equation (10), the residual should demonstrate no visible trends or patterns and instead be symmetrically distributed about the zero axis. The presence of systematic patterns indicates model deficiencies that require improvements, such as the inclusion of higher-order terms, data transformation, or hyperparameter tuning, to fit the underlying relationship better. To assess how effectively models capture underlying data patterns and to affirm the absence of bias, residual analysis, as illustrated in Fig. 22, compares model performance under manually configured and optimised hyperparameters.

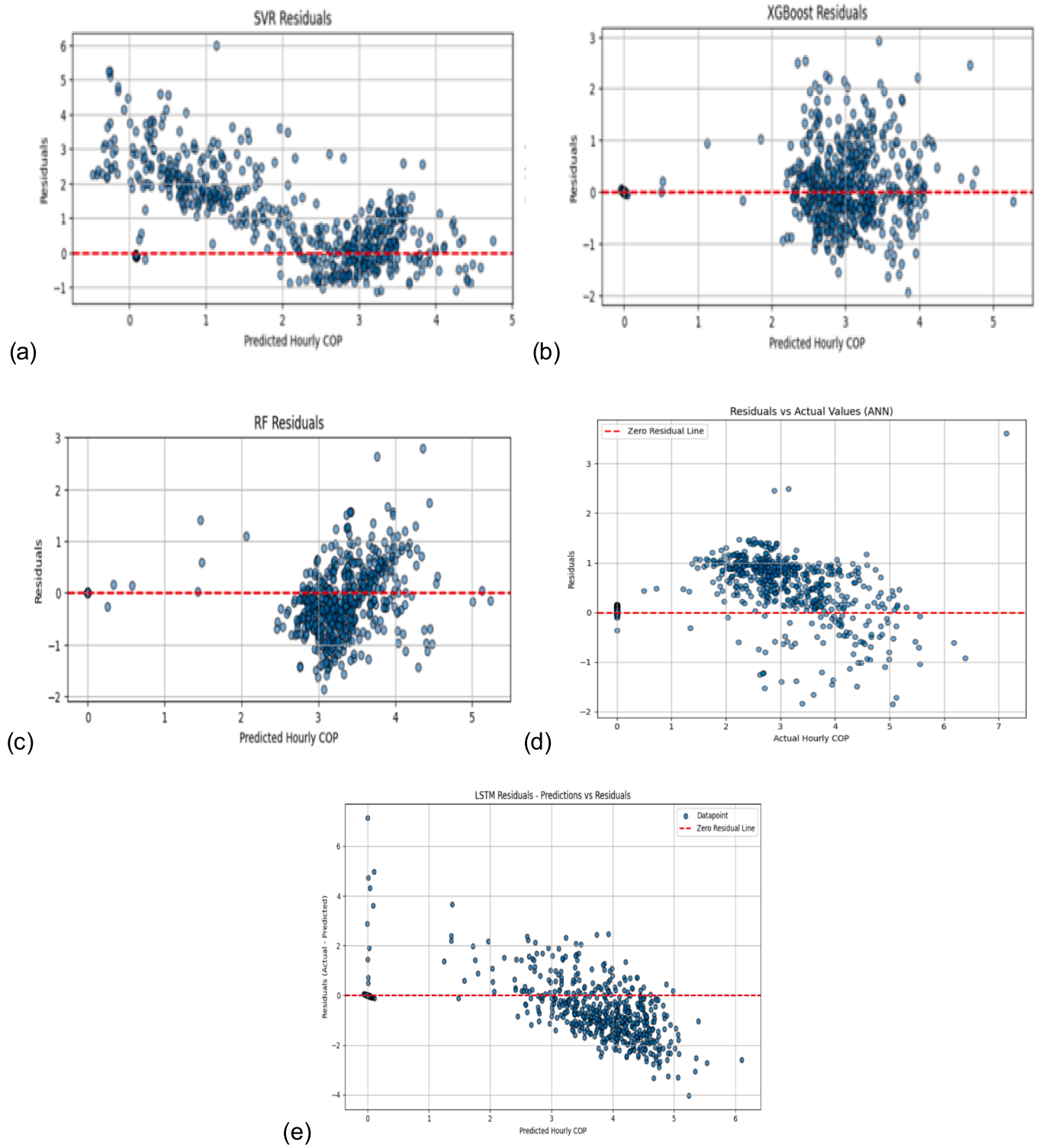


Fig. 22. Residual before the model optimisation (a) SVR (b) XGBoost (c) RF (d) ANN (e) LSTM.

$$Residuals(e_i) = Actual(Y_i) - Predicted(\widehat{Y}_i) \quad (10)$$

3.7. Model performance under hyperparameter optimisation

Fig. 23 shows the performance of each model after the parameter tuning on the test dataset. The plot describes the percentage distributions of evaluation metrics for the models. LSTM performs better over other models across all the metrics, followed by ANN. This is likely due to the ability of the LSTM model to handle long-term dependencies in sequential data effectively. Additionally, nuances like underprediction that are not as apparent in the conventional error metrics, such as MAE, MSE, and RMSE, are revealed by NMBE and CVRMSE. The LSTM and ANN models exhibiting negative NMBEs are likely due to the algorithms predicting lower values than the actual target values, as observed by O'Neill and Costello [100]. It is worth noting that R^2 is referred to in this study as a relative model comparison tool rather than treated as an

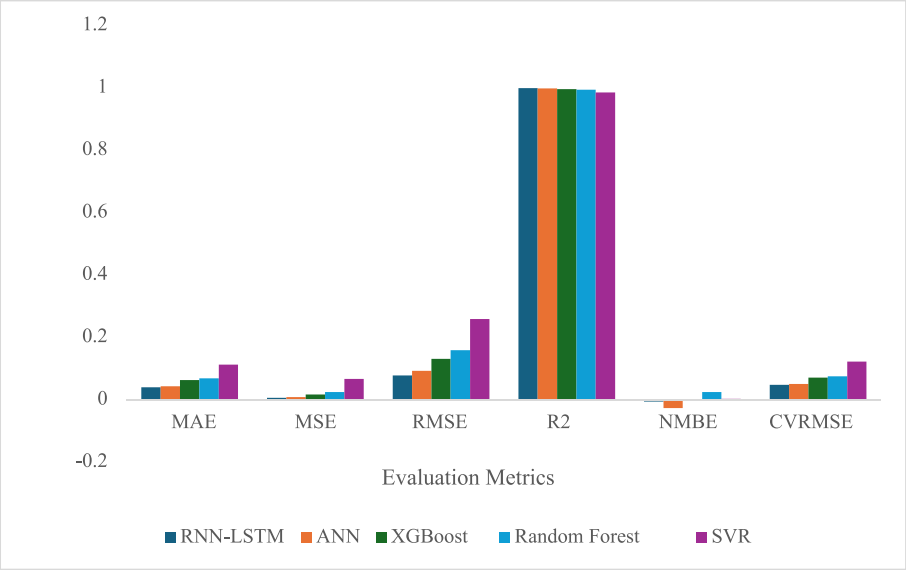


Fig. 23. Analysis of error metrics and coefficient of determination for different algorithms featuring NMBE and CVRMSE.

