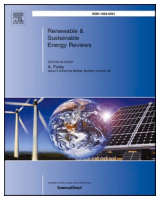
Renewable and Sustainable Energy Reviews 189 (2024) 113998

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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



# A systematic review of data analytics applications in above-ground geothermal energy operations

Paul Michael B. Abrasaldo *, Sadiq J. Zarrouk, Andreas W. Kempa-Liehr

*Department of Engineering Science and Biomedical Engineering, The University of Auckland, Private Bag 92019, Auckland, New Zealand*

ARTICLE INFO

*Keywords:*
Geothermal energy
Data analytics
Machine learning
Feature selection
Time-series
Data-quality

ABSTRACT

The advent of reliable and inexpensive sensors and advancements in general computing have made data-heavy algorithms feasible for operational, real-time decision-making applications in the geothermal energy industry. This systematic review aims to provide a starting point for researchers interested in developing data-driven systems, tools, and frameworks to enhance the performance and reliability of above-ground geothermal energy operations.

The approach and results of the review are presented to answer the following research questions: how has data analytics been applied in above-ground geothermal operations, what data sets have been used in such studies, which types of machine learning or artificial intelligence algorithms have been used in geothermal studies, and at which stages of geothermal development have studies been applied. Published research articles were retrieved from four literature databases: the International Geothermal Association (IGA) online library, ScienceDirect, SpringerLink, and IEEE Xplore. A total of 830 publications were retrieved using the same search query across the selected databases, from which 63 research papers were selected based on a set of inclusion and exclusion criteria. A full-text evaluation of the selected research papers revealed that machine learning has been used in geothermal for design optimisation, performance monitoring, performance optimisation, fault detection, and other applications. Most of the trained models (95 %) were of the artificial neural network family with other model types generally used as performance benchmarks. The systematic review revealed significant potential for further research and applications in the areas of feature selection, systematic time-series feature engineering, and model evaluation.

## 1. Introduction

Geothermal energy has been a reliable source of natural heat and has been utilised for bathing, washing, cooking, and space heating in many cultures worldwide since ancient times [1–3]. The 20th century saw geothermal energy grow into one of the most reliable and renewable sources of electricity, backed by a global push for clean energy development. This growth in electricity production using geothermal energy peaked in the late 1980s until the early 1990s, after a combined effort from the public and private sectors to develop liquid-dominated geothermal resources worldwide [4]. Further increase in the installed capacities of existing steam-dominated fields in the USA, Italy, and Indonesia after the Second World War combined with new installations in the "wet steam" fields of New Zealand, the Philippines, Mexico, Turkey, and other parts of Asia and Europe resulted in the fastest growth of geothermal energy production for electricity [1,5]. Interest in geothermal energy continued to rise due to the public's increasing awareness regarding the adverse effects of burning fossil fuels on the population's quality of life and several economic crises driven by the demand and supply of oil in the world markets [6–11].

From the mid-1990s until the present, the annual growth rate of geothermal energy installed capacity has slowed to an average of 4 % per year, significantly lower than in the mid-80s, wherein yearly growth rates of more than 10 % were common (Fig. 1). Known economic, political, and technical challenges have continued to hound geothermal energy development throughout the decades. The more extensive and easily accessible geothermal resources have already been explored and exploited. Many other potential sites are in remote areas, needing significant civil work to access and conduct more detailed surveillance studies to prove economic viability [12]. Any developer intending to kickstart a geothermal power plant project must navigate pre-development requirements such as securing land rights, regulatory compliance, fund sourcing, and other factors that vary across different

* Corresponding author.
*E-mail address:* pabr612@aucklanduni.ac.nz (P.M.B. Abrasaldo).

https://doi.org/10.1016/j.rser.2023.113998
Received 11 June 2023; Received in revised form 11 October 2023; Accepted 21 October 2023
Available online 10 November 2023

**Nomenclature**

| | |
|---|---|
| $b$ | Linear regression coefficient |
| $L$ | Number of hidden layers |
| $n$ | Number of samples |
| $p$ | Pressure |
| $R^2$ | Coefficient of determination |
| $T$ | Temperature |
| $t$ | Time |
| $u$ | System control parameter |
| $w$ | Model weights |
| $X$ | Model input matrix |
| $x$ | Model input vector |
| $y$ | Model output |

*Abbreviations*

| | |
|---|---|
| AGDHS | Afyon geothermal district heating system |
| AI | Artificial intelligence |
| ANN | Artificial neural network |
| BPNN | Backpropagation neural network |
| CNN | Convolutional neural network |
| CVRMSE | Coefficient of the variation of the root mean square error |
| EC | Exclusion criteria |
| DT | Decision Tree |
| FFNN | Feed-forward neural network |
| GRNN | General regressor neural network |
| IC | Inclusion criteria |
| IEEE | Institute of Electrical and Electronics Engineers |
| IGA | International Geothermal Association |
| LDA | Latent Dirichlet allocation |
| LR | Linear regression |
| MAPE | Mean absolute percentage error |
| MEI | Minimum equivalence interval |
| ML | Machine learning |
| MLR | Multivariate linear regression |
| MSE | Mean squared error |
| NN | Neural network |
| ORC | Organic Rankine cycle |
| PCA | Principal component analysis |
| PID | Proportional integral derivative |
| PRISMA | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| PWF | Present worth factor |
| RMSE | Root mean square error |
| RNN | Recurrent neural network |
| ROCAUC | Area under the receiver operating characteristic curve |
| RQ | Research question |
| SSC | Steam consumption coefficient |
| STRidge | Sequential threshold ridge regression |
| SVM | Support vector machines |

*Greek letters*

| | |
|---|---|
| $\varphi$ | Activation function |
| $\delta$ | Delta rule or error term |
| $\eta$ | Learning rate or efficiency |

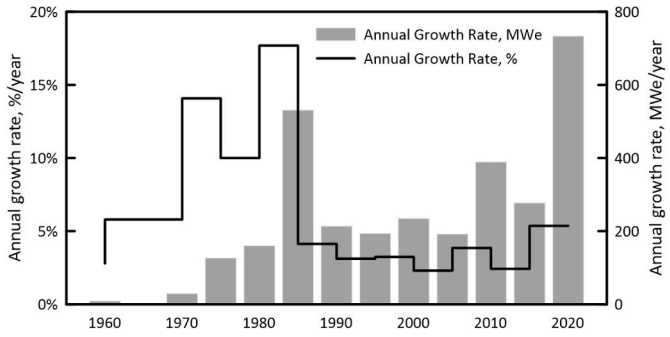



**Fig. 1.** Global growth of electricity produced from geothermal energy sources. Adapted from Bertani [13,14] and Huttrer [15,16].

countries.

The effects of long-term geothermal energy extraction in many of the larger and older geothermal projects built during the peak of geothermal energy growth have started to surface in recent years. The Geysers geothermal site in California, USA, decommissioned several turbine units during the 1990s due to significant pressure drawdown in the reservoir caused by over-extraction [17]. Aside from reservoir pressure decline, the Mahanagdong field in the Philippines experienced reservoir cooling caused by reinjected fluid flooding the main production areas, resulting from "stressed" production levels [18]. There have been reports of geysers and other natural surface features that disappeared in parts of New Zealand as a side-effect of geothermal energy extraction [19,20]. In all these resource management challenges, geothermal field operators worldwide have continued to build upon existing knowledge and technology to approach a more sustainable electricity production from geothermal energy.

Data and data-processing technologies have been essential to the growth of geothermal resource management strategies over the years [21–27]. The different stages of problem identification, solution development, and performance monitoring require good-quality data to increase the likelihood of a successful geothermal project. The role of data has only grown as new data capture and storage technologies have been deployed across geothermal sites. A significant number of reports have already been made on successful applications of artificial intelligence (AI), machine learning (ML), and other data analytics approaches, which usually require large volumes of data [28]. Understandably, such studies have focused on characterising the subsurface geothermal reservoir. Still, there is significant merit in pursuing applications of these data-driven approaches to solving problems that arise within surface facility operations due to the unique conditions posed by geothermal energy production.

Therefore, this systematic review paper investigates the different applications of data science and analytics in the context of above-ground geothermal operations. This work will highlight the type of data used and the primary use cases or problems being solved in such studies. Any reported performance, reliability, or efficiency enhancement from such work will also be presented. This systematic review can be treated as an introduction for new researchers, particularly geothermal scientists and engineers, to artificial intelligence and data analytics in above-ground geothermal operations. Other review papers relating to machine learning applications in geothermal are present, such as the works of Okoroafor et al. [28] and Muther et al. [29]. These review papers, however, primarily focused on AI or ML applications related to the subsurface aspects of geothermal energy production.

## 2. Research methodology

A systematic review is a structured approach to conducting literature surveys that follow a relatively strict methodological framework that promotes the reproducibility of such studies [30]. This method allows for the extraction and synthesis of information that can aid researchers

in identifying gaps in knowledge of a specific discipline concerning a set of defined research questions and protocols. The method used in this work follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist, which is an "evidence-based minimum set of items for reporting in systematic reviews and meta-analyses" [31]. Since the PRISMA checklist was constructed mainly for the field of medicine, some of the checklist items related to the meta-analyses criteria are not considered in this review.

The systematic review process used in this paper is comprised of several steps. The process started with formulating the research questions this review aims to answer. The search process and strategy are then described, including the keywords and search strings used to find relevant and available publications from the literature databases. The metadata of the retrieved publications then went through topic modelling, which was carried out to help refine the inclusion and exclusion criteria. These criteria are then explicitly defined to assist with filtering relevant literature. Publication data such as the title, abstract, year of publication, and full text are extracted for the papers that reach this stage. Information from the selected publications relevant to the predefined research questions is then synthesised and presented in the latter sections of this paper. The steps taken for this systematic review and the identified risks for bias are described in detail in the following subsections.

### 2.1. Research questions

This systematic review was motivated to provide new researchers with an introduction to the field of data science and its existing applications in geothermal energy operations. Thus, the following research questions (RQs) are formulated, which this paper aims to answer.

- RQ1: What methods, techniques, algorithms, or approaches from data science and analytics have been applied in the geothermal setting?
- RQ2: What domains did the data used in the reported applications originate from?
- RQ3: What was the primary task (e.g., forecasting, classification, anomaly detection) in the existing applications?
- RQ4: Which stage of geothermal development and surface facility components were targeted by such applications?

RQ1 investigates the prevalent algorithms and approaches used in the geothermal industry's existing data science or analytics applications worldwide. This question further looks into the model design decisions that researchers made considering the type of data being used and target applications in geothermal. RQ2 allows us to identify the equally important information of which domain the datasets used in such application were sourced. RQ3 takes it one step further by recording what type of problem and solution approach researchers used in the studies. At the same time, RQ4 gives an insight into the common areas or stages of geothermal development that have been studied using data-heavy approaches.

### 2.2. Publication search

Four databases that researchers in the industry commonly access are used to search for relevant articles about data analytics applications in geothermal. These databases are.

- International Geothermal Association (IGA) conference paper database
- ScienceDirect
- SpringerLink
- IEEE (Institute of Electrical and Electronics Engineers) Xplore

The search process begins by defining the search string used to search the publication databases. Since many keywords related to the topic at hand, a boolean query was formulated to cover most, if not all, of the desired studies:

("artificial intelligence" OR "machine learning" OR "neural network" OR "data science") AND "geothermal".

The "geothermal" keyword at the end of the search string above was not necessary when searching the IGA database but was essential when querying the other three literature sources. The initial search results using the above query string returned 830 publications from the four publication databases.

### 2.3. Filtering and selection

An intermediate step of topic modelling was conducted based on the initial results of the search query to ensure that the retrieved publications were within the scope of this review. By uncovering patterns of word use and linking documents with similar themes, topic models can discover the underlying structure of a group of documents [32]. In this review, Latent Dirichlet Allocation (LDA) is used to identify and group keywords that frequently occur together, which may describe the general content of each publication returned by the initial query. The title and abstract of each publication returned by the initial search query were modelled using LDA implemented in scikit-learn [33].

The results of the LDA model with four topics based on the initial publication search query results are visualised in Fig. 2 using the pyLDAvis library [34,35], with the intertopical distance showing the marginal topic distribution. The topic modelling results shown in Fig. 2c–f presents the top 30 most relevant terms for each topic and can be used to interpret prevailing themes. It can be observed that Topics 1 and 2 have top words related to geothermal and data use. Topic 1 is interpreted to be applications of ML/AI that relate to the overall performance of a geothermal field with relevant words such as "field", "performance", "temperature", and "flow" appearing in the top terms. However, Topic 2 can be interpreted to be more focused on the ML/AI studies for managing geothermal energy systems. Specific studies focus on the economic aspect of electricity generation, including forecasting overall power demand and other similar themes. Looking at the top 30 keywords of Topic 3, one might classify that topic as "classification of volcanic events" since words such as "volcano", "classification", "event", and "eruption" occur in that cluster. Topic 4 may contain many of the studies fit for inclusion in this review since the top terms for this topic relate to the application of AI/ML in the control and optimisation of power plant systems. This topic modelling step has shown that a handful of papers may not fall under the purview of this review, such as studies more focused on the economic aspect of energy production and theoretical studies related to hybrid electricity production. However, interested readers can look at review papers [36–39] which investigate some of the representative studies of the excluded articles. Publications under Topic 3 were excluded from this review, while further screening of studies under Topics 1, 2, and 4 was conducted to select the relevant papers for this study.