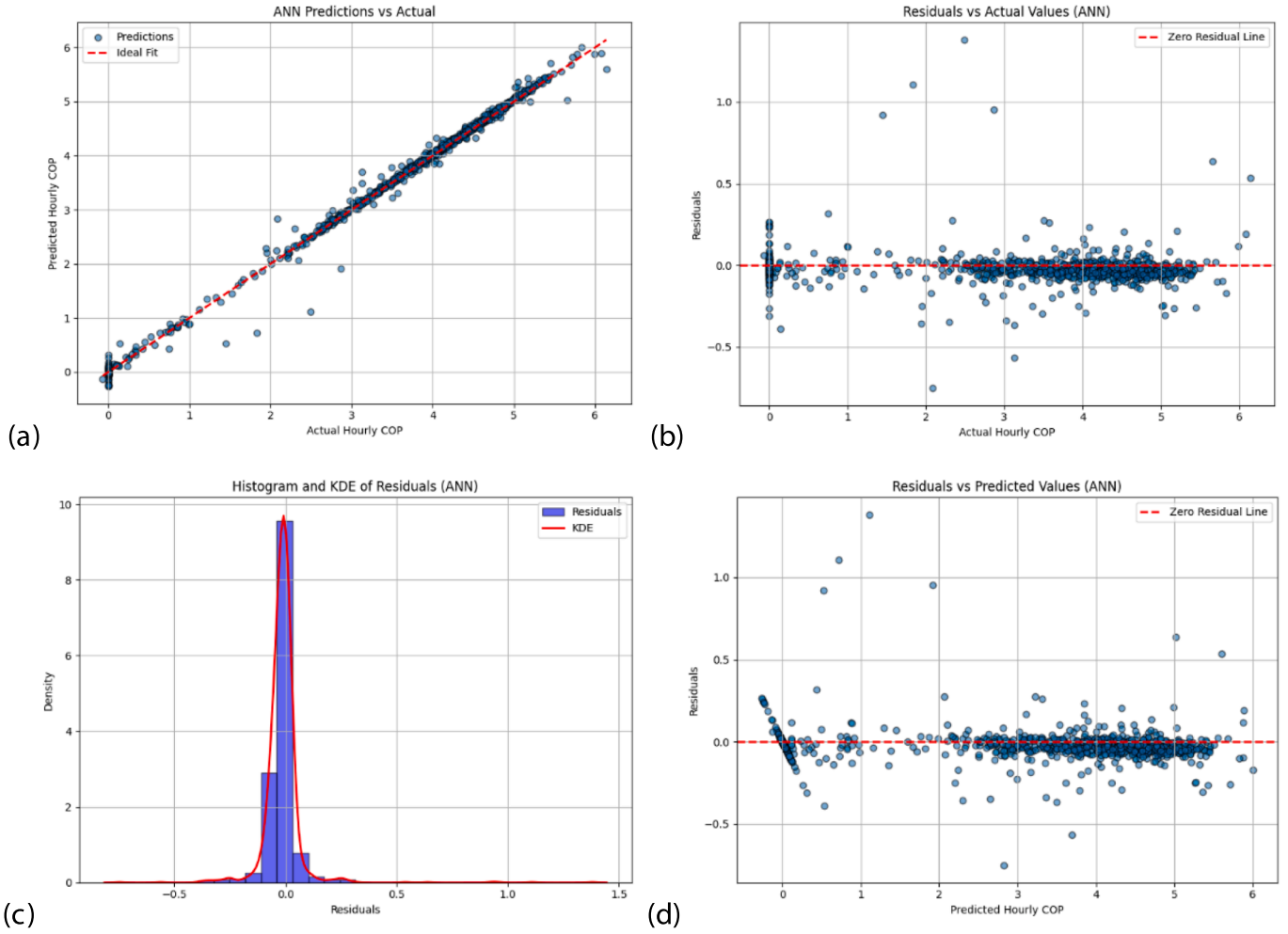


Fig. 24. ANN-based model performance after hyperparameter tuning. (a) Predictions vs Actual (b) Residuals vs actual values (c) Residual Histogram-KDE plot (d) Residual vs predicted values. Best Hyperparameters: {'activation': 'tanh', 'hidden_layers': 2, 'optimiser': 'adam', 'units': 32} and Test Set Evaluation Metrics – MAE = 0.071, MSE = 0.011, RMSE = 0.107, R^2 = 0.998, NMBE = -2.652%, and CVRMSE) = 5.073%.

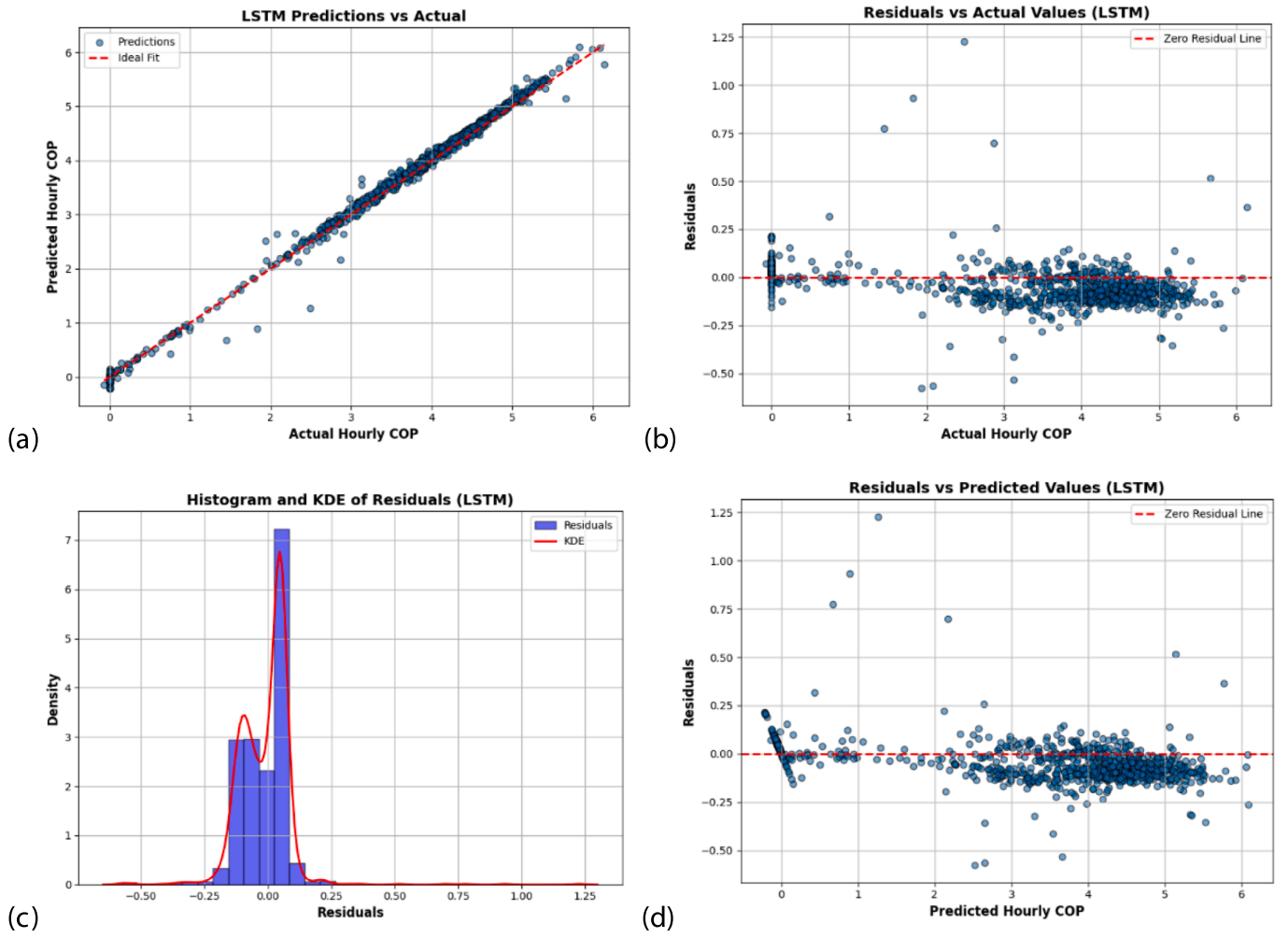


Fig. 25. LSTM-based model performance following hyperparameter tuning. (a) Predictions vs Actual (b) Residuals vs actual values (c) Residual Histogram-KDE plot (d) Residual vs predicted values. `Best Hyperparameters: {'activation': 'tanh', 'hidden_layers': 1, 'lstm_units': 150, 'optimiser': 'adam'}` and `Test Set Evaluation Metrics: MAE = 0.046, MSE = 0.008, RMSE = 0.088, $R^2$ = 0.998, NMBE = 1.46%` and `CVRMSE) = 4.170%.`

absolute performance measure because of the non-linear characteristics of the datasets.