The following inclusion and exclusion criteria were formulated based on the objectives of this review and the results of the topic modelling.

| | |
|---|---|
| Inclusion criteria (IC): | |
| IC1 | Papers that are related to geothermal energy production |
| IC2 | Papers that utilise artificial intelligence, machine learning, or data science/analytics approaches |
| IC3 | Papers that look at above-ground components or systems of geothermal energy operations |
| Exclusion criteria (EC): | |
| EC1 | Papers whose full texts are not available |
| EC2 | Papers that are not original research articles (e.g., reviews) |
| EC3 | Papers that are not related to geothermal energy production |
| EC4 | Papers that involve theoretical optimisation of geothermal hybrid systems |

After duplicate removal and identification of relevant and irrelevant

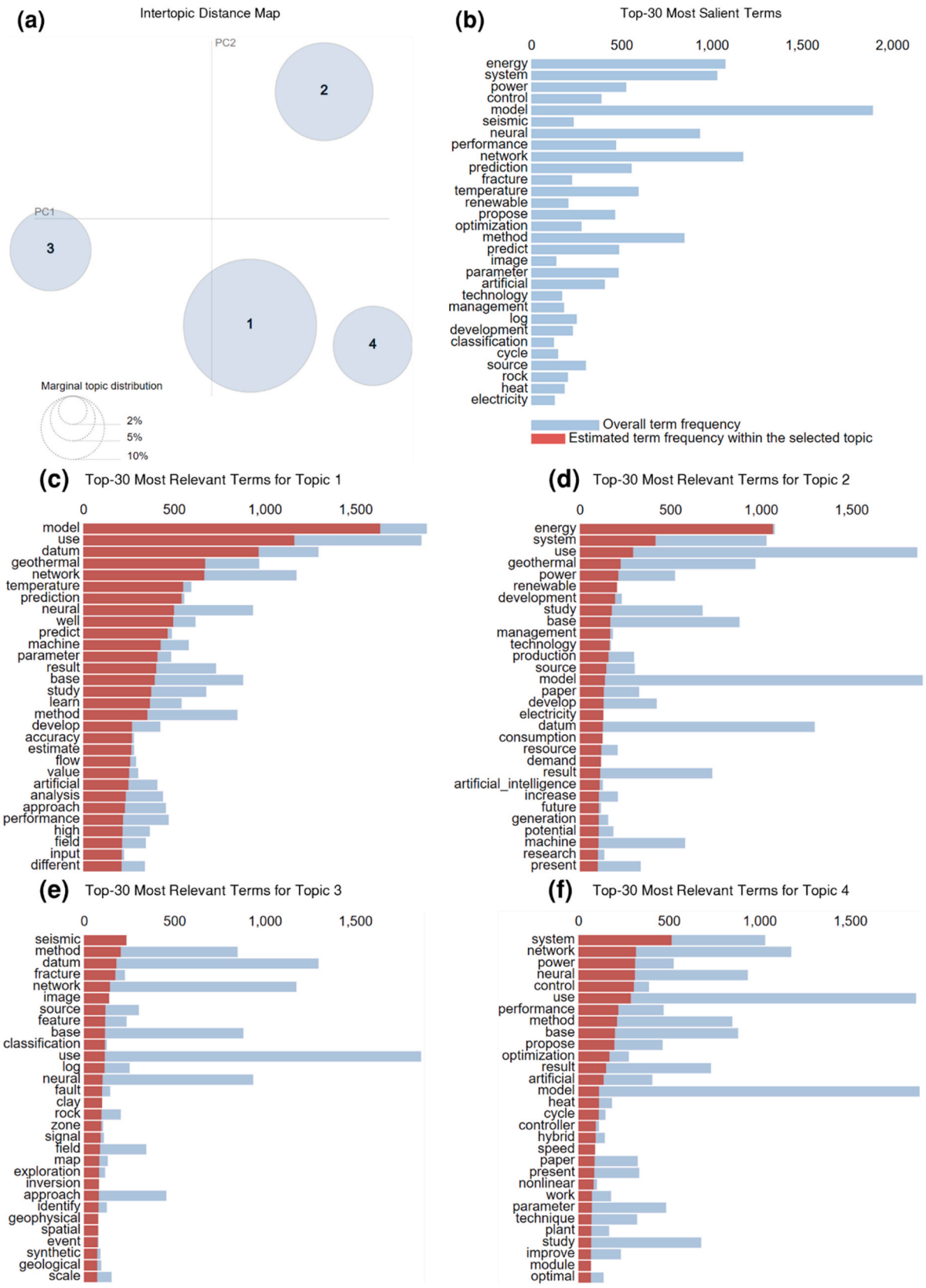



**Fig. 2.** Topic modelling results using LDA showing (a–b) intertopical distance and marginal topic distribution of each topic and (*c*–f) the top 30 most relevant terms for Topics 1–4, respectively, along with the estimated term frequencies. Topics 1, 2, and 4 may cover publications relevant to this systematic review, while papers under topic 3 can be excluded.

papers based on the inclusion and exclusion criteria, an initial selection of 344 documents was made from the original 830 search query results. A final count of 63 research articles was selected for inclusion in this systematic review after a full-text assessment was done to determine their ability and contribution to answering the research questions. A PRISMA flow diagram [31] presenting the number of papers retrieved at each stage of the review process is shown in Fig. 3.

## 3. Results

The data extracted from the publications selected for this review are analysed to answer the research questions. The publications were further classified based on their actual or potential application in operating geothermal power plants, whether for improving the overall efficiency, performance, or reliability of individual components or larger systems. It was observed that more than 90 % of the publications are based on applying varying forms of neural networks as surrogate models for nonlinear, dynamic, complex systems. These surrogate or proxy models are then used for further studies, such as monitoring the condition of a system or individual components, optimising operations for improved economic benefit, detecting faulty operations, or other fundamental studies. Table 1 summarises the main themes of the retrieved articles after a full-text evaluation was done on each one.

The information synthesised from the selected publications is presented in the following sections to answer the formulated research questions. The first subsection deals with RQ1 and RQ2, wherein data analytics approaches used in the geothermal space are discussed, and the data used for such studies are highlighted. The latter subsections attempt to answer RQ3 and RQ4 to identify the primary type of problems the applications have been used on and at which stage of the geothermal energy production life cycle it was utilised.

The earliest publication in the list of articles included for data synthesis was published in 2010. Still, it was not until 2020 that significant growth in research articles that dealt with data science or analytics in above-ground geothermal was observed. This succeeding subsections discuss the different design decisions made by various researchers to apply data science and analytics methods in the above-ground geothermal setting, including data types and sources, model types, and feature selection methods.

### 3.1. Data types and sources

There were 39 (62 %) of the selected studies that developed data models based on experimental or numerically simulated input data. The remaining papers relied on operational data measured during the operation of machinery or energy systems.

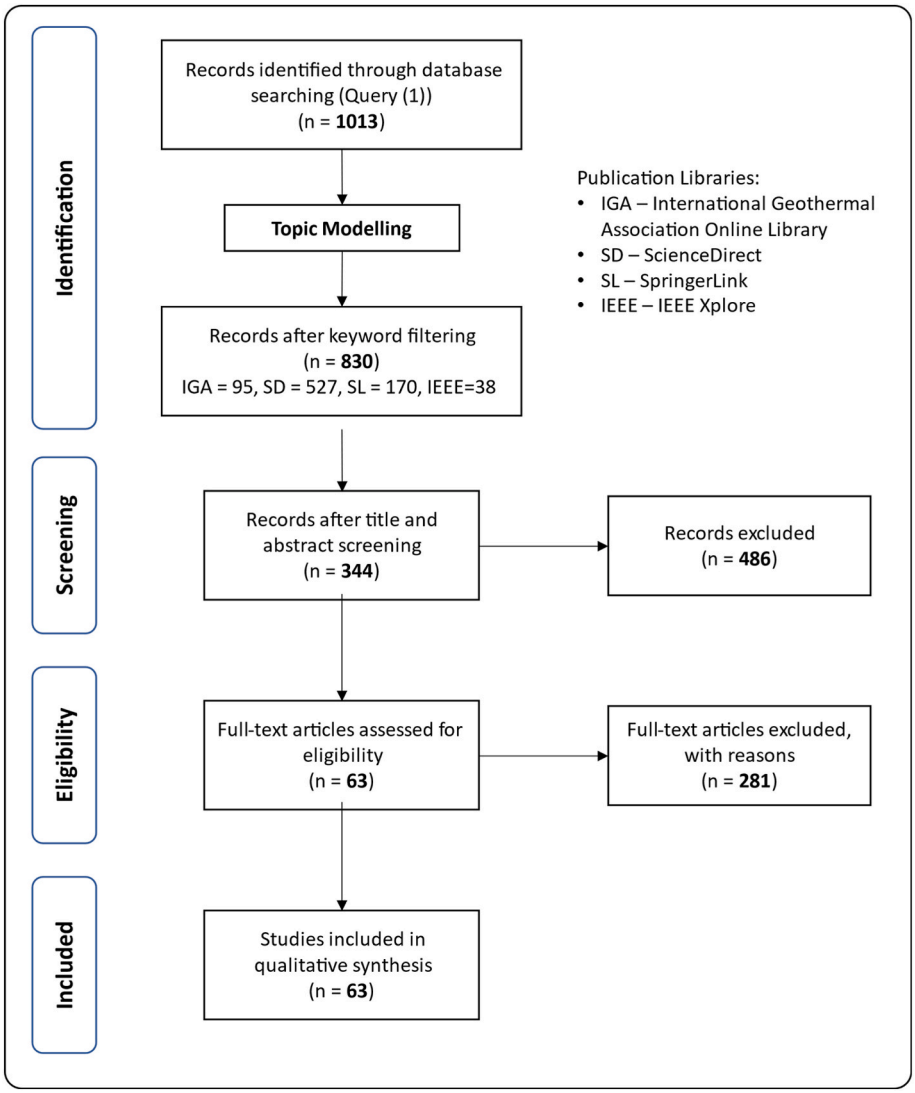



**Fig. 3.** Extended PRISMA flow diagram visualising the publication selection process used in this systematic review. Sixty-three (63) publications were selected for a full-text review and synthesis from the 830 studies retrieved from four (4) research databases.

**Table 1**
Main research themes observed in the articles retrieved from the publication databases are tabulated against the data types used in the papers and the type of ML models that were trained on the input data. Note that there are articles that fall under multiple research themes, have used more than one data type, and utilised more than one ML algorithm in the studies.

| Data Type | AI/ML Algorithm | Design optimisation | Performance monitoring | Performance optimisation | Fault detection | Others | Total No. Of Refs. |
|---|---|---|---|---|---|---|---|
| Experimental | NN | [40–45] | [46–53] | [52,54,55] | [56] | [57–59] | 22 |
| | SVM | | [51,53] | | | [60] | |
| | LR & non-LR | [45] | [51] | | | [60] | |
| | DT & ensembles | | | | [61] | [60] | |
| Simulated | NN | [40–44,62–65] | [66–70] | [54,70–75] | [76] | [77] | 23 |
| | SVM | | [69] | | | | |
| | LR & non-LR | | [69] | | | [78] | |
| | DT & ensembles | | [69] | | | | |
| Operational | NN | [79] | [66,70,80–100] | [70,98–101] | [97,102] | | 27 |
| | SVM | | [94,96,97] | | [97] | | |
| | LR & non-LR | | [78,95,96,98] | [98] | | | |
| | DT & ensembles | | [94,97] | | [97] | | |

On the other hand, eleven (11) studies relied on experimental data to construct data models and had sourced their model inputs from data publications, e.g., as seen in Chang et al. [46] (Table 2) and Huster et al. [40]. Design and performance data for ground-sourced heat pumps previously reported in several publications were utilised by Xu et al. [45] to train a neural network model. Six (6) of the studies that used published data utilised libraries such as REFPROP [103] and CoolProp [104] to generate working fluid properties for Organic Rankine cycle (ORC) systems, which were then used as inputs for thermodynamic and machine learning models [40–43,46,54]. The remaining articles in this group of studies looked at alternative methods to study and predict two-phase flow behaviour based on published look-up tables of experimental data [44,57,58], as well as strategies for handling missing data in a global geochemical database [60].

The publication search yielded twelve (12) research articles that used experimental data measured from test rigs and laboratory setups, which were then used as inputs into the training and evaluation of data models. These studies used data extracted from sensors installed along critical points of a system, such as ground source heat pumps [45,47,48] and ORC systems [51–53,55] (Fig. 4), as well as sensors on individual components such as turbines [49,56], pumps [50,61], and pipelines [59].

Synthetic data from numerical simulations have similarly been used in the literature to train surrogate machine learning models. In such cases, the objective of the research was to develop an alternative model that is as accurate as the physics-based models but requires a fraction of the computational cost to run (Table 3). Models developed in this manner were beneficial for sensitivity and optimisation studies that would have been less feasible with the computationally expensive physics-based models [66,68,75,93].