3.8. Model diagnostics and residual analysis following hyperparameter optimisation

Figs. 24 to 28 show model diagnostic plots, which comprise frequently used plots in regression. These include predicted vs. actual, residual vs. actual values, residual vs. predicted values, and error distribution (Histogram/KDE) for each figure. These plots reflect the models' prediction performance and residual patterns, particularly on the unseen data, showing bias, variance, and the capacity of the models to generalise. In all cases, the coefficient of determination (R^2) is interpreted as a relative comparison tool among the models, but not as a measure of goodness of fit. The best hyperparameters influencing the performance of each model consist of discrete and categorical values. For instance, the best hyperparameters for the ANN model are: {'activation': 'tanh', 'hidden_layers': 2, 'optimiser': 'adam', 'units': 32} while the resulting error metrics are evaluated as – MAE = 0.043, MSE = 0.009, RMSE = 0.093, R2 = 0.998, NMBE = -2.652%, and CVRMSE) = 5.073%. The best Hyperparameters and the test set evaluation metric values are captioned in the figures corresponding to each model. ANN and LSTM prediction models exhibit high accuracy (R2 = 0.998 and 0.999), as reflected in their prediction vs. actual plots in Figs. 24 and 25, respectively. The two residual plots in Fig. 25 indicate that the errors cluster into two distinct regions of the input space, accounting for the bimodal distribution observed in the histogram-KDE of the LSTM presented in the plot. The three ensemble algorithms, RF (Fig. 26), XGBoost (Fig. 27), and SVR (Fig. 28), perform reasonably well but require further tuning to match the performance of the deep-learning algorithms (ANN and LSTM).

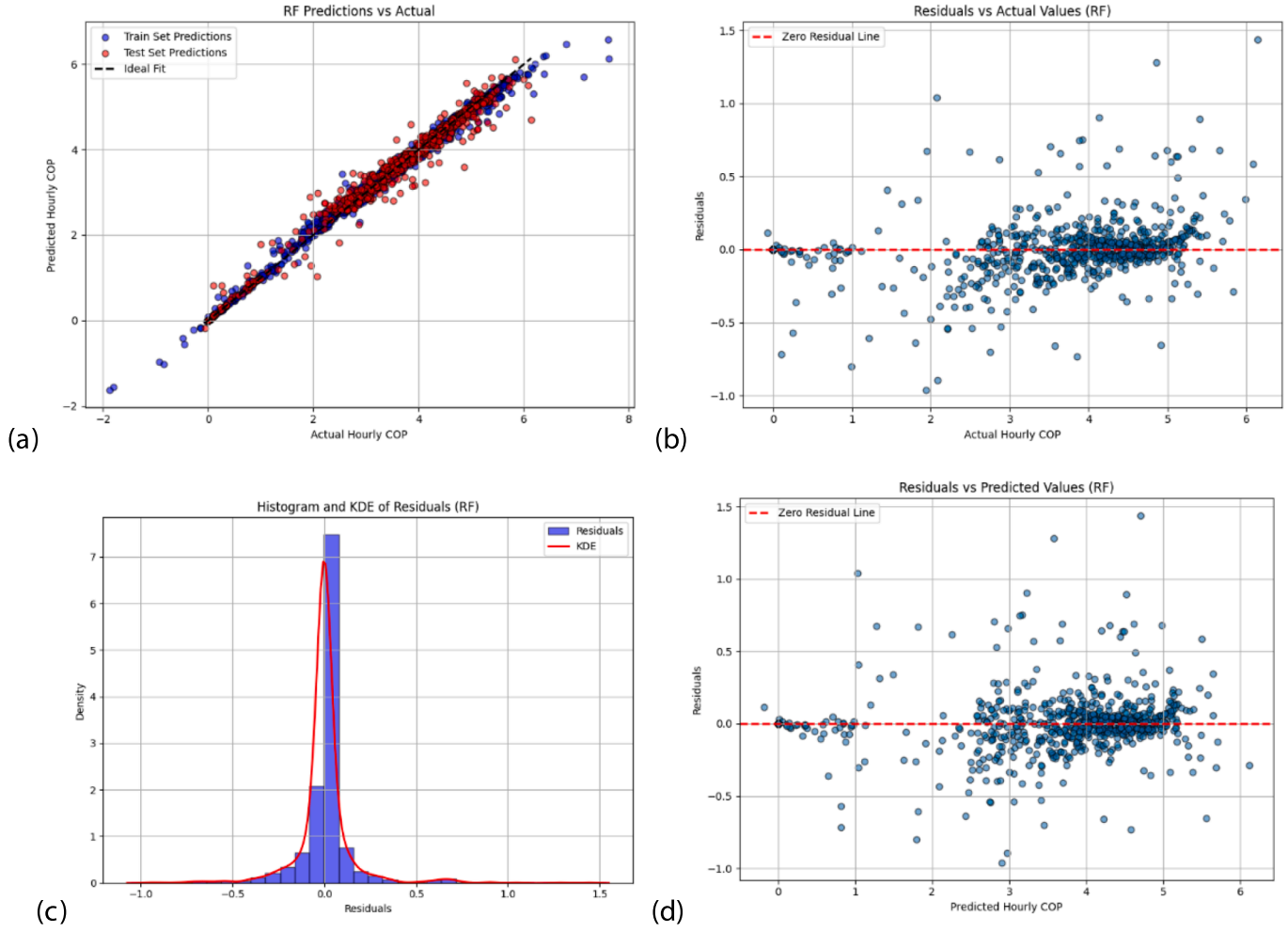


Fig. 26. RF-based model performance after hyperparameter tuning. (a) Predictions vs Actual (b) Residuals vs actual values (c) Residual Histogram-KDE plot (d) Residual vs predicted values. Best Hyperparameters: {'max_depth': 10, 'n_estimators': 100} and Test Set Evaluation Metrics - MAE = 0.069, MSE = 0.025, RMSE = 0.159, R^2 = 0.994, NMBE = 0.025% and CVRMSE = 7.544%.