The selected literature showcased twenty-six (26) research studies that utilised data from sensors installed within operational energy systems [66,70,79–102]. Time-series data are downloaded from the sensors and used to model individual components, e.g., as seen in a study by Siratovich et al. [98]. More commonly, the measured data are sourced from multiple components critical to estimating and monitoring a system's performance (Table 4). It is important to note that data from these studies have not been made available by the authors to other researchers, mainly due to the confidential nature of the source datasets.

Data downloaded from installed sensors generally provide information on the state of the system or component being monitored. This data type is prevalent in industrial applications such as power plants and is

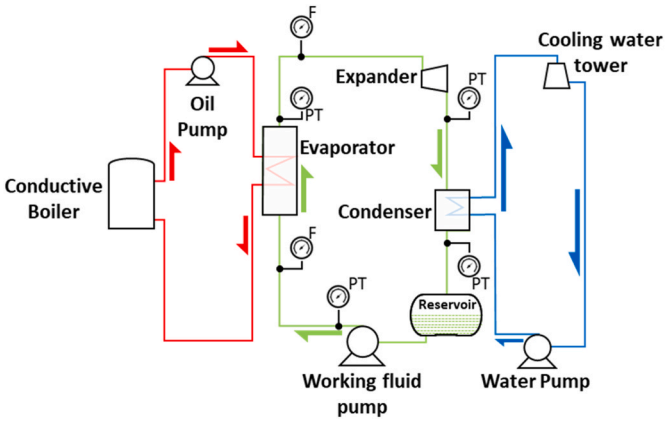



**Fig. 4.** Schematic diagram of the experimental ORC setup used by and adapted from Yan et al. [51]. Data from critical components along the three fluid circuits consisting of the conductive oil heat source (red), motive fluid (green), and cooling water (blue) are used to develop machine learning models.

**Table 2**
Subset of the look-up table used as the primary source of ANN model inputs in a study by Chang et al. [46]. The original dataset was published by Loewenberg et al. [105] in an earlier study.

| *Mass flux* | *Heat flux* | *Pressure* | *Tube Diameter* | *Bulk Enthalpy (kJ /kg)* | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *kg/ $m^2$s* | *kW/$m^2$* | *MPa* | *mm* | 1200 | 1400 | 1600 | 1800 | … | 2700 |
| | | | | *Wall Temperature °C* | | | | | |
| 1000 | 300 | 24 | 8 | 299 | 337 | 366 | 384 | … | 433 |
| 1000 | 300 | 24 | 10 | 299 | 337 | 366 | 384 | … | 433 |
| 1000 | 300 | 24 | 15 | 299 | 337 | 366 | 384 | … | 433 |
| 1000 | 300 | 24 | 20 | 299 | 337 | 366 | 384 | … | 433 |
| 1000 | 300 | 25 | 8 | 299 | 337 | 366 | 384 | … | 433 |
| 1000 | 300 | 25 | 10 | 299 | 337 | 366 | 384 | … | 433 |
| 1000 | 300 | 25 | 15 | 299 | 337 | 366 | 384 | … | 433 |
| 1000 | 300 | 25 | 20 | 299 | 337 | 366 | 384 | … | 433 |

**Table 3**
Representative runtimes for the recurrent neural network proxy model versus a full-size reservoir model used by Jiang et al. [93]. Each row represents the time it takes for the model to complete one Newtonian iteration as part of the optimisation process to maximise the future performance of a geothermal field by varying well controls.

| Number of Control Parameters | Runtime for Neural Network Proxy Model (s) | Runtime for Reservoir Model (s) |
|---|---|---|
| 14 | 0.806 | 15 |
| 42 | 0.864 | 43 |
| 84 | 0.862 | 85 |
| 168 | 0.865 | 169 |

often logged with timestamps (e.g., Table 5). Information stored in these time-series data can be used to evaluate the condition of a system or phenomena over time and are thus valuable inputs in data models that make predictions of future performance or occurrence of certain events [106]. However, more than 65 % of the studies that used data logged by sensors in the selected literature did not use the time component of the data as part of their models. More often, the raw time-series values are aggregated to obtain a set of parameters measured within a specific time frame and frequency, which are then used to define the state of a system [106]. These regularly spaced temporal values are then passed as inputs into a neural network model to estimate one or more performance metrics of the system for that point in time (e.g. Refs. [75,82–84,99,100]).

The literature has shown that the most common application of time-series data is in the creation of forecasting models. These are models developed to use past values of one or more monitored parameters to predict some future value [66,68,74,76,93,98,102]. In such studies, a subset taken from the past values of the input time series is used as the input vector into a neural network model to predict one- or multiple-time steps into the future. For example, in the dynamic neural network model developed by Liu et al. [102], the past 11 values of a time series were used to make predictions 1-step and 48-steps ahead (Fig. 5).

### 3.2. Model types and algorithms

Approximately 95 % of the selected articles utilised variations of artificial neural networks for their studies, followed by support vector machines, which appeared in 11 % of the papers (Table 6). Support vector machines, principal component analysis, and several forms of decision trees and ensembles, such as random forest models and gradient-boosted decision trees, were also seen in the rest of the publications. Nine (9) studies in the selected literature trained multiple machine learning models of different types and compared their performance when trained to do the same task. Table 7 lists a comparison of model performance observed in the selected literature that employed multiple model types or algorithms. The succeeding sections describe the different model types in the selected literature and how such models were applied in geothermal operations.

#### 3.2.1. Artificial neural networks

As described by McCulloch & Pitss in 1943 [99], an artificial neuron is a mathematical representation of how a biological neuron processes various excitatory and inhibitory inputs (Fig. 6). The flow of information in an artificial neuron begins with an input vector $\boldsymbol{X} = \{x_1, x_2, \ldots, x_m\}$ being presented to the neuron. Each input is then multiplied by a set of weights $w_m$ and the products are linearly combined before being passed into an activation function $\varphi$ which produces the output $y$. An Artificial Neural Network (ANN) is a network of interconnected single artificial neurons that model experiential knowledge to resemble the learning process of the human brain (Fig. 6b). This artificial neural network can be trained by iteratively adjusting the weights at various points of the network to minimise the difference between network-computed outputs and the target values. Commonly used as a "black box" [107] modelling technique, ANNs can numerically represent various individual components or whole systems without fully knowing the underlying relationships between model inputs and outputs.

In general, the literature has shown that the following steps are necessary to develop a good-performing ANN model.

- Pre-processing input data to remove outliers such as those measured during power plant shutdown [102].
- Selection of an ANN architecture and learning algorithm
- Partitioning of input data into training, validation, and testing datasets
- Model selection based on validation performance
- Model evaluation based on testing on the unseen dataset for testing
- Sensitivity analysis or parametric studies to determine how model inputs contribute to learning the desired outputs

The feed-forward backpropagation neural network (BPNN) was the most common ANN architecture used in the literature. In this approach, the output of a neuron in the $l^{th}$ hidden layer after being presented with the input vector $x^l = \{x_0, x_1, x_2, \ldots, x_m\}$ is:

$$x_j^l = \varphi\left(\sum_{i=1}^{N^{l-1}} w_{ji}^l x_i^l\right), \text{where } l = 0, 1, 2\ldots \tag{1}$$

where the superscript $l$ refers to the current layer, with $l = 0$ being the input layer and $l = L$ referring to the output layer. The subscript index $i$ refers to the output of the neuron from the previous layer $(l-1)$, while the subscript $j$ refers to the index of the neuron in the $l^{\text{th}}$ hidden layer, and $N^{l-1}$ is the number of neurons in the previous layer. After computing the output of the network $x_j^L = y_j$, the error term or delta-rule at the last hidden layer is computed as:

$$\delta_j^L = \varphi'\left(\sum_{i=1}^{N^L} w_{ji}^L x_i^L\right)(d_j - y_j) \tag{2}$$

Where $d_j$ is the desired output for the $j^{th}$ neuron in the output layer and $\varphi'$ is the derivative of the activation function $\varphi$ evaluated for the neurons at the $L^{\text{th}}$ or output layer. Moving backwards in the network, the delta terms for a neuron in the hidden layers are calculated as:

$$\delta_j^l = \varphi'\left(\sum_{i=1}^{N^l} w_{ji}^l x_i^l\right)\left(\sum_{k=1}^{N^{l+1}} \delta_k^{l+1} w_{jk}^{l+1}\right) \tag{3}$$

Finally, the value of the neural weights is adjusted based on the current weight values and the pre-selected learning factor $\eta$:

$$w_{ij}^{l,new} = w_{ij}^{l,old} + \eta \delta_j^l x_j^l \tag{4}$$

Studies exploring different learning algorithms in combination with the BPNN structure were present in the selected publications. Levenberg-Marquardt, Scaled Conjugate Gradient, Pola-Ribiere Conjugate Gradient, and Genetic algorithms have been observed in the literature [47,52,55,71,74,79,83,85,101]. However, variations to the standard BPNN were also used in the literature, such as the dynamic neural network utilised by Liu et al. [102] to train a model using time-series data from a geothermal power plant. A multi-stage BPNN was successfully developed in Ref. [65] to determine optimal designs for a geothermal Kalina cycle power plant based on model-estimated system performance and economic metrics. In conjunction with backpropagation, fuzzy logic was incorporated by Şencan Şahin et al. [84] and Sun et al. [86], resulting in neural networks that performed well when used for exergy and energy analyses of power plant systems.

A recurrent neural network (RNN) is another variation of ANNs that takes advantage of contextual information from past inputs to map sequential inputs to sequential outputs [109]. This type of neural

**Table 4**
A subset of references from the selected literature using time-series data logged by sensors installed across critical points and components of geothermal energy systems.

| Ref | Coefficient of performance | Electric current | Energy efficiency | Energy rate | Enthalpy | Exergy efficiency | Exergy rate | Flow rate | Heating capacity | Operation mode | Power | Pressure | Pump speed | Relative humidity | Specific steam consumption | Temperature | Thermal efficiency | Torque | Valve opening | Sampling Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [66] | – | – | – | – | – | – | – | ✓ | – | – | – | ✓ | ✓ | – | – | ✓ | – | – | – | – |
| [70] | – | – | – | – | – | – | – | ✓ | – | – | ✓ | ✓ | ✓ | – | – | ✓ | – | – | – | 1 h |
| [79] | – | – | – | – | – | – | – | – | – | – | ✓ | ✓ | – | – | – | ✓ | – | – | – | – |
| [80] | – | – | – | – | – | – | – | ✓ | – | ✓ | ✓ | ✓ | – | – | – | ✓ | – | – | ✓ | 5 min |
| [81] | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | ✓ | – | – | – | 13 s |
| [82] | – | – | – | – | – | ✓ | – | ✓ | – | – | – | ✓ | – | – | – | ✓ | – | – | – | 1 day |
| [83] | – | – | ✓ | – | – | ✓ | – | ✓ | – | – | ✓ | ✓ | – | – | – | ✓ | – | – | – | 1 week |
| [84] | – | – | – | ✓ | – | – | ✓ | ✓ | – | – | – | – | – | – | – | ✓ | – | – | – | – |
| [85] | – | – | – | – | – | – | – | – | ✓ | – | ✓ | ✓ | – | – | – | ✓ | – | – | – | – |
| [86] | ✓ | – | – | – | – | – | – | – | – | – | – | – | – | – | – | ✓ | – | – | – | 15 min |
| [87] | – | – | – | – | – | – | – | – | – | – | ✓ | ✓ | – | – | ✓ | ✓ | – | – | – | 5 min |
| [88] | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | ✓ | ✓ | – | – | – |
| [89] | – | – | – | – | – | – | – | – | – | – | ✓ | ✓ | – | – | ✓ | ✓ | – | – | – | – |
| [90] | – | – | – | – | – | ✓ | – | ✓ | – | – | ✓ | ✓ | – | – | – | ✓ | – | – | – | 1 h |
| [91] | – | – | – | – | – | – | – | ✓ | – | – | – | ✓ | – | – | – | ✓ | – | – | – | – |
| [92] | – | – | – | – | – | – | – | – | – | – | – | ✓ | – | – | – | ✓ | ✓ | ✓ | – | – |
| [93] | – | – | – | – | ✓ | – | – | – | – | – | – | ✓ | – | – | – | – | – | – | – | 1 week |
| [94] | – | – | – | – | – | – | – | ✓ | – | – | ✓ | ✓ | – | ✓ | – | ✓ | – | – | – | 1 h |
| [95] | – | – | – | – | – | – | – | ✓ | – | – | ✓ | – | – | ✓ | – | ✓ | – | – | – | 1 min |
| [96] | – | – | – | – | – | – | – | – | – | – | ✓ | ✓ | – | – | – | ✓ | – | – | – | 1 h |
| [97] | – | ✓ | – | – | – | – | – | – | – | – | – | ✓ | – | – | – | ✓ | – | – | – | – |
| [98] | – | – | – | – | ✓ | – | – | ✓ | – | – | ✓ | ✓ | – | – | – | ✓ | – | – | – | – |
| [99] | – | – | – | – | – | ✓ | – | ✓ | – | – | – | ✓ | – | – | – | ✓ | – | – | – | 1 h |
| [100] | – | – | ✓ | ✓ | – | – | – | ✓ | – | – | – | ✓ | – | – | – | ✓ | – | – | – | 1 week |
| [101] | – | – | – | – | – | ✓ | – | – | – | – | – | – | – | – | – | ✓ | – | – | – | 1 week |
| [102] | – | – | – | – | – | – | – | ✓ | – | – | – | ✓ | ✓ | – | – | ✓ | – | – | – | – |

**Table 5**
Exemplary time-series data used by Şencan Şahin & Yazici [84] to train neural network models to estimate the energy rates of a geothermal heating system.

| Year | Month | Day | Mass Flow rate, kg/s | Temperature, °C | Energy, kW | ANN-predicted Energy, kW |
|---|---|---|---|---|---|---|
| 2006 | 12 | 17 | 124.7 | 93.7 | 48,968.06 | 48,966.45 |
| 2006 | 12 | 18 | 127.0 | 93.7 | 49,873.27 | 49,872.55 |
| 2007 | 1 | 22 | 138.6 | 93 | 54,012.59 | 54,010.45 |
| 2007 | 1 | 23 | 138.5 | 93 | 53,969.30 | 53,968.20 |
| 2007 | 2 | 22 | 136.3 | 93.4 | 53,343.72 | 53,342.25 |

network is beneficial for time-series data, as presented in the studies by Jiang et al. [66,68,93] (Fig. 5). Similarly, a Convolution neural network (CNN) is a more complex architecture of ANN that was applied by Jiang et al. [66,93] to train models that can reliably estimate the short- and long-term performance of a geothermal resource.