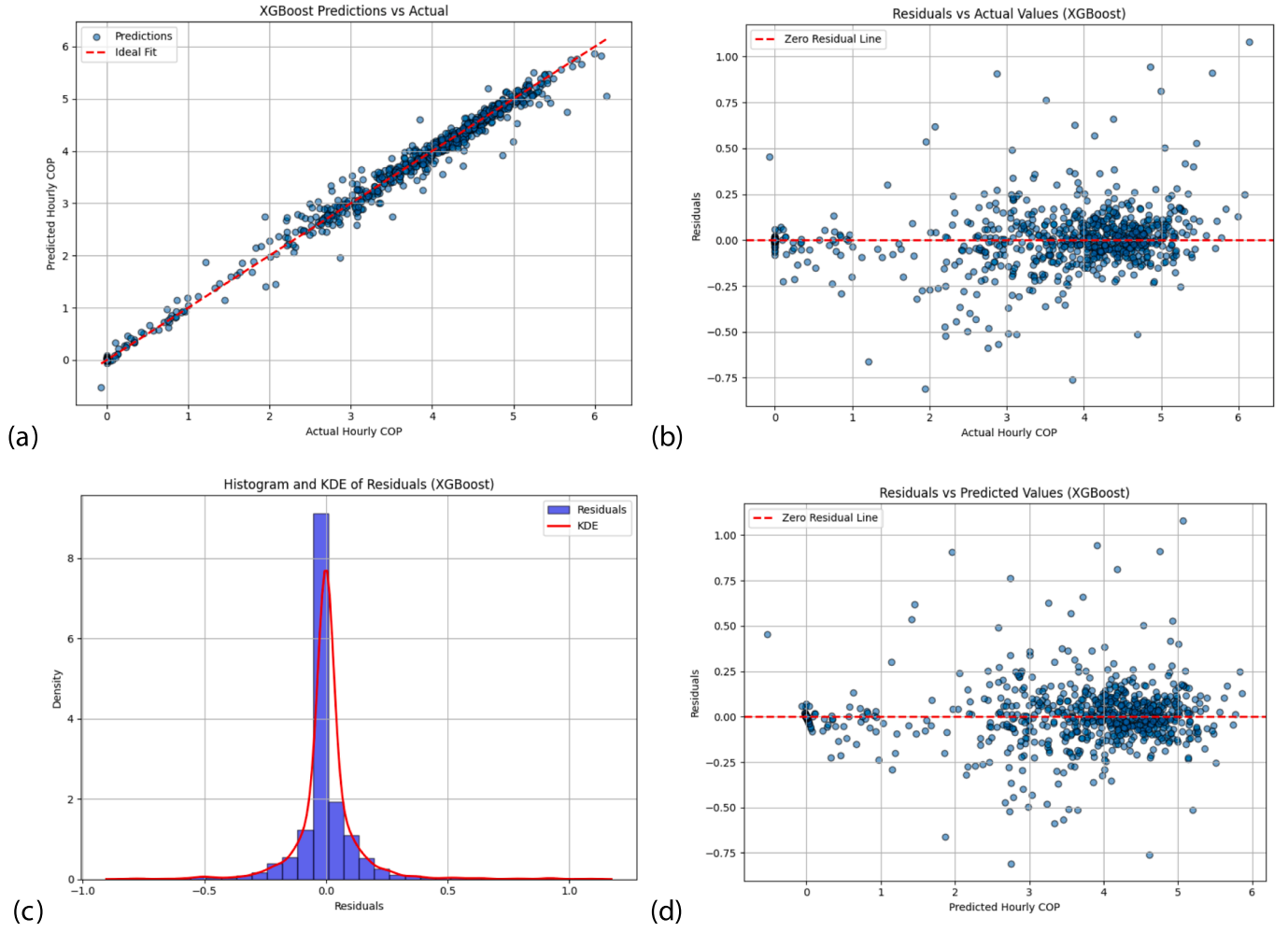


Fig. 27. XGBoost-based model performance after hyperparameter tuning. (a) Predictions vs Actual (b) Residuals vs actual values (c) Residual Histogram-KDE plot (d) Residual vs predicted values. `Best Hyperparameters: {'max_depth': 10, 'n_estimators': 100}` and `Test Set Evaluation Metrics` – $MAE = 0.063$, $MSE = 0.017$, $RMSE = 0.131$, $R^2 = 0.996$, $NMBE = -0.022\%$, and $CVRMSE = 1.287\%$.

3.9. *Effect of hyperparameter tuning on model performance*

Fig. 29 (a to f) shows the relative improvement in model performance resulting from hyperparameter optimisation within their architecture, as evidenced by the percentage increase in R^2 and the reduction in error metric values. The percentage improvement from optimising the model's hyperparameters using GridsearchCV and k-fold validation was estimated. Comparing R^2 and error metrics obtained before and after ML parameter optimisation allows an explicit comparison between the models and reveals the best performers. Percentage improvements in model performance are as follows: ANN recorded 22 % improvement in R^2, 22 % reduction in MAE, 21 % reduction in MSE, 22 % reduction in RMSE, 23 % reduction in CVRMSE, and 24 % reduction in NMBE; LSTM recorded 21 % improvement in R^2, 22 % reduction in MAE, 21 % reduction in MSE, 23 % reduction in RMSE, 19 % reduction in CVRMSE, and 20 % reduction in NMBE; XGBoost recorded 13 % improvement in R^2, 19 % reduction in MAE, 20 % reduction in MSE, 19 % reduction in RMSE, 19 % reduction in CVRMSE, and 20 % reduction in NMBE; RF recorded 13 % improvement in R^2, 18 % reduction in MAE, 19 % reduction in MSE, 17 % reduction in RMSE, 18 % reduction in CVRMSE, and 18 % reduction in NMBE; SVR recorded 31 % improvement in R^2, 19 % reduction in MAE, 19 % reduction in MSE, 19 % reduction in RMSE, 17 % reduction in CVRMSE, and 18 % reduction in NMBE. ANN and LSTM consistently demonstrate improved error reduction and a higher coefficient of determination (R^2) across all evaluation metrics, establishing them as the best ML models for COP prediction based on the primary method adopted in this study. The XGBoost, RF, and SVR perform relatively well, and their performance can be further improved to match ANN and LSTM.

While prior studies have explored predictive modelling for heat pump systems using data-driven and application of ML methods, such as [23] and [101], this study advances the field by integrating real-time experimental data and uncertainty-aware KPIs to assess ML models. Unlike previous approaches that relied on existing datasets, employed single baseline models, and empirical tuning, our framework demonstrates statistically coherent performance across regimes and improved interpretability.

Table 7 summarises optimised key performance indicator (KPI) results obtained through the systematic optimisation pipeline, benchmarked against the existing standards. All evaluated MLs' performance not only aligns with the uncertainty limit of IoT equipment but also conforms to ASHRAE guideline 14. These metrics underscore the novelty and practical relevance of our method, particularly in the context of real-time system operation and control.

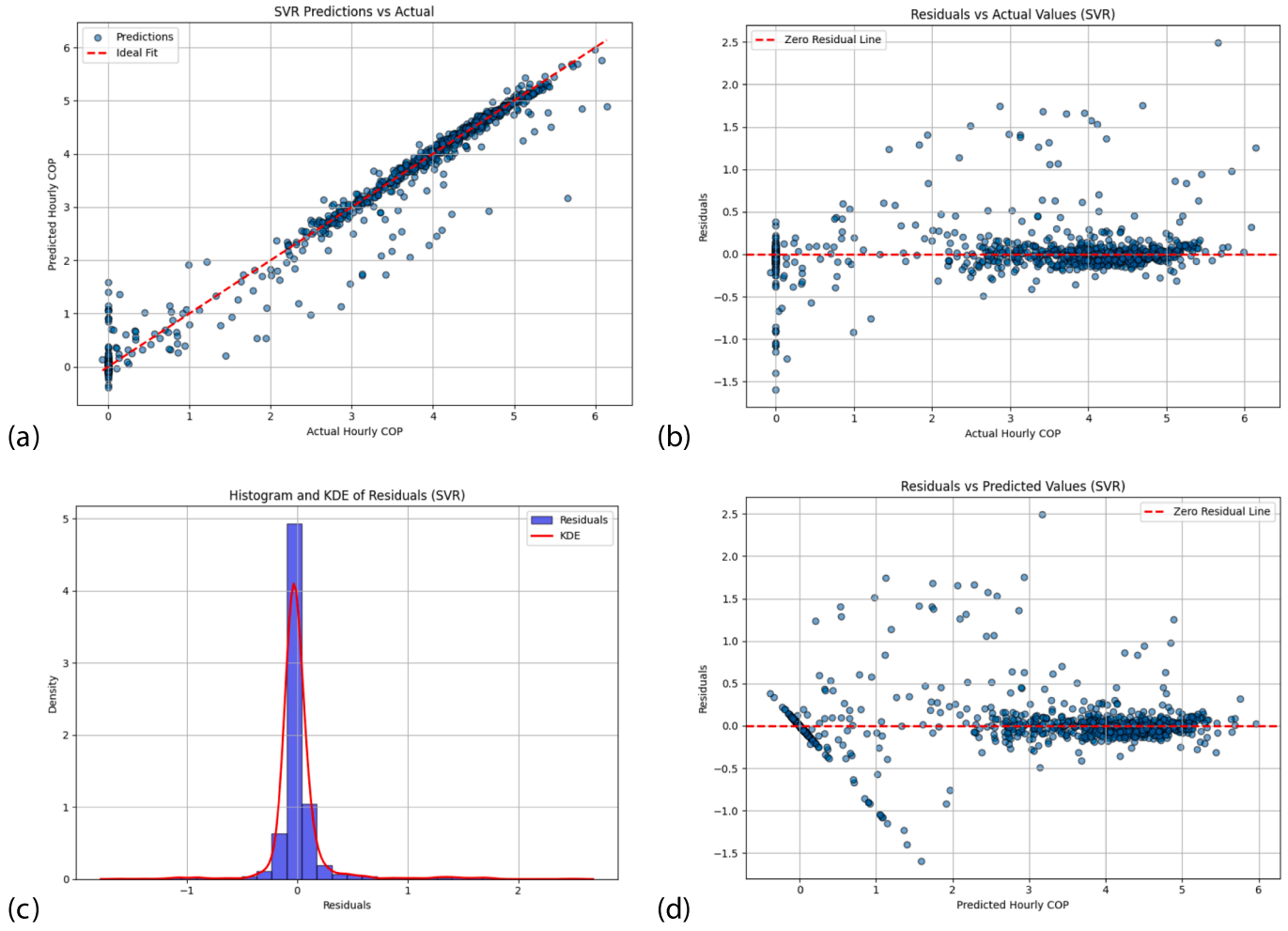


Fig. 28. SVR-based model performance after hyperparameter tuning. (a) Predictions vs Actual (b) Residuals vs actual values (c) Residual Histogram-KDE plot (d) Residual vs predicted values. Fitting five folds for each of the 48 candidates, totalling 240 fits. `Best Hyperparameters: {'C': 100, 'gamma': 'scale', 'kernel': 'rbf'}` and `Test Set Evaluation Metrics: MAE = 0.113, MSE = 0.067, RMSE = 0.259,` R^2 `= 0.985, NMBE = 0.303%,` and `CVRMSE = 12.250%`.

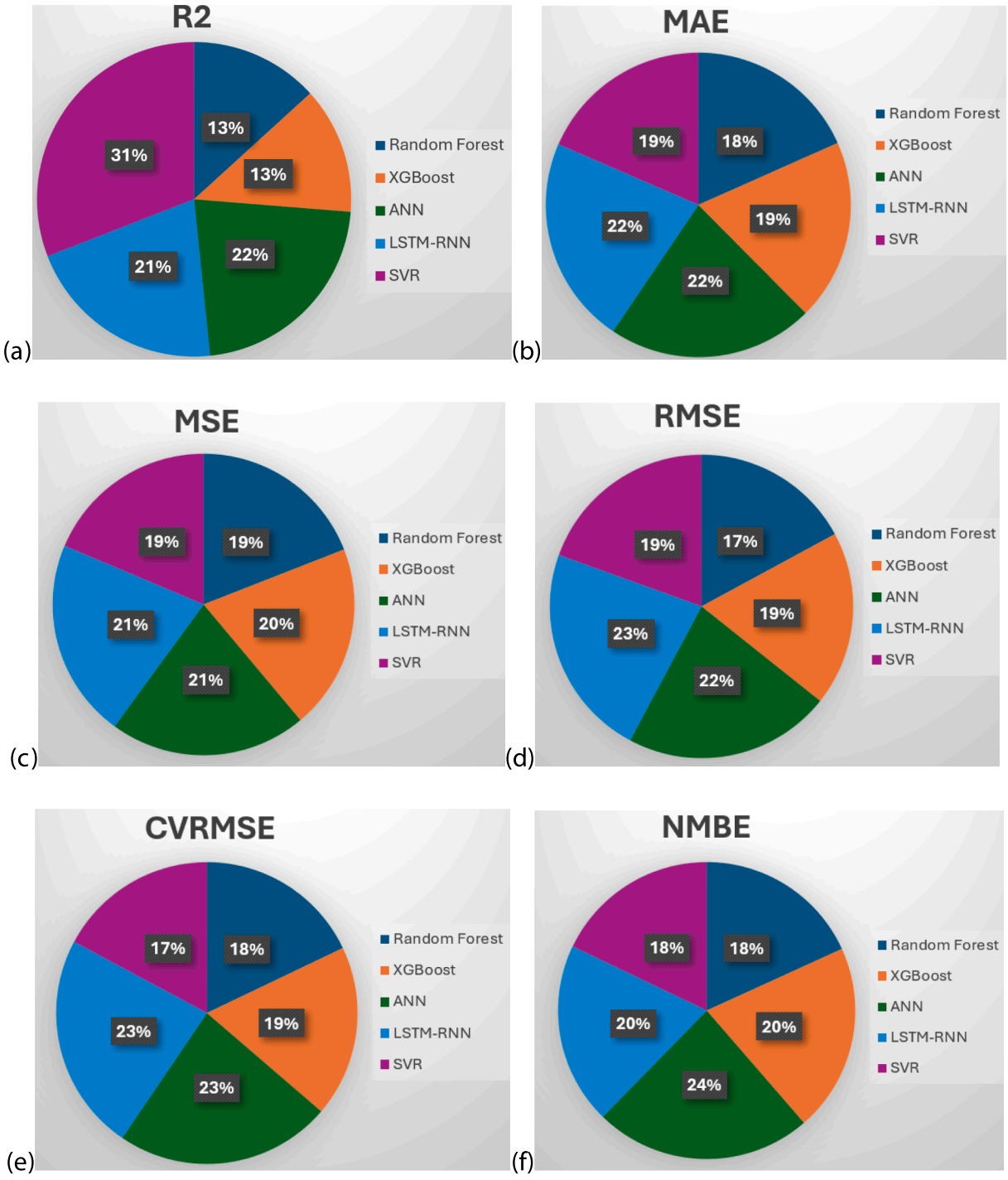


Fig. 29. Model performance improvement after hyperparameter tuning (a) R^2,(b) MAE, (c) MSE, (d) RMSE, (e) CVRMSE, (f) NMBE.

4. Conclusions

This study presented a comprehensive data-driven methodology for evaluating and optimising the operational performance of residential air-to-water heat pumps (A2WHPs) using real-time IoT-enabled monitoring and machine learning. By combining field-based experiments with advanced analytics, the study bridges the gap between theoretical modelling and actual system behaviour under dynamic residential conditions. Eight ML models were rigorously evaluated using structured preprocessing, principal component analysis, and hyperparameter tuning via GridSearchCV and k-fold cross-validation. Among these, artificial neural networks (ANN) and long short-term memory networks (LSTM) consistently achieved superior prediction of coefficient of performance (COP), achieving an accuracy of R^2 values of 0.998 and 0.999, respectively, and minimal error across all evaluation metrics, including MAE, MSE, RMSE, CVRMSE and NMBE. Including diagnostic tools such as residual histograms and KDE plots added interpretability to model outputs, helping to detect prediction bias and assess generalisation across unseen data. Beyond predictive accuracy, the findings emphasise the value of integrating ML techniques with real-time sensor data to inform fault detection, energy optimisation, and performance diagnostics in residential heating systems. This approach provides a replicable and scalable solution for smart building management, supporting broader energy transition goals by enhancing system-level

efficiency and mitigating operational uncertainty. The methodology in this study can be extended to diverse housing typologies, varied climatic zones, and other low-carbon heating technologies.