The selected literature has shown that ANNs have been used to develop data-driven models to represent individual parts of the geothermal surface facility, such as heat exchangers [46,76], pipe networks [62], pumps [50], turbines [49,56], separators [44], and individual or group of wells [66–69,73,93]. However, most published work in this field used ANN to model whole systems to condense the complexity of modelling individual highly nonlinear, interconnected components. ANNs have been used to model conventional geothermal power plants [80,87,89,90,97], binary cycle power plants [51–55,63,65,70,71,79,88,91,92,96,102] (Fig. 7), and geothermal-powered heating or cooling systems [45,47,48,58,62,75,81–84,86,94,95,99–101]. Such models of complex systems are then used for condition monitoring, design optimisation, performance optimisation, and fault detection.

### 3.2.2. Support vector machines

In 1992, Boser et al. [110] proposed a new pattern classification method that aims to maximise the margin between the hyperplane separating data classes and the training data. The margin is defined as the smallest distance of an observation to the proposed separating hyperplane from each data class. Support vector machines (SVMs) apply the kernel trick to transform the features into higher dimensions and project the decision boundary to the original dimension. SVMs can be used for classification and regression tasks. Recent advances in computing technology and new application interfaces provided by open-source software such as scikit-learn [33] have made SVM more accessible to researchers.

In the selected list of publications, SVMs are typically included in a list of machine learning algorithms when investigating which method results in better predictive performance. Zulkarnain et al. [97] compared the performance of SVM models for fault detection in geothermal power plants against ANNs and tree-based models. Similarly, Dong et al. [53] benchmarked the performance of SVMs against ANNs in predicting the performance of an ORC system based on measured system parameters. Santamaría-Bonfil et al. [60] compared the performance of SVMs, decision trees, and other machine learning algorithms in imputing missing values for a global geochemical database. Yan et al. [51] showed that SVMs performed better in estimating the thermal efficiency and net power output of an ORC system compared to neural networks and linear regression models (Fig. 8).

### 3.2.3. Linear and nonlinear regression

In linear regression, a model is used to fit the data while varying the coefficients of a linear model to minimise the residual sum of squares between the actual data and the values predicted by the model [111,112]. This model approximates the relationship between a dependent variable $y$, one or more independent variables $x_1, .., x_j$, and the bias $w_0$. A geometric representation of simple linear regression is shown in Fig. 9, while the general multiple linear regression equation is given as:

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_j x_j \quad (5)$$

**Table 6**
Main machine learning model types and algorithms used in the selected above-ground geothermal publications.

| Subsection | Model types and algorithms | References | Publication Count |
|---|---|---|---|
| 3.2.1 | Artificial Neural Networks | [40–59,62–77,79–102] | 60 |
| 3.2.2 | Support Vector Machines | [51,53,60,69,94,96,97] | 7 |
| 3.2.3 | Linear and Nonlinear Regression | [45,51,60,69,78,95,96,98] | 8 |
| 3.2.4 | Decision Trees and Ensembles | [60,61,69,94,97] | 5 |

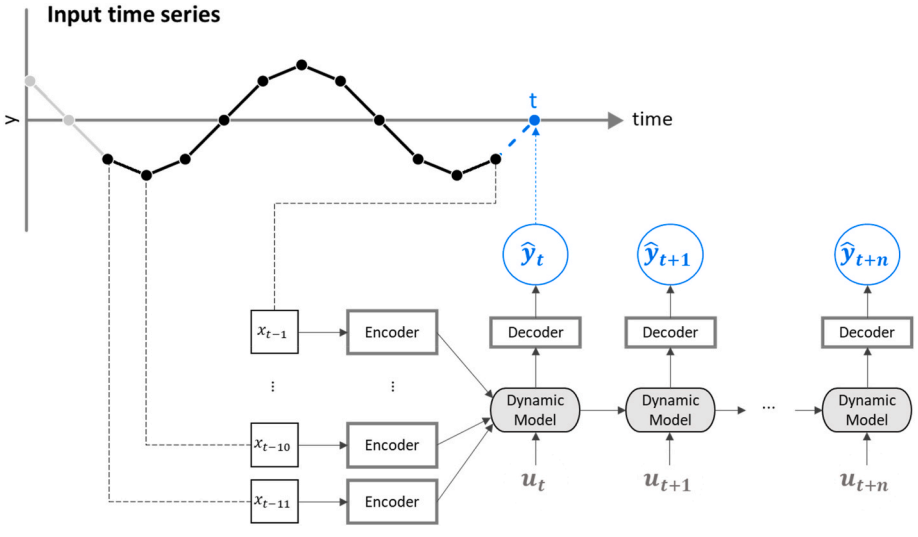



**Fig. 5.** RNN structure showing naïve feature engineering wherein model inputs **x** are the past eleven (11) time-series values from the input feature vector to make predictions $\hat{\mathbf{y}}$ at **n** time steps ahead (modified after [102]). The model also accounts for control variables **u** that may operationally affect the values of the model inputs and outputs.

**Table 7**
Summary of model performance comparison between different model types in the selected literature. Values presented in the table are relative to the best-performing model with a value of 0 % means no difference with the best model.

| Reference | Evaluation metric | Neural Networks | Support Vector Machines | Linear Regression | Decision Trees | Random Forest |
|---|---|---|---|---|---|---|
| Dong et al. [53] | RMSE | 0 % | 21 % | – | – | – |
| Muchamad et al. [69] | MSE | 0 % | 463 % | 13 % | 80 % | 19 % |
| Park et al. [95] | CVRMSE (%) | 0 % | – | 103 % | – | – |
| Santamaria-Bonfil et al. [60] | MEI | – | 9 % | 36 % | 0 % | 17 % |
| Wibowo et al. [96] | RMSE | 0 % | 48 % | 145 % | – | – |
| Xu et al. [45][ | MAPE (%) | 0 % | – | 275 % | – | – |
| Yan et al. [51] | RMSE | 75 % | 0 % | 165 % | – | – |
| Yan et al. [94] | MAPE (%) | 0 % | 14 % | – | 31 % | – |
| Zulkarnain et al. [97] | ROCAUC[a] | 0 % | −1% | – | −25 % | – |

**MAPE**: Mean Absolute Percentage Error; **CVRMSE**: Coefficient of the Variation of the Root Mean Square Error; **RMSE**: Root Mean Square Error; **ROCAUC**: Area Under the Receiver Operating Characteristic Curve; **MSE**: Mean Squared Error; **MEI**: Minimum Equivalence Interval.
[a] Best performance corresponds to the model with the highest ROCAUC.

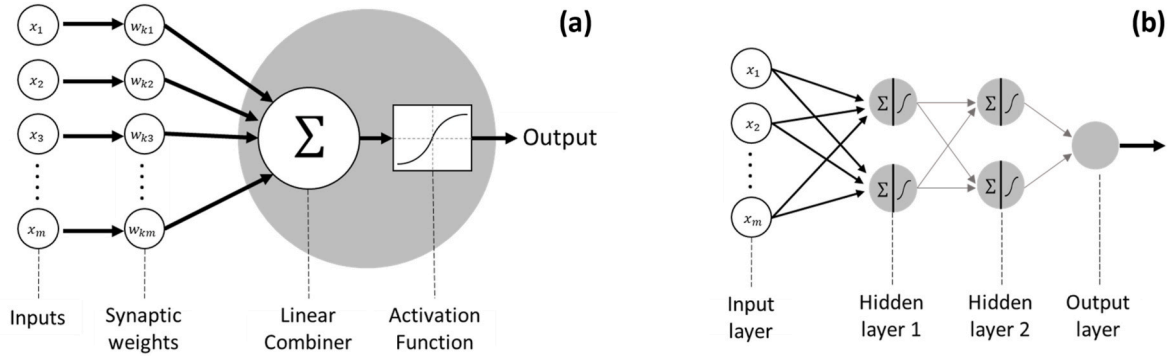



**Fig. 6.** (A) A high-level diagram of a single artificial neuron after Kalogirou [108] and (b) an example of a neural network with two hidden layers.

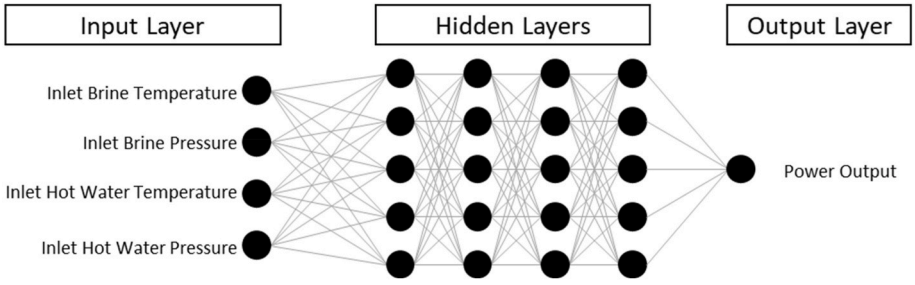



**Fig. 7.** *Multi-layer backpropagation neural network developed by Wibowo* et al. *to predict the power output of a geothermal-powered ORC power plant based on measured system parameters (adapted from* Ref. [96]*).*

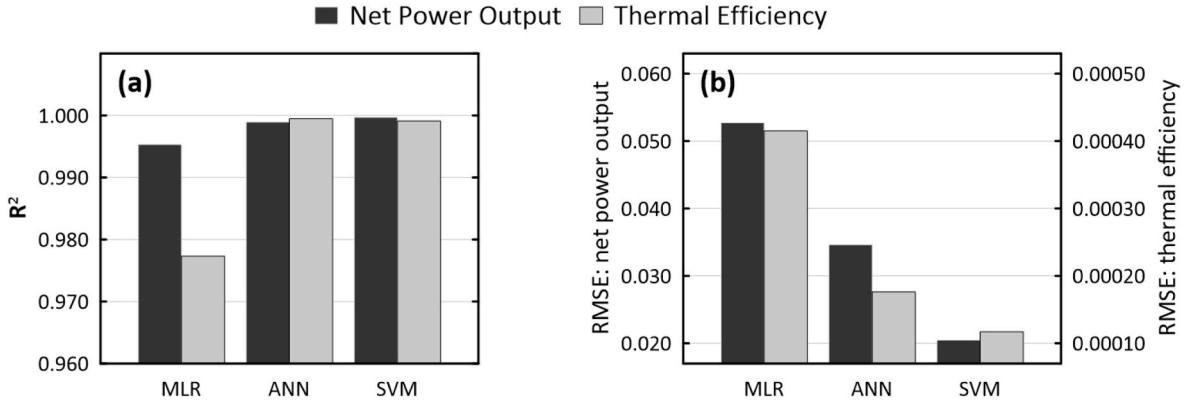



**Fig. 8.** Comparison of (a) coefficients of determination and (b) root mean squared errors for the models trained by Yan et al. [51].

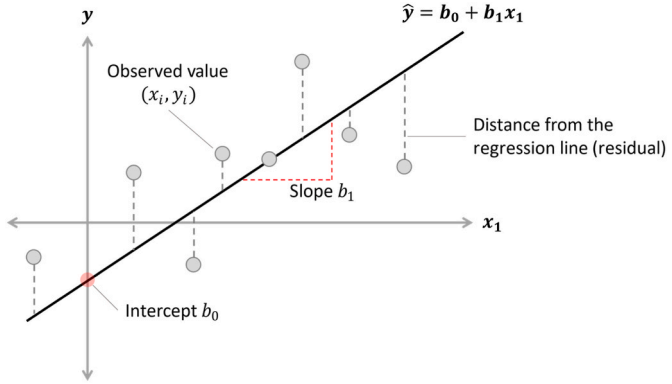



**Fig. 9.** Geometric representation of the simple linear regression equation showing the samples (dots) and the regression line. This is a special case of Eq. (5) when $j = 1$.

where $\hat{y}$ is the estimated value of the dependent variable and the constants $w_0, w_1, \ldots w_p$ can be estimated using information from observed values of $y$ and $x_1, \ldots, x_j$.