Future research should investigate the integration of this approach with reinforcement learning for adaptive control and real-time optimisation, enabling intelligent, autonomous residential energy systems aligned with smart grid and decarbonisation goals. Incorporating physics-based models into A2WHP performance predictions may enhance model generalisability. Although the machine learning models in this study performed well within the training horizon, their extrapolation to unseen seasonal conditions remains constrained, a common challenge with purely data-driven models. Ongoing further work will compare the performance of these ML models against deep learning architectures, including Physics-informed long short-term memory (PI-LSTM) and Physics-informed neural network (PINN), to address this limitation. This benchmarking will support the selection of models best suited for energy optimisation, control, and fault detection and diagnosis (FDD) in residential settings. Ultimately, this work will contribute to building a robust, scalable framework for energy-efficient and thermally comfortable homes, with potential for integration in building energy management (BEM) systems.

CRediT authorship contribution statement

Rasheed B. Ayoola: Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Olusegun M. Ilori:** Writing – review & editing, Supervision, Funding acquisition. **Noel Perera:** Writing – review & editing, Supervision. **Monica Mateo-Garcia:** Writing – review & editing, Supervision. **Kabir Akinyemi:** Writing – review & editing, Formal analysis. **David Boyd:** Writing – review & editing, Supervision. **Mike Leonard:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors gratefully acknowledge the PhD funding support provided by the Faculty of Computing, Engineering and the Built Environment (CEBE), Birmingham City University, through the Faculty Dean Scholarship Award.

Data availability

Data will be made available on request.

References

[1] European Heat Pump Association (EHPA), 2025. Market Report 2025: Executive Summary. Brussels: EHPA. Available at: https://www.ehpa.org/wp-content/uploads/2025/07/EHPA-Market-Report-2025-executive-summary.pdf [Accessed 23 July 2025].

[2] C.Masternak, S. Meunier, V. Reinbold, D. Saelens, C. Marchand, Y. Leroy, Potential of air-source heat pumps to reduce environmental impacts in 18 European countries, Energy 292 (2024) 130487.

[3] Wang, X., Xia, L., Bales, C., Zhang, X., Copertaro, B., Pan, S. and Wu, J., 2020. A systematic review of recent air source heat pump (ASHP) systems assisted by solar thermal, photovoltaic and photovoltaic/thermal sources. Renewable energy, 146, pp.2472-2487. 1.

[4] Dunbabin and Wickins. (2012). *Detailed analysis from the first phase of the Energy Saving Trust's heat pump field trial.* www.energysavingtrust.org.uk/Media/node_1422/Getting-warmer-a-field-trial-of-heat-pumps-PDF.

[5] Department for Energy Security and Net Zero (DESNZ) (formerly BEIS), 2023. Interim Heat Pump Performance Data Summary—Electrification of Heat Demonstration Project. Energy Systems Catapult on behalf of DESNZ. Available at: [URL if known] (Accessed July 2025).

[6] B. Gordon, Linear electronic analog/digital conversion architectures, their origins, parameters, limitations, and applications, IEEE Transactions on Circuits and Systems 25 (7) (1978) 391–418.

[7] K. Djebko, D. Weidner, M. Waleska, T. Krey, B. Kamble, S. Rausch, D. Seipel, F. Puppe, Design and Implementation of a Decision Integration System for Monitoring and Optimizing Heating Systems: Results and Lessons Learned, Energies 17 (2024) 6290.

[8] Z. Jiao, X. Du, Z. Liu, L. Liu, Z. Sun, G. Shi, Sustainable Operation and Maintenance Modeling and Application of Building Infrastructures Combined with Digital Twin Framework, Sensors 23 (2023) 4182.

[9] I. Dincer, C. Acar, Smart energy systems for a sustainable future, Applied Energy 194 (2017) 225–235.

[10] C. Yu, Y.P. Zhu, H. Luo, Z. Luo, L. Li, Design assessments of complex systems based on design-oriented modelling and uncertainty analysis, Mechanical Systems and Signal Processing 188 (2023) 109988.

[11] N. Boccara, N. Boccara, Modeling Complex Systems, Vol. 1, Springer, New York, 2010.

[12] A.S.R. Subramanian, T. Gundersen, T.A. Adams, Modeling and simulation of energy systems: a review, Processes 6 (12) (2018) 238.

[13] V. Mistry, The Role of IoT in Enhancing HVAC Control Systems, J. Biosens. Bioelectron. Res 115 (2023) 2–5.

[14] Rajith, A., Soki, S. and Hiroshi, M., 2018, April. Real-time optimised HVAC control system on top of an IoT framework. In *2018 Third international conference on fog and mobile edge computing (FMEC)* (pp. 181-186). IEEE.

[15] W. Yaïci, E. Entchev, M. Longo, A. Annuk, in: August. Internet of Things (IoT) Monitoring and Control for Smart Heating and Cooling in a Residential Building, IEEE, 2023, pp. 387–392.

[16] X. Gao, Y.A. Shardt, in: August. Mutual Information Induced Slow-Feature Analysis of Non-Linear Dynamic Systems and the Application in Soft Sensors, IEEE, 2022, pp. 319–324.

[17] A.S. Abdelwahed, A. Zekry, G.M. Amer, R.M. Tawfeek, in: Internet of Things Based Energy Management System Using Real-Time Consumption Data, IEEE, 2022, pp. 1–6.

[18] M. Corrà, E. Fusari, A. Ferrari, D. Macii, in: A System Based on IoT Platforms and Occupancy Monitoring for Energy-Efficient HVAC Management, IEEE, 2019, pp. 347–352.

[19] Chew, Michael Yit Lin, and Ke Yan. "Intelligent and Automated Fault Detection and Diagnosis Strategy for HVAC Systems Based on Maintainability Rules for Construction 4.0." In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pp. 1946-1951. IEEE, 2021.

[20] H. Dong, B. Liu, D. Ye, G. Liu, Interpretability as Approximation: Understanding Black-Box Models by Decision Boundary, Electronics 13 (22) (2024) 4339.

[21] Yoshikawa, Y., Kimura, M., Shimizu, R. and Saito, Y., 2024. Explaining Black-box Model Predictions via Two-level Nested Feature Attributions with Consistency Property. *arXiv preprint arXiv:2405.14522*.

[22] Pawar, A., Shenoy, K.M., Prabhu, S. and Rai, D.G., 2023, October. Performance analysis of machine learning algorithms: Single Model VS Ensemble Model. In *Journal of Physics: Conference Series* (Vol. 2571, No. 1, p. 012007). IOP Publishing.

[23] M. Chesser, P. Lyons, P. O'Reilly, P. Carroll, Air source heat pump in-situ performance, Energy and Buildings 251 (2021).

[24] H. Alizadegan, B. Rashidi Malki, A. Radmehr, H. Karimi, M.A. Ilani, Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction, Energy Explor. Exploit. (2024) 01445987241269496.

[25] Nouvel, R., COTRADO SEHGELMEBLE, M. and Pietruschka, D., 2015. European mapping of seasonal performances of air-source and geothermal heat pumps for residential applications. In Proceedings of International Conference CISBAT, *Future buildings and Districts Sustainability from Nano to Urban Scale*, EPFL, LESO-PB, 2015, pp. 543–548.

[26] O. Ruhnau, L. Lundström, L. Dürr, F. Hunecke, (2023). Empirical weather dependency of heat pump load: Disentangling the effects of heat demand and efficiency. In Proceedings of the 19th International Conference on the European Energy Market (EEM), Lappeenranta, Finland, 1–5 June 2023.

[27] T. Brudermueller, F. Breer, T. Staake, in: *For Energy-Efficient Buildings, Cities, and Transportation*, 2023, pp. 228–231.

[28] H.U. Cho, Y. Nam, E.J. Choi, Y.J. Choi, H. Kim, S. Bae, J.W. Moon, Comparative analysis of the optimised ANN, SVM, and tree ensemble models using Bayesian optimisation for predicting GSHP COP, Journal of Building Engineering 44 (2021) 103411.

[29] DUMAN, N., YÜKSEK, A.G., BUYRUK, E. and CANER, M., 2024. Performance Prediction of Ground Source Heat Pump System With Data-Driven Modelling: Application of Machine Learning Techniques.

[30] M. Ma, O. Pektezel, V. Ballerini, P. Valdiserri, E. Rossi di Schio, Performance predictions of Solar-Assisted Heat Pumps: Methodological Approach and Comparison between Various Artificial Intelligence Methods, Energies 17 (22) (2024) 5607.

[31] S.L. Tangwe, M. Simon, Quantification of the viability of residential air source heat pump water heaters as potential replacement for geysers in South Africa, Journal of Engineering, Design and Technology 17 (2) (2019) 456–470.

[32] D. Leiria, H. Johra, A. Marszal-Pomianowska, M.Z. Pomianowski, P.K. Heiselberg, Using data from smart energy meters to gain knowledge about households connected to the district heating network: a danish case, Smart Energy 3 (2021) 100035.

[33] OpenEnergyMonitor. (2024). *Level 3 Heat Pump Monitoring Bundle (emonHP)*. [online] Available at:https://shop.openenergymonitor.com/level-3-heat-pump-monitoring-bundle-emonhp/?utm_source=chatgpt.com [Accessed 1 Aug. 2025].

[34] OpenEnergyMonitor: Open-source energy monitoring platform. https://openenergymonitor.org.

[35] L. Xu, J.D. Choi, Modeling task interactions in document-level joint entity and relation extraction. Proceedings of the 2022 Conference of the North American Chapter of the Association for, Computational Linguistics: Human Language Technologies (2022) 5409–5416.

[36] J. Zhang, Y. Xu, H. Chen, L. Xing, A novel building heat pump system semi-supervised fault detection and diagnosis method under small and imbalanced data, Engineering Applications of Artificial Intelligence 123 (2023) 106316.

[37] M. Hoffmann, L. Kotzur, D. Stolten, M. Robinius, A review on time series aggregation methods for energy system models, Energies 13 (3) (2020) 641.

[38] S.A. Alasadi, W.S. Bhaya, Review of data preprocessing techniques in data mining, Journal of Engineering and Applied Sciences 12 (16) (2017) 4102–4107.

[39] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC Trends in Analytical Chemistry 132 (2020) 116045.

[40] Morvan, M.L. and Vert, J.P., 2017. Supervised quantile normalisation. *arXiv preprint arXiv:1706.00244*.

[41] J. Raymaekers, P.J. Rousseeuw, Transforming variables to central normality, Machine Learning 113 (8) (2024) 4953–4975.

[42] V. Romanuke, Impact of starting outlier removal on accuracy of time series forecast ING. *Maritime Technical*, Journal. (2022).

[43] F. Ritter, A procedure to clean, decompose, and aggregate time series, Hydrology and Earth System Sciences 27 (2) (2023) 349–361.

[44] G. Agapito, C. Zucco, M. Cannataro, COVID-warehouse: a data warehouse of Italian COVID-19, pollution, and climate data, Int. J. Environ. Res. Public Health 17 (15) (2020) 5596.

[45] D. Singh, B. Singh, Investigating the impact of data normalisation on classification performance, Applied Soft Computing 97 (2020) 105524.

[46] F.L. Huang, W. Wiedermann, B. Zhang, Accounting for heteroskedasticity resulting from between-group differences in multilevel models, Multivariate Behavioral Research 58 (3) (2022) 637–657.

[47] Sladekova, M. and Field, A.P., 2024. *Quantifying heteroscedasticity in linear models using quantile LOWESS intervals* [online].

[48] H. Tian, X. Wu, Z. Chen, J. Huang, A new approach for improving the robustness of regression models via response variable transformation, Mathematics 11 (17) (2023) 3782.

[49] Hyndman, R.J., 2018. *Forecasting: principles and practice*. OTexts.

[50] Wen, Q., Zhang, Z., Li, Y. and Sun, L., 2020, August. Fast RobustSTL: Efficient and robust seasonal-trend decomposition for time series with complex patterns. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2203-2213). Ramallo-González, A.P., 2013. *Modelling, simulation and optimisation methods for low-energy buildings*. University of Exeter (United Kingdom).

[51] S.T.L. Cleveland, A seasonal-trend decomposition procedure based on Loess (with discussion), J. off. Stat 6 (3) (1990).

[52] Jin, L., Spurlock, C.A., Borgeson, S., Lazar, A., Fredman, D., Todd, A., Sim, A. and Wu, K., 2021. Investigating Underlying Drivers of Variability in Residential Energy Usage Patterns with Daily Load Shape Clustering of Smart Meter Data. *arXiv preprint arXiv:2102.11027*.

[53] Kodra, E. and Ganguly, A.R., 2014. Asymmetry of projected increases in extreme temperature distributions. Sci. Rep., 4, 5884 [online].

[54] C. Hanmer, D. Shipworth, M. Shipworth, E. Carter, Household thermal routines and their impact on space heating demand patterns. European Council for an Energy Efficient Economy (ECEEE), 2017.

[55] F. Acito, Ensemble Models, in: *Predictive Analytics with KNIME: Analytics for Citizen Data Scientists*, Cham, Springer Nature Switzerland, 2023, pp. 255–265.

[56] Saxena, P.S., 2024. An Approach through Different Mathematical Models to Enhance the Utility in Different Areas of Machine Learning. In *Deep Learning Concepts in Operations Research* (pp. 124-129). Auerbach Publications.

[57] B. Sekeroglu, A. Altan, P.D. Barua, Comparative evaluation and comprehensive analysis of machine learning models for regression problems, Academia Letters, Article (2022) 5939.

[58] I. K. Nti, S. Akyeramfo-Sam, B. Bediako-Kyeremeh, S. Agyemang, Prediction of social media effects on students' academic performance using Machine Learning Algorithms (MLAs), Journal of Computers in Education 9 (2021) 195–223.

[59] D. Alladi, Machine Learning Algorithms for Predictive Analytics: A Review and Evaluation, SSRN Tech, Rep (2016).

[60] E. Agliari, F. Alemanno, M. Aquaro, A. Fachechi, Regularisation, early-stopping and dreaming: a Hopfield-like setup to address generalisation and overfitting, Neural Netw. 177 (2024) 106389.

[61] N.S. Rao, in: June. Study of Overfitting by Machine Learning Methods Using Generalization Equations, IEEE, 2023, pp. 1–8.

[62] N.C. Chesnaye, M. van Diepen, F. Dekker, C. Zoccali, K.J. Jager, V.S. Stel, Non-linear relationships in clinical research, Nephrology Dialysis Transplantation p. gfae187 (2024).