Multivariate linear regression (MLR) has been used in the selected literature to predict the performance of whole power plant systems and compare its predictive performance against other algorithms. Park et al. [95] compared the predictive performance of MLR and ANN in modelling the performance of a geothermal-powered heating system using parameters measured by sensors deployed at critical points of the system. Xu et al. [45] trained linear (Eq. (6)) (Table 8) and nonlinear regression models to represent another geothermal-powered heating system using measured pressures and temperatures. The authors then compared its performance to an ANN representation of the system (Fig. 10).

$$Q = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6 \quad (6)$$

Other studies in the selected literature fall outside the previous group of publications that used the standard linear and nonlinear models for regression. In a novel approach, Jin et al. [78] used a special form of regression analysis called STRidge or sequential threshold ridge regression to learn the governing equations of heat conduction and conductive-convective heat transfer. Siratovich et al. [98] have used linear regression models to represent some of the simpler components in a geothermal power plant as part of their proposed framework for constructing digital twins of whole power plant systems.

### 3.2.4. Decision trees and ensembles

Tree-based models learn from the data using a set of if-then-else decision rules to predict the value of a target and can be used for regression and classification tasks [113]. Decision or classification trees consist of nodes acting as "roots" or "leaves." A root node tests the incoming data based on a discrete function of input attribute values and partitions the input space based on a criterion. A leaf node, on the other hand, does not have outgoing edges but stands as terminal points of the tree and assigns the partitioned input space to the appropriate target value or assigns a likelihood of that target attribute having a particular value. Most tree-based models are easy to interpret and can be visualised to better represent the rules learned by the algorithm [114]. Hyperparameters such as the depth of the tree or the minimum number of samples for each node can be tuned to avoid the potential pitfalls of this method (e.g., overfitting). An example decision tree utilised in the selected literature is presented in Fig. 11. Although decision trees are easier to understand relative to the black-box neural network models and the mathematical equations of linear regression models, the main drawback is the tendency of overfitting, especially when using very deep decision trees [113].

An ensemble of decision tree models can be used to reduce the risk of overfitting by training more generalizable models that cope better with unseen data. Random forest involves the use of several shallow decision trees trained on random subsets of the training data, making each tree slightly different from one another [113]. The final prediction of the whole random forest model is then an aggregation of the individual predictions of each decision tree. Another ensemble algorithm using decision trees is the gradient boosting machine wherein training data that are harder to "learn" are given a higher weight or importance. An iterative process of creating shallow decision trees to match the training data with different importances assigned to each input lead to robust predictions by gradient-boosted decision trees but can also be highly influenced by outliers in the training data [113].

In 2019, Zulkarnain et al. [97] used gradient-boosted decision trees to develop a fault detection system for a geothermal power plant based on the measured parameters from its water-cooling system. The classification performance of the model to detect whether the system is in a normal or an abnormal state was compared with other algorithms. Although the gradient-boosted decision tree model performed well during the training phase, it showed apparent overfitting after performing poorly during the validation stage compared with the other models. Similarly, Castellanos et al. [61] constructed a fault identification framework for an electrical submersible pump (ESP) system based on a chain of simple decision trees. The model was trained on data gathered from experiments using a ten-stage ESP to simulate, monitor, and label the expected failure modes. The authors concluded that this decision tree structure is suitable for practical fault detection and diagnosis applications based on its excellent performance in detecting and classifying the faults in the pump system (Fig. 11).

An innovative method to fill in missing values in a global compositional geochemical database was successfully applied by Santamaria-Bonfil et al. [60] using decision trees and random forest models, among other machine learning algorithms.

## 3.3. Feature selection and interpretability

The feature selection task is essential in creating parsimonious, equally accurate and understandable models. Primarily, feature selection aims to reduce the number of model parameters to the variables that have the most influence on model performance [115]. The value of this step is seen not only in dimensionality reduction, thereby lowering computation costs, but also in improving the interpretability of the resulting models [94].

Of the 63 studies included in this review, only eight (8) applied

**Table 8**
Example regression coefficients for the best linear regression model (Eq. (6)) to predict the ground source heat pump energy rates based on the system design parameters from various experiments (after Xu et al. [45]).

| Feature variable | | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|---|
| **Feature description** | Intercept (Constant) | Soil thermal conductivity [W/m°C] | Vertical well depth [m] | Well diameter [mm] | U-tube thickness [mm] | Water flow rate [$mm^3/h$] | Water temperature difference [°C] |
| **Coefficient variable** | $w_0$ | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ |
| **Coefficient value** | 35.245 | −0.463 | −0.219 | 0.016 | −4.613 | 20.550 | 6.960 |

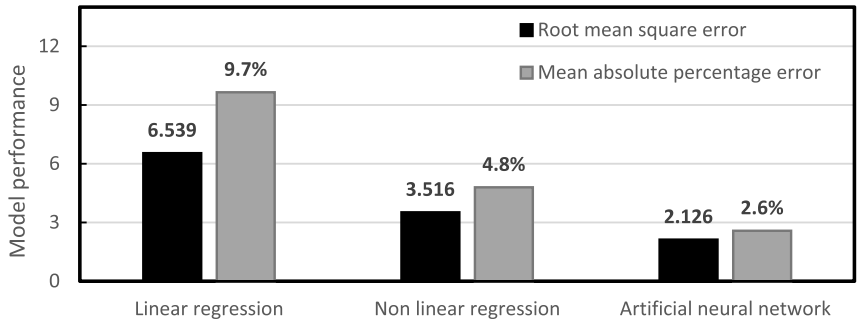



**Fig. 10.** Comparison of model performance in predicting heat transfer rate from a geothermal heating system (after Xu et al. [45]).

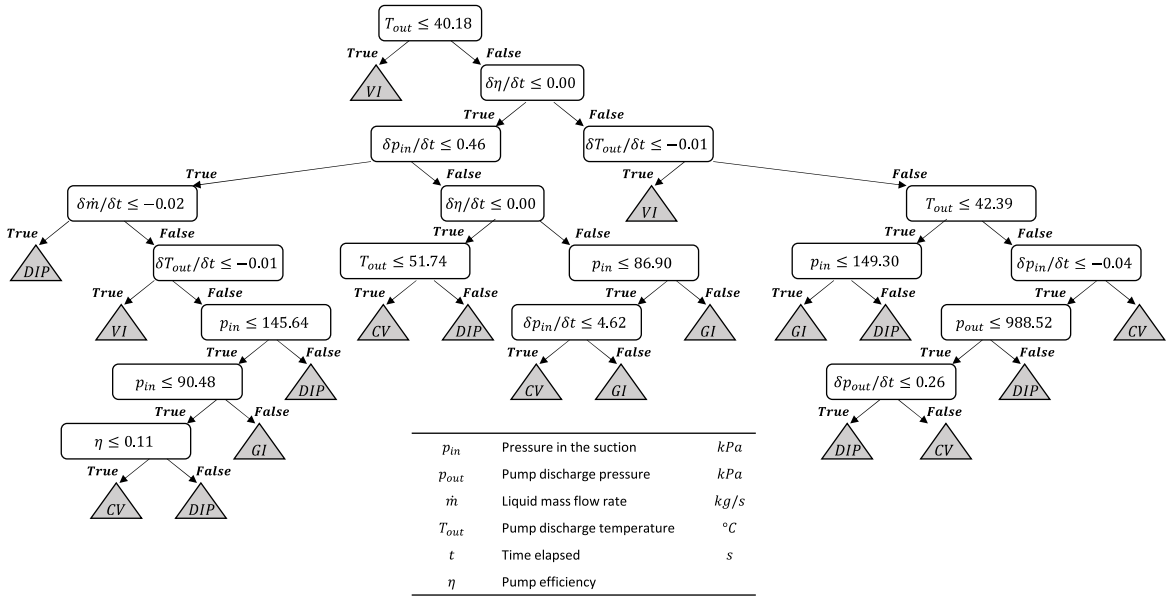



**Fig. 11.** Resulting decision tree for classifying the fault type of a submersible pump (after Castellanos et al. [61]). The decision tree was designed to detect the following premature failure faults in the pump: choke valve closure (CV), decreasing input pressure (DIP), gas content increase (GI), and fluid viscosity increase (VI).

feature selection methodologies in their approaches. The rest of the publications relied on domain expert knowledge to select the input parameters for their models. Studies by Yan et al. [94] and Park et al. [95] utilised the relationship of the input parameters with the target variables, as expressed by their Pearson correlation coefficients, to determine the parameter subset with the most impact on the target variable

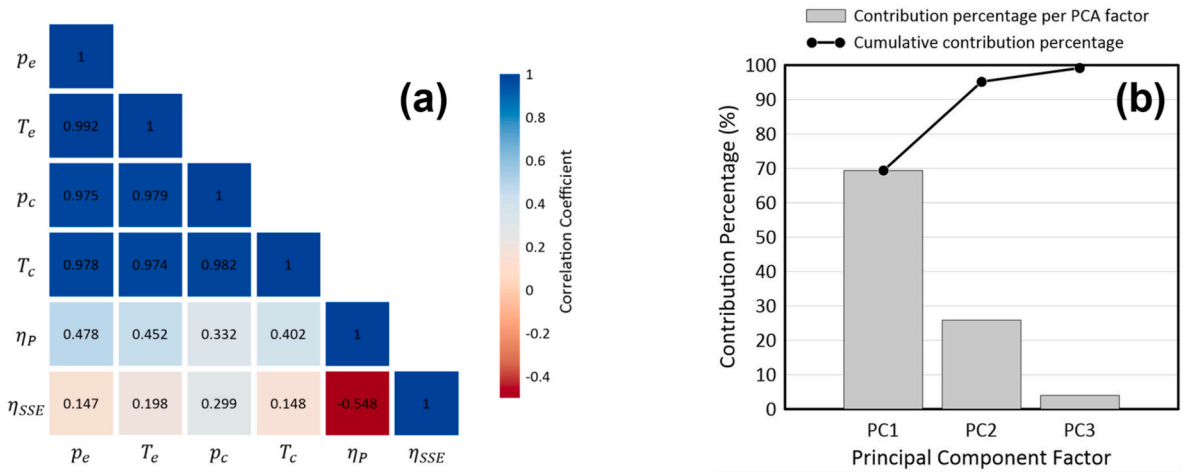



**Fig. 12.** (a) Correlation coefficients between system parameters used as model inputs for predicting net power output and thermal efficiency of an ORC system. (b) Contribution percentages of the principal component factors show that the first two principal components cover 95.21 % of the variance in the data. A key parameter subset was identified among the correlated system parameters based on the best-performing models built using the principal components [51].

(Fig. 12a). It was seen in works by Park et al. [95], Yan et al. [51], and Ping et al. [92] that principal component analysis (PCA) can be effectively applied to reduce the feature space and improve model interpretability (Fig. 12b). Taking it a step further, PCA was augmented by including partial mutual information to determine the degree of interdependence between input parameters in the studies by Ping et al. [91,92]. Sensitivity and parametric studies were used in the literature to determine how input parameters affected the model outcomes and, consequently, determine which had the most influence on performance [44,63,81].

### 3.4. Primary modelling tasks and selected use cases

As part of adopting a relatively new type of technology in the geothermal energy industry, it is of great interest to note at which stage of the geothermal project development life cycle [9] the applications are used and what type of challenges are being tackled using such methods. The selected literature has shown that data science techniques have been applied in the pre-development and operational phase of geothermal above-ground facilities for the purpose of design optimisation (Sec 3.2.1), condition monitoring (Sec 3.2.2), performance optimisation (Sec 3.2.3), and fault detection (Sec 3.2.4). There are existing works in the other stages of geothermal project development, but these are mostly related to characterising and evaluating the subsurface resource [28]. Fig. 13 shows a Venn diagram depicting the distribution of published work among the main categories of use cases of data science or analytics observed in the selected literature.

#### 3.4.1. Design optimisation

There were eleven (11) publications included in this review that trained neural networks and other machine learning models to create surrogate models used for design optimisation [40–45,62–65,79]. The first group of papers focused on using the developed models to identify the optimal set of system design parameters that would yield the highest system outputs and efficiencies. Another group of publications took it further by including economic metrics such as payback period, levelized cost of electricity, and investment cost in the design evaluation and optimisation process.

Using design and performance data from existing publications, ANNs can be trained to estimate the energy and exergy efficiency [63], net power output [41], and other performance metrics [44,45] of an individual component or a whole power plant system design. These models are further used to evaluate variations in plant design decisions, such as using recuperators in an ORC plant [63]. Several studies have been devoted to identifying the optimal working fluid [41,42,64] for a binary plant. Some notable research in this group of publications includes works by Huster et al. [41] showing that by using surrogate ANNs, a working fluid mixture of isobutane and isopentane was identified that can allow a geothermal-powered ORC to generate 17 % more power compared to purely using isobutane. Similarly, Peng et al. [42,64] conducted a study that included the use of group contribution methods to extend the application of their ANN to working fluids that have known molecular structures but whose thermodynamic properties are unknown (Fig. 14).

The selected literature has shown that data models are effective surrogates for thermodynamic process models when evaluating the economics of a particular system design. For example, Kayfeci et al. [62] performed a life cycle cost analysis using ANN proxy models to determine the financially optimal pipe insulation specifications for district heating systems anywhere in the world. Arslan and Yetik [79] applied a similar economic evaluation process using ANNs to arrive at an optimal design for a 64 MW ORC plant with a potential economic benefit of USD $125 M for the existing wells in the Simav geothermal system in western Anatolia. For the same Simav geothermal system, Senturk Acar [65] developed and applied ANNs in multiple stages to determine an optimal 25 MW Kalina cycle power plant with a net present value of $113 M (Fig. 15). In two studies by Huster et al. [40,43], it was shown that a structured framework to apply ANNs as thermodynamic surrogate models in conjunction with an optimisation tool to determine the optimal working fluid of an ORC plant to maximise the plant output, levelized cost of electricity, investment cost, and break-even period.

#### 3.4.2. Condition monitoring

The bulk of the publications included in this review, which is 34 out of 63 papers, relates to the training and application of ANNs and other machine learning models to estimate the current performance of an energy system based on sensor-measured data. The models have been used in a range of geothermal-powered energy systems, such as district heating systems, binary cycle plants, and conventional steam turbines. This particular use of data models can be a valuable tool for operators to flag potential degradation of the system, which could allow a more proactive approach when conducting maintenance activities.

One of the energy systems that had early applications of ANNs are geothermal heating systems. These are complex, non-linear systems with seasonal variations that have been studied using analytical and numerical approaches [82]. The Afyon geothermal district heating system (AGDHS) in Turkey is one of the most studied systems within the selected literature, with five (5) publications related to it. Average weekly data from the 2006–2010 heating season was used by Keçebaş and Yabanova [83] and Keçebaş et al. [82] to train ANNs to predict the exergy and energy efficiency of the system based on the ambient temperature and the pressures, temperatures, and flow rates of the working fluid at pre-defined locations within the system. An attempt to improve these models by applying fuzzy logic together with neural networks was conducted by Şencan Şahin and Yazici [84] but resulted in the standard ANN model outperforming the model with fuzzy logic. A financial model was then developed by Keçebaş et al. [100] using the ANNs to show that the current AGDHS had a present worth factor (PWF) of 1.43, significantly below the 7.9 PWF value required for the system to be profitable. The trained ANN models were then utilised by Keçebaş and Yabanova [99] to show that the implementation of an automated control strategy using proportional integral derivative (PID) controllers could allow the AGDHS to increase the overall heat production by 13 % without the need for further investments (Fig. 16). Other studies that applied similar

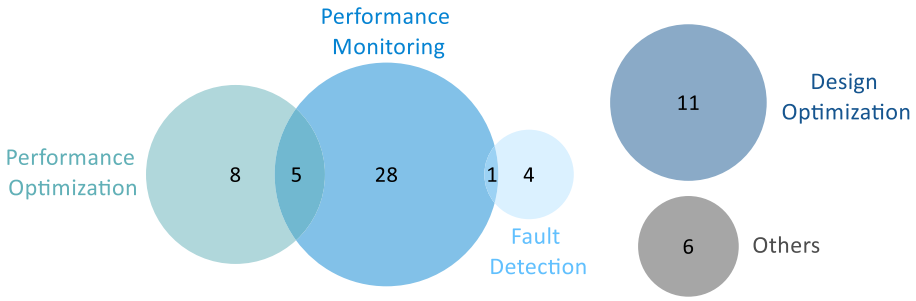



**Fig. 13.** Venn diagram showing the distribution of the publications according to the primary use case of data science or analytics mentioned in the research papers.

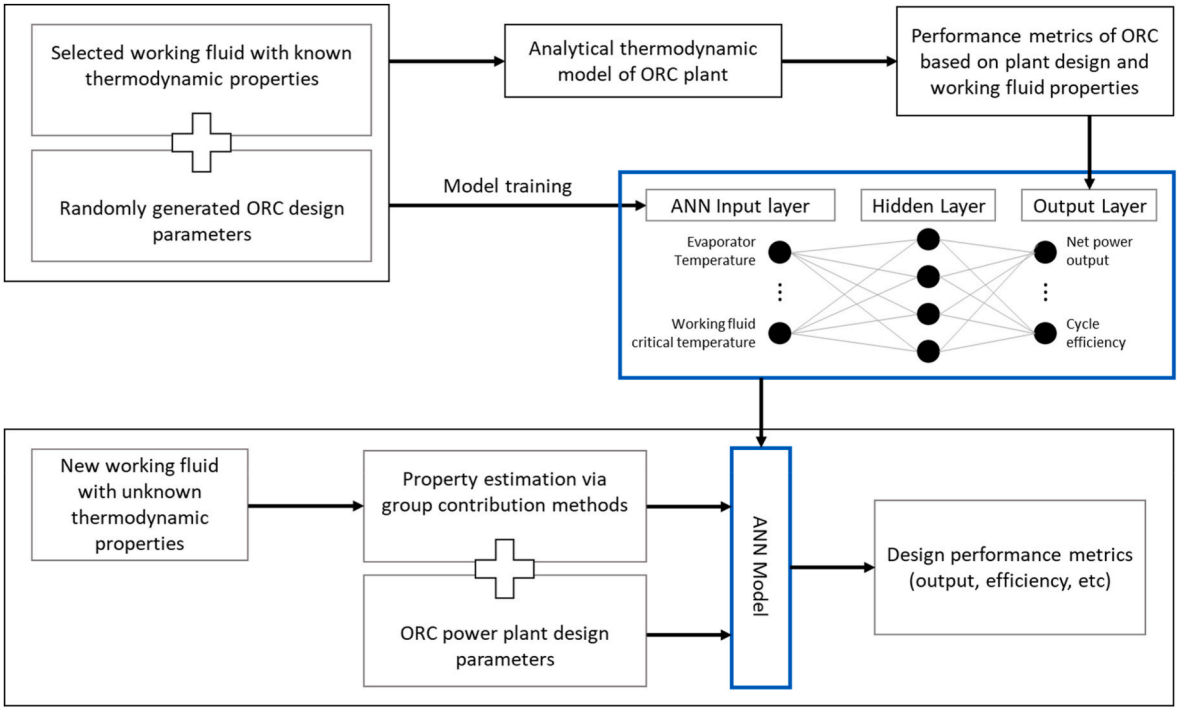



**Fig. 14.** Schematic diagram of the process developed by Peng et al. [42] to combine ANNs with group contribution methods to estimate the performance of an ORC design with working fluids of unknown thermodynamic properties.

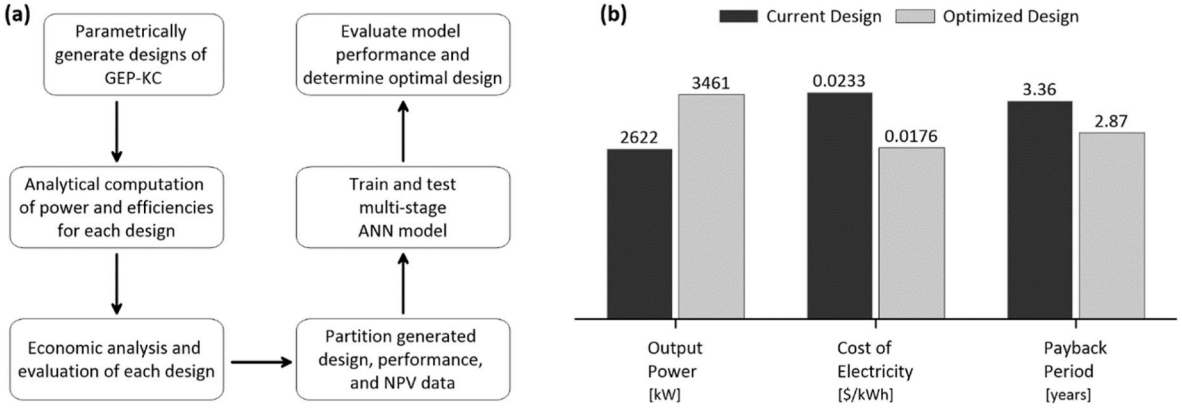



**Fig. 15.** (a) Development and testing of multi-stage ANN for optimising a generic geothermal energy powered Kalina cycle (GEP-KC) after Senturk Acar [65] and (b) comparison of economic metrics between existing and optimised design for the Afyon geothermal binary power plant in Turkey [75].

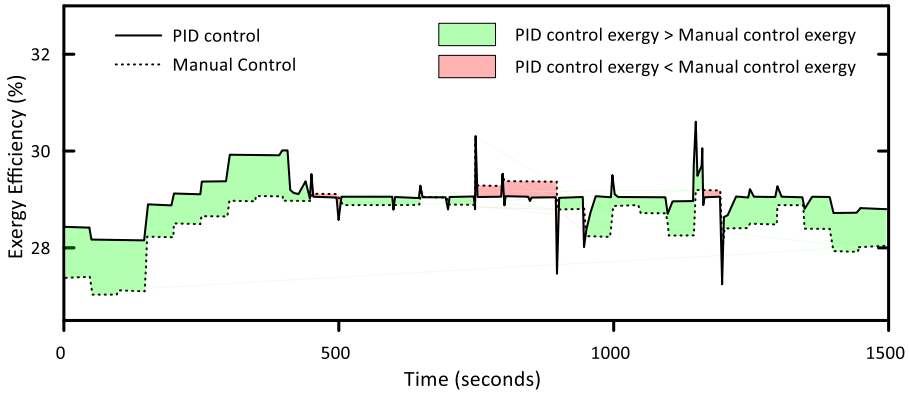



**Fig. 16.** Comparison of geothermal district heating system exergy efficiency estimates by ANN model based on different control strategies (modified after Keçebaş et al. [99]).

methods with data from different geothermal heating systems across the world can be seen in Refs. [48,85,86,94,95].

Organic Rankine Cycle binary power plants generate electricity from the excess heat stored in separated brine or geothermal fluid extracted from low-to medium-temperature reservoirs [116]. In this review, nine (9) looked at the application of ANNs and other ML models to assess the performance of ORC systems and the selection of optimal working fluids for such systems. A study by Wibowo et al. [96] showed positive results in using ANNs, support vector machines, and ridge regression to predict the output of a geothermal ORC plant in Indonesia based on measured system temperatures and pressures (Fig. 7). The ANNs developed by Wibowo et al. [96] achieved less than 10 % root mean square and mean absolute errors on the validation dataset, indicating that the models were not overfitted and were robust enough to be applied on unseen data. A similar approach was taken by Ling et al. [70] to develop an ANN model as a proxy for an ORC plant operated by Cyrq Energy Inc. The ANN performed well when first trained and tested using data generated from a process model but displayed subpar performance when using real data from the field, as seen by the doubling of the errors from 0.03 using the synthetic dataset to 0.069 using the field data. Yilmaz et al. [88] developed an accurate ANN model for ORC plants with an internal heat exchanger or recuperator. The conducted research showed that such models are accurate enough for operational use and forecasting of plant performance (Fig. 17). The rest of the studies in this group used data from experimental test rigs to identify critical system parameters that can be used to develop ANNs that can accurately predict the performance metrics of ORC systems [46,49,51–53,91].

Compared to geothermal heating and ORC systems, fewer publications looked at machine learning applications for conventional geothermal power plants, as only six (6) publications were found in the search process. Ruliandi [87] and Ruliandi & Priyangga [89] developed ANN models to estimate the steam consumption coefficient (SSC) of the Unit 4 Kamojang geothermal power plant in Indonesia using temperature and pressure measurements from sensors scattered across the facility. The studies showed that excluding data gathered during plant commissioning improved the performance of the models, which is most likely due to the drastically different operating conditions required during commissioning tests versus regular operation. The models were able to replicate the expected increasing trend of the SSC over time due to natural wear and tear of the plant components and can be used to indicate when the plant should be taken out for preventive maintenance [87]. In another paper, Ruliandi et al. [89] trained several ANNs on the same input data but having different outputs representing the exergy efficiency of individual components and that of the overall power plant. The models showed good performance when predicting the overall plant exergy efficiency, with MAPE of 0.87 %, but exhibited systematic underestimation in the prediction of the exergy efficiency of the plant (Fig. 18). The models could have performed better when predicting the individual efficiencies of critical plant components, with average errors of 12.5 %. This concept of training component-specific models was applied by Siratovich et al. [98] to develop a framework that would allow the creation of a digital twin for any geothermal power plant by connecting the individual component models using first principal thermodynamics. In this framework, each component will be represented by a different machine learning model with the type of model depending on the complexity of the expected input and output parameters. Looking at component-specific models, Zulkarnain et al. [97] were able to use K-means clustering to label monitoring data for the cooling system of a geothermal power plant and train several machine learning models to predict whether the component is working in normal or abnormal conditions (see Fig. 19).