[63] R.J. Knell, On the analysis of non-linear allometries, Ecological Entomology 34 (1) (2009) 1–11.

[64] Fox, J. and Stine, R., 2001. Detecting Non-linearity in 3D Dynamic Graphs. *Available from: citeseerx. ist. psu. edu [20 August 2013]*.

[65] B.J. Stolz, J. Tanner, H.A. Harrington, V. Nanda, Geometric anomaly detection in data, Proceedings of the National Academy of Sciences 117 (33) (2020) 19664–19669.

[66] X. Guo, A.P. Goumba, Air source heat pump for domestic hot water supply: Performance comparison between individual and building scale installations, Energy 164 (2018) 794–802.

[67] G. Dudek, A comprehensive study of random forest for short-term load forecasting, Energies 15 (20) (2022) 7547.

[68] J. Che, J. Wang, Short-term load forecasting using a kernel-based support vector regression combination model, Applied Energy 132 (2014) 602–609.

[69] R.C. Chen, C. Dewi, S.W. Huang, R.E. Caraka, Selecting critical features for data classification based on machine learning methods, Journal of Big Data 7 (1) (2020) 52.

[70] S. Hochreiter, Long short-term memory, Neural Computation MIT-Press. (1997).

[71] B.H. Shekar, G. Dagnew, in: February. Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data, IEEE, 2019, pp. 1–8.

[72] D.M. Belete, M.D. Huchaiah, Grid search in hyperparameter optimisation of machine learning models for prediction of HIV/AIDS test results, International Journal of Computers and Applications 44 (9) (2022) 875–886.

[73] V. Ignatenko, A. Surkov, S. Koltcov, Random forests with parametric entropy-based information gains for classification and regression problems, PeerJ Computer Science 10 (2024) e1775.

[74] J. Zhu, A. Zhang, H. Zheng, Research on Predictive Model based on Ensemble Learning, Highlights in Science, Engineering and Technology 57 (2023) 311–319.

[75] R. Sahu, S. Srivastava, R. Kumar, in: November. Modelling of a Non-Linear Dynamic System Using Long Short-Term Memory, IEEE, 2023, pp. 1016–1021.

[76] H. Agarwal, G. Mahajan, A. Shrotriya, D. Shekhawat, Predictive Data Analysis: Leveraging RNN and LSTM Techniques for Time Series Dataset, Procedia Comput. Sci. 235 (2024) 979–989.

[77] Y.O. Ouma, R. Cheruyot, A.N. Wachera, Rainfall and runoff time-series trend analysis using LSTM recurrent neural network and wavelet neural network with satellite-based meteorological data: case study of Nzoia hydrologic basin, Complex & Intelligent Systems (2021) 1–24.

[78] L. Scorzato, Reliability and Interpretability in Science and Deep Learning, Minds and Machines 34 (3) (2024) 27.

[79] N. Boyko, K. Boksho, P. Telishevskyi, Neural Networks: Training with Backpropagation and the Gradient Algorithm, in: In *2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC s&t)*, 2022, pp. 543–550.

[80] Granderson, J., Touzani, S., Custodio, C., Sohn, M., Fernandes, S. and Jump, D., 2015. Assessment of automated measurement and verification (M&V) methods.

[81] E.A.U. Malahina, G.R. Iriane, Y.S. Belutowe, P. Katemba, J. Asmara, A Grid-search Method Approach for Hyperparameter Evaluation and Optimisation on Teachable Machine Accuracy: a Case Study of Sample size Variation, Journal of Applied Data Sciences 5 (3) (2024) 1008–1025.

[82] Y. Zhang, V. Cheng, D.S. Mallapragada, J. Song, G. He, A model-adaptive clustering-based time aggregation method for low-carbon energy system optimisation, IEEE Transactions on Sustainable Energy 14 (1) (2022) 55–64.

[83] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459.

[84] P. Kynčlová, P. Filzmoser, K. Hron, Compositional biplots including external non-compositional variables, Statistics 50 (5) (2016) 1132–1148.

[85] Gniazdowski, Z., 2023. On the Correlation between Random Variables and their Principal Components. *arXiv preprint arXiv:2310.06139*.

[86] Wang, G., Lauri, F., Wang, P. and Luo, H., 2022. An Adaptive Neighborhood Partition Full Conditional Mutual Information Maximisation Method for Feature Selection. *arXiv preprint arXiv:2210.12002*.

[87] J.A. Barradas-Palmeros, E. Mezura-Montes, R. Rivera-López, H.G. Acosta-Mesa, Computational cost reduction in wrapper approaches for feature selection: a case of study using permutational-based differential evolution, in: In *2024 IEEE Congress on Evolutionary Computation (CEC)*, 2024, pp. 1–8.

[88] López-De-Castro, M., García-Galindo, A. and Armañanzas, R., 2024. Conformal Recursive Feature Elimination. *arXiv preprint arXiv:2405.19429*.

[89] P.D. Isles, A random forest approach to improve estimates of tributary nutrient loading, Water Research 248 (2024) 120876.

[90] V. Lytvynenko, O. Naumov, M. Voronenko, O. Naumov, I. Lurie, P. Radovenchuk, in: September. Application of Mutual Information Estimation Methods for Feature Selection in Regression Problems of Aviation Traffic Forecasting, IEEE, 2024, pp. 784–789.

[91] M.I. Prasetiyowati, N.U. Maulidevi, K. Surendro, Feature selection to increase the random forest method performance on high dimensional data, International Journal of Advances in Intelligent Informatics 6 (3) (2020) 303–312.

[92] J.L. Speiser, M.E. Miller, J. Tooze, E. Ip, A comparison of random forest variable selection methods for classification prediction modeling, Expert Systems with Applications 134 (2019) 93–101.

[93] F. Kuran, G. Tanırcan, E. Pashaei, Performance evaluation of machine learning techniques in predicting cumulative absolute velocity, Soil Dynamics and Earthquake Engineering 174 (2023) 108175.

[94] A. Yafouz, N. AlDahoul, A.H. Birima, A.N. Ahmed, M. Sherif, A. Sefelnasr, M. F. Allawi, A. Elshafie, Comprehensive comparison of various machine learning algorithms for short-term ozone concentration prediction, Alexandria Engineering Journal 61 (6) (2022) 4607–4622.

[95] ASHRAE (2023). *ASHRAE Guideline 14-2023: Measurement of Energy, Demand, and Water Savings*. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. Available at: https://webstore.ansi.org [Accessed 3 Aug. 2025].

[96] M. Revilla-León, A. Gohil, A.B. Barmak, A. Zandinejad, A.J. Raigrodski, J. Alonso Pérez-Barquero, Best-fit algorithm influences on virtual casts' alignment discrepancies, Journal of Prosthodontics 32 (4) (2023) 331–339.

[97] R. Judkoff, J. Neymark, Model validation and testing: The methodological foundation of ASHRAE Standard 140 [Conference paper], ASHRAE Transactions 112 (Part 2) (2006) 367–376.
[98] S. Hossain, G. Kaur, in: May. Stock Market Prediction: Xgboost and LSTM Comparative Analysis, IEEE, 2024, pp. 1–6.
[99] C.R. Standridge, Beyond lean, Simulation in Practice, 2013.
[100] O'Neill, O. and Costello, F., 2023. Systematic Bias in Sample Inference and its Effect on Machine Learning. *arXiv preprint arXiv:2307.01384*.
[101] J. Shin, J. Lee, Y. Cho, A COP prediction model of hybrid geothermal heat pump systems based on ANN and SVM with hyper-parameters optimisation, Applied Sciences 13 (13) (2023) 7771.