The rest of the studies on condition monitoring focused on the overall performance of the geothermal wells producing and injecting the geothermal fluids. In multiple works, Jiang et al. [66,68,93] explored the development and application of RNNs to estimate the short- and long-term performance of geothermal wells. The authors of the said studies highlighted the challenge of making predictions outside the bounds of the training data, which is expected in geothermal operations where the resource may be degrading over time and will eventually operate outside of known initial conditions. Still, the studies showed that using multi-scale recurrent neural networks makes it possible to make reliable short- and long-term predictions, provided that the time series is relatively stationary. Although geothermal wells are mostly considered part of the subsurface portion of geothermal operations, their performance can greatly affect surface operations and are vital to

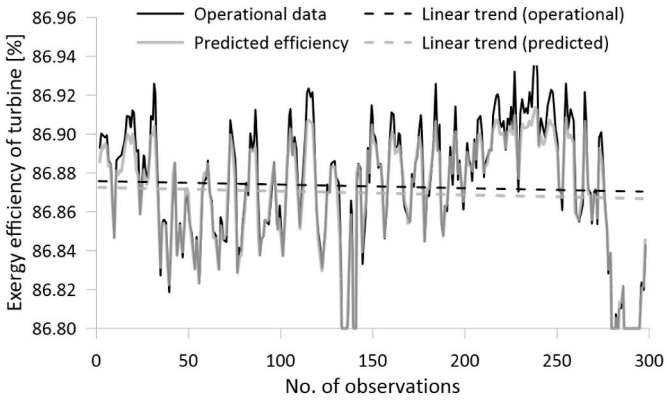



**Fig. 18.** Model-predicted turbine efficiencies during the training phase showing systematic underestimation by the ANN (modified after Ruliandi et al. [90]).

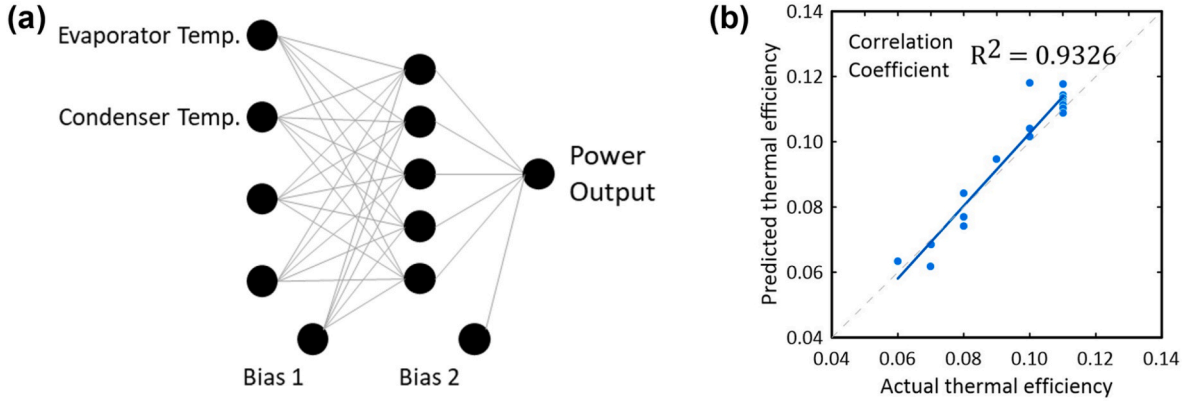



**Fig. 17.** (a) ANN structure showing model inputs to predict the output of an ORC plant with an internal heat exchanger and (b) modelling results showing good validation accuracy (modified after Yilmaz et al. [88]).

ensuring the safe and efficient operations of surface facilities. This was the primary motivation in the works by Harry et al. [67,69] where an investigation was conducted to use machine learning algorithms to study the decline rates in geothermal wells operating under different scenarios. The authors were able to develop accurate feed-forward (FFNN) and general regressor neural network (GRNN) models with errors of less than 5 % for the best model while highlighting the potential bias posed by overfitted models (Table 9).

### 3.4.3. Performance optimisation

Aside from monitoring the condition and performance of individual components and whole geothermal systems, surrogate machine-learning models have been developed to optimise or enhance existing operations. Previously, Siratovich et al. [98] showcased their proposed framework for modelling above-ground geothermal systems. Moreover, the authors in that work took this a step further by applying their framework in a case study to show how reinforcement learning algorithms can be used to increase generation in a geothermal power plant. The autonomous agent was allowed to change the operating conditions of the production wells and pipeline pressures, and a corresponding reward or penalty was given to the agent depending on the model-predicted performance of the plant. The model-controlled operation resulted in a hypothetical increase in generation by 1.8 MW when compared against a baseline forecast using existing operating conditions (see Fig. 19).

The data-driven models were trained for parametric optimisation to find potential operational adjustments that can be implemented to improve the performance metrics of existing ORC systems [52,54,55,70]. The same process of using proxy models has been applied to geothermal-powered heating systems [75,99–101]. The operational enhancement of individual components has been shown to work using trained proxy models, such as for turbines [72] and geothermal wells [73].

### 3.4.4. Fault detection

The following subset of published works tackled challenges in detecting faulty operations in various parts of geothermal systems. The ANNs developed by Lalot and Pálsson [76] were primarily used to detect fouling in heat exchangers based on forecasted temperature difference values across the heat exchanger wall. The models performed the task successfully, and the authors recommended transitioning to a predictive maintenance scheme for the heat exchangers instead of the existing periodic, fixed maintenance intervals. Liu et al. [102] similarly used ANN trained on past values of various operational parameters monitored by sensors to make single- and multiple-step forecasts and anticipate abnormal behaviour across system. A different approach was taken by Rodriguez et al. [56] in a regression task to estimate the useful life of turbine blades by training ANNs on data obtained from finite element simulations of the component. Zulkarnain et al. [97] compared the performance of several machine learning models in processing multiple input streams to detect faulty operations in the water-cooling system of a geothermal power plant in Indonesia (Fig. 20a). The automatic detection of abnormal conditions improves the existing manual fault detection process, which proved challenging when simultaneously looking at multiple operating parameters. Castellanos et al. [61] used a chain of decision trees to detect and classify faulty operations in a submersible pump system operated in a two-phase fluid environment. The decision to use multiple decision trees instead of a single one decreased the overall accuracy of the model but improved performance in classifying the fault types when abnormal operation is detected (Fig. 20b).

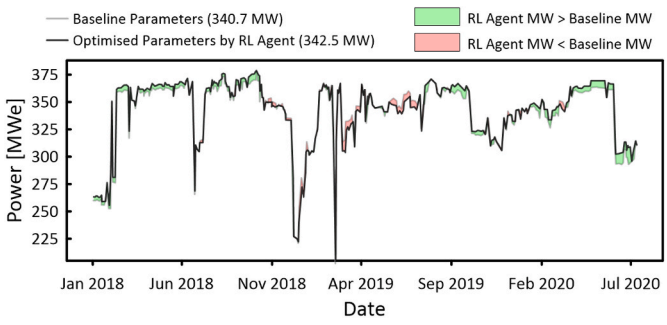



**Fig. 19.** Comparison of the baseline forecast model and the reinforcement learning (RL) agent optimisation results (after Siratovich et al. [98].

## 3.5. Modelling best practices

The publications included in this review has clearly shown the potential benefits of developing accurate and reliable data-driven models for above-ground geothermal operations. Still, the existing models and machine learning pipelines could greatly benefit from a renewed focus in systematic time-series feature engineering and selection, cross validation, and data rebalancing.

### 3.5.1. Systematic time-series feature engineering and selection

Only a few studies included in the review (e.g. Refs. [66,68,76,81,82,84,93]) applied algorithms and methods specifically tailored for time-series data. Although it is possible to accurately perform tasks such as time-series regression or classification based on the raw data values (naïve time-series feature engineering), there are instances where a feature-based representation of a time series can better display the unique characteristics of the dynamics contained in the data [117]. Thus, pursuing further research in applying a feature-based approach to time-series analysis of above-ground geothermal facility data is of significant merit.

Rigorous feature engineering and selection is another aspect that can be a focus of further research using operational geothermal data. Most of the selected literature took advantage of the ability of ANNs to glean intrinsic relationships in large datasets, with the majority (55 out of 63) of the studies relying on domain expertise to select input parameters for the models. A systematic and replicable feature selection methodology can greatly improve model performance and address model interpretability even while using “black-box” type models such as ANNs and complex tree-based models [112].

### 3.5.2. Cross-validation

Only six (6) publications included in this review indicated the use of cross-validation techniques as part of the model development and evaluation process [60,61,73,80,83,95]. The process of cross-validation is essential to ensuring that the trained models are not overfitted and

**Table 9**
Performance metrics of neural network models developed by Harry et al. [69] showing training and prediction errors for the best models relative to values taken from wellbore model simulations.

| Scenario | Training MAPE (%) | | Decline Rate Predictions (%) | | | Prediction MAPE (%) | |
|---|---|---|---|---|---|---|---|
| | FFNN | GRNN | Wellbore Model | FFNN | GRNN | FFNN | GRNN |
| Normal operation, no issues | 1.6 | 1.6 | 0 | 0 | 0 | – | – |
| Increasing reservoir pressure | 1.2 | 1.4 | 13.9 | 13.9 | 4.3 | 0.1 | 69.4 |
| Near-wellbore scaling | 1.0 | 1.2 | 17.7 | 17.5 | 3.8 | 1.4 | 78.8 |
| Wellbore and near-wellbore scaling | 1.1 | 1.1 | 14.8 | 14.1 | 6.8 | 4.4 | 53.8 |

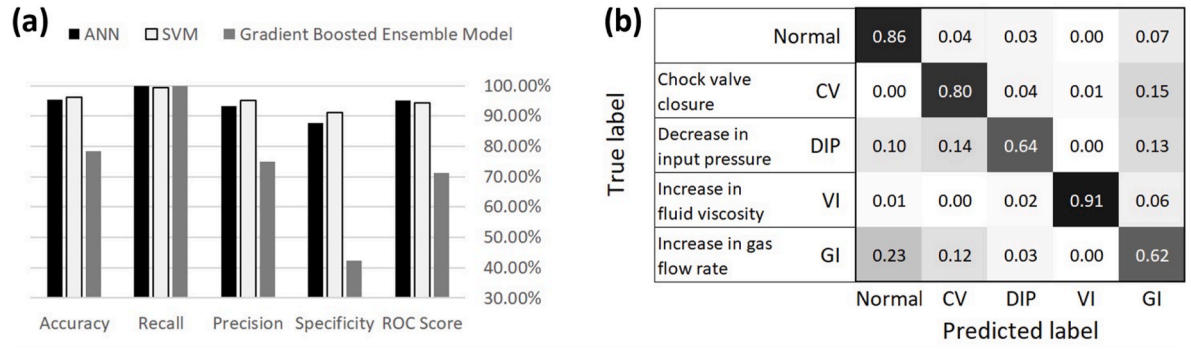



**Fig. 20.** (a) Classification performance of ML models for fault detection of water cooling system in a geothermal power plant (modified after Zulkarnain et al. [97]) and (b) confusion matrix showing the 5-fold cross-validation results for the chain of decision trees used to detect faults in a submersible pump (adapted from Castellanos et al. [61]).

perform robustly on unseen data. In essence, cross-validation is done by grouping the input data into several groups, with one group set aside to evaluate the model performance while the model is trained on the other groups. This process is repeated a certain number of times, with a different data group acting as the testing data for each iteration [113].

Aside from the standard *k-fold* cross-validation technique described previously, there are also other approaches to cross-validation that can be useful in models developed for geothermal applications. Stratified *k-fold* cross-validation is a technique that ensures the distribution of different classes is close to equal among the different folds. This method was successfully applied by Castellanos et al. [61] in their fault-detection study. When working with time series, a rolling cross-validation technique can be applied to assess the performance of the models accurately while honouring the potential temporal correlations in the dataset [118].

### 3.5.3. *Rebalancing of the input data*

Applications such as fault detection and condition monitoring in industrial operations rely on datasets that are inherently imbalanced, with one class representing normal operations and a minority class referring to an anomalous state of the system [119]. In such cases, standard classifier algorithms tend to be biased toward predicting the occurrence of the majority class since many performance metrics are designed to favour models that make the most number of correct predictions. The minority class tend to be ignored by the models and are often misclassified.

Data resampling can be applied to ensure a close to equal distribution of classes in the input data. This process will ensure that standard classifier algorithms and performance metrics can be applied while minimising bias on the original majority class. For example, Dempsey et al. [120] applied random undersampling on seismic monitoring data when developing the models used to forecast eruptive activity at Whakaari in New Zealand. Undersampling aims to arrive at a more balanced input dataset by eliminating samples from the majority class, while oversampling achieves the same objective by replicating samples from the minority class [119].

## 4. Conclusions and scope of future research

This systematic review has shown that significant research has been done on applying data science and analytics to tackle and solve problems related to the operation of geothermal surface facilities. An in-depth review has been conducted to show the primary model design decisions researchers have made to develop AI and ML models and the use cases for such models in above-ground geothermal operations.

The review has shown that the approach used in most applications of the selected literature is the development of surrogate or proxy models to represent the operation of an individual component of a plant (e.g., turbines [49,56,72,98], wells [67–69,73,93], flash plants [44,98], heat exchangers [46,76]) or whole geothermal systems [47,48,52,53,70,81–84,86–89,91,92,94,95,98–100]). Furthermore, most published data analytics applications used the standard backpropagation architecture of artificial neural networks in their models with one or more hidden layers. At the same time, usage of more advanced forms of ANN is found in the literature (e.g., ANN with fuzzy logic [47,84,86], recurrent neural networks [66,68,93], as well as convolutional neural networks [66,93]).

In terms of the source or model input data, only 24 out of the 63 papers included in this review used operational data to train and evaluate the AI/ML models. The remaining 39 papers used published experimental and synthetic data from numerical simulations. The models developed from these datasets were used as proxies for more complex thermodynamic numerical models. The proxy models were utilised in the design stage to inform design decisions, such as selecting the working fluid in binary power plants. Meanwhile, the same proxy models were used in operational plants to detect faulty conditions and optimise system performance and overall project economics.

Although considerable research has been done in applying AI and ML methods in above-ground geothermal applications, this review has shown knowledge areas that can be expanded to further growth in this field. The improved accessibility of large commercial and education computing resources can be harnessed to efficiently handle the terabytes of data gathered from geothermal operations and the algorithms needed to process them. Emphasis on a feature-based approach to time-series feature engineering and automated feature selection is another promising avenue for research that can aid modellers in developing accurate and understandable data models.

## CRediT authorship contribution statement

**Paul Michael B. Abrasaldo:** Conceptualization, Investigation, Data curation, Writing – original draft, preparation, revision. **Sadiq J. Zarrouk:** Supervision, Methodology, Writing – review & editing. **Andreas W. Kempa-Liehr:** Supervision, Methodology, Writing – review & editing.

## 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.

## Data availability

The authors do not have permission to share data.

## Acknowledgements

The authors would like to express their appreciation to the New Zealand Ministry of Business, Innovation and Employment (MBIE) for their financial support through the Empowering Geothermal Energy funds.

## References

[1] Fridleifsson IB. Prospects for geothermal energy worldwide in the new century. In: Proceedings world geothermal congress; 2000.
[2] Tutua-Nathan T. Maori tribal rights to ownership and control: the geothermal resource in New Zealand. Appl Geogr 1992;12:192–8. https://doi.org/10.1016/0143-6228(92)90007-A.
[3] Lund JW. Worldwide utilization of geothermal energy - 2005. GRC Transactions 2005;29.
[4] Lund JW. 100 Years of geothermal power product. Thirtieth Workshop on Geothermal Reservoir Engineering; 2005.
[5] Lund JW. World status of geothermal energy use overview 1995-1999. In: Proceedings world geothermal congress; 2000. 2000.
[6] Ragnarsson A. Geothermal development in Iceland 1995-1999. In: Proceedings world geothermal congress 2000; 2000. Tohoku, Japan.
[7] Fridleifsson IB. Geothermal research and development in Iceland 1982. New Zealand Geothermal Workshop; 1982.
[8] Yamaguchi F, Kawazoe S. Process of geothermal energy development in Japan. World Geothermal Congress; 1995.
[9] Suryantoro S, Dwipa S, Ariati R, Darma S. Geothermal deregulation and energy policy in Indonesia. In: Proceedings world geothermal congress; 2005. 24–9.
[10] Castrejon-Campos O. Evolution of clean energy technologies in Mexico: a multi-perspective analysis. Energy for Sustain. Dev. 2022;67:29–53. https://doi.org/10.1016/J.ESD.2022.01.003.
[11] Sussman D, Javellana SP, Benavidez PJ. Geothermal energy development in the Philippines: an overview. Geothermics 1993;22:353–67. https://doi.org/10.1016/0375-6505(93)90024-H.
[12] Gehringer M, Loksha V. Geothermal handbook: planning and financing power generation. 2012. Washington DC.
[13] Bertani R. Geothermal power generation in the world 2005–2010 update report. In: Proceedings world geothermal congress; 2010.
[14] Bertani R. Geothermal power generation in the world 2010-2014 update report. In: Proceedings world geothermal congress; 2015.
[15] Huttrer GW. The status of world geothermal power generation 1995-2000. In: Proceedings world geothermal congress; 2000. p. 2000.
[16] Huttrer GW. Geothermal power generation in the world 2015-2020 update report. In: Proceedings world geothermal congress 2020+1; 2021.
[17] Goyal KP, Conant TT. Performance history of the Geysers steam field, California, USA. Geothermics 2010;39:321–8. https://doi.org/10.1016/J.GEOTHERMICS.2010.09.007.
[18] Salonga ND, Dacillo DB, Siega FL. Providing solutions to the rapid changes induced by stressed production in Mahanagdong geothermal field, Philippines. Geothermics 2004;33:181–212. https://doi.org/10.1016/J.GEOTHERMICS.2003.08.008.
[19] Allis RG, Lumb T. Preservation of the rotorua geysers: conflicts and issues. Trans Geoth Resour Counc 1990;14.
[20] Mongillo MA, Allis RG. Continuing changes in surface activity at Craters of the Moon thermal area, Wairakei. In: Proceedings 10th New Zealand geothermal workshop; 1988. p. 1988.
[21] Hyodo M, Kitao K, Furukawa T. Development of database system for lost circulation and analysis of the data. In: Proceedings world geothermal congress; 2000. 2000.
[22] Akin S. Reservoir characterization by integrated pressure-transient and tracer-concentration/time data analysis. In: Proceedings. Twenty-Fifth Workshop on Geothermal Reservoir Engineering; 2000.
[23] Gonzalez LF, Aguiar AC, Karplus M. Data mining microseismicity associated to the blue mountain geothermal site. In: PROCEEDINGS, 47th workshop on geothermal reservoir engineering; 2022.
[24] Taverna N, Buster G, Huggins J, Rossol M, Siratovich P, Weers J, et al. Data curation for machine learning applied to geothermal power plant operational data for GOOML: geothermal operational optimization with machine learning. In: PROCEEDINGS, 47th workshop on geothermal reservoir engineering; 2022.
[25] Weers J, Frone Z, Huggins J, Vimont A. The data foundry: secure collaboration for the geothermal industry. In: PROCEEDINGS, 45th workshop on geothermal reservoir engineering; 2020.
[26] Adityatama D, Purba D, Marza S, Muhammad F, Asokawaty R, Kusumawardani R, et al. The significance of drilling data management to improve geothermal drilling planning and operation in Indonesia. In: Proceedings world geothermal congress; 2021. 2020+1.
[27] Witcher JC, Whittier J, Morgan R. New Mexico geothermal data base. Trans Geoth Resour Counc 1990;14.
[28] Okoroafor ER, Smith CM, Ochie KI, Nwosu CJ, Gudmundsdottir H, Jabs, Aljubran M. Machine learning in subsurface geothermal energy: two decades in review. Geothermics 2022;102:102401. https://doi.org/10.1016/j.geothermics.2022.102401.
[29] Muther T, Syed FI, Lancaster AT, Salsabila FD, Dahaghi AK, Negahban S. Geothermal 4.0: AI-enabled geothermal reservoir development- current status, potentials, limitations, and ways forward. Geothermics 2022;100:102348. https://doi.org/10.1016/j.geothermics.2022.102348.
[30] Kofod-Petersen A. How to do a structured literature review in computer science. 2015.
[31] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097. https://doi.org/10.1371/journal.pmed.1000097.
[32] Blei DM, Lafferty JD. Text mining. Classification, clustering, and applications. New York: Chapman and Hall/CRC; 2009.
[33] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[34] Mabey B, Susol M. pyLDAvis: Python library for interactive topic model visualization. Port of the R LDAvis package. 2016. https://github.com/bmabey/pyLDAvis.
[35] Sievert C, Shirley K. LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014. p. 63–70. https://doi.org/10.3115/v1/W14-3110.
[36] Sengar S, Liu X. Optimal electrical load forecasting for hybrid renewable resources through a hybrid memetic cuckoo search approach. Soft Comput 2020;24:13099–114. https://doi.org/10.1007/s00500-020-04727-9.
[37] Abd El-Aziz RM. Renewable power source energy consumption by hybrid machine learning model. Alex Eng J 2022;61:9447–55. https://doi.org/10.1016/J.AEJ.2022.03.019.
[38] Kialashaki A, Reisel JR. Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks. Appl Energy 2013;108:271–80. https://doi.org/10.1016/J.APENERGY.2013.03.034.
[39] Entchev E, Yang L, Ghorab M, Rosato A, Sibilio S. Energy, economic and environmental performance simulation of a hybrid renewable microgeneration system with neural network predictive control. Alex Eng J 2018;57:455–73. https://doi.org/10.1016/J.AEJ.2016.09.001.
[40] Huster WR, Schweidtmann AM, Lüthje JT, Mitsos A. Deterministic global superstructure-based optimization of an organic Rankine cycle. Comput Chem Eng 2020;141:106996. https://doi.org/10.1016/j.compchemeng.2020.106996.
[41] Huster WR, Schweidtmann AM, Mitsos A. Globally optimal working fluid mixture composition for geothermal power cycles. Energy 2020;212:118731. https://doi.org/10.1016/j.energy.2020.118731.
[42] Peng Y, Su W, Zhou N, Zhao L. How to evaluate the performance of sub-critical Organic Rankine Cycle from key properties of working fluids by group contribution methods? Energy Convers Manag 2020;221:113204. https://doi.org/10.1016/j.enconman.2020.113204.
[43] Huster WR, Schweidtmann AM, Mitsos A. Working fluid selection for organic rankine cycles via deterministic global optimization of design and operation. Optim Eng 2020;21:517–36. https://doi.org/10.1007/s11081-019-09454-1.
[44] Fadaei M, Ameri MJ, Rafiei Y, Ghorbanpour K. A modified semi-empirical correlation for designing two-phase separators. J Pet Sci Eng 2021;205:108782. https://doi.org/10.1016/j.petrol.2021.108782.
[45] Xu X, Liu J, Wang Y, Xu J, Bao J. Performance evaluation of ground source heat pump using linear and nonlinear regressions and artificial neural networks. Appl Therm Eng 2020;180:115914. https://doi.org/10.1016/j.applthermaleng.2020.115914.
[46] Chang W, Chu X, Binte Shaik Fareed AF, Pandey S, Luo J, Weigand B, et al. Heat transfer prediction of supercritical water with artificial neural networks. Appl Therm Eng 2018;131:815–24. https://doi.org/10.1016/j.applthermaleng.2017.12.063.
[47] Esen H, Inalli M. ANN and ANFIS models for performance evaluation of a vertical ground source heat pump system. Expert Syst Appl 2010;37:8134–47. https://doi.org/10.1016/j.eswa.2010.05.074.
[48] Benli H. Performance prediction between horizontal and vertical source heat pump systems for greenhouse heating with the use of artificial neural networks. Heat Mass Tran 2016;52:1707–24. https://doi.org/10.1007/s00231-015-1723-z.
[49] Kim J-S, Kim D-Y, Kim Y-T. Experiment on radial inflow turbines and performance prediction using deep neural network for the organic Rankine cycle. Appl Therm Eng 2019;149:633–43. https://doi.org/10.1016/j.applthermaleng.2018.12.084.
[50] Huang R, Zhang Z, Zhang W, Mou J, Zhou P, Wang Y. Energy performance prediction of the centrifugal pumps by using a hybrid neural network. Energy 2020;213:119005. https://doi.org/10.1016/j.energy.2020.119005.
[51] Yan D, Yang F, Yang F, Zhang H, Guo Z, Li J, et al. Identifying the key system parameters of the organic Rankine cycle using the principal component analysis based on an experimental database. Energy Convers Manag 2021;240:114252. https://doi.org/10.1016/j.enconman.2021.114252.
[52] Yang F, Zhang H, Hou X, Tian Y, Xu Y. Experimental study and artificial neural network based prediction of a free piston expander-linear generator for small scale organic Rankine cycle. Energy 2019;175:630–44. https://doi.org/10.1016/j.energy.2019.03.099.
[53] Dong S, Zhang Y, He Z, Deng N, Yu X, Yao S. Investigation of support vector machine and back propagation artificial neural network for performance prediction of the organic rankine cycle system. Energy 2018;144:851–64. https://doi.org/10.1016/j.energy.2017.12.094.
[54] Langiu M, Dahmen M, Mitsos A. Simultaneous optimization of design and operation of an air-cooled geothermal ORC under consideration of multiple operating points. Comput Chem Eng 2022;161:107745. https://doi.org/10.1016/j.compchemeng.2022.107745.

[55] Yang F, Cho H, Zhang H, Zhang J, Wu Y. Artificial neural network (ANN) based prediction and optimization of an organic Rankine cycle (ORC) for diesel engine waste heat recovery. Energy Convers Manag 2018;164:15–26. https://doi.org/10.1016/j.enconman.2018.02.062.

[56] Rodríguez JA, Hamzaoui YEl, Hernández JA, García JC, Flores JE, Tejeda AL. The use of artificial neural network (ANN) for modeling the useful life of the failure assessment in blades of steam turbines. Eng Fail Anal 2013;35:562–75. https://doi.org/10.1016/j.engfailanal.2013.05.002.

[57] Azizi S, Ahmadloo E, Awad MM. Prediction of void fraction for gas–liquid flow in horizontal, upward and downward inclined pipes using artificial neural network. Int J Multiphas Flow 2016;87:35–44. https://doi.org/10.1016/j.ijmultiphaseflow.2016.08.004.

[58] Lin Z, Liu X, Lao L, Liu H. Prediction of two-phase flow patterns in upward inclined pipes via deep learning. Energy 2020;210:118541. https://doi.org/10.1016/j.energy.2020.118541.

[59] Serra PLS, Masotti PHF, Rocha MS, de Andrade DA, Torres WM, de Mesquita RN. Two-phase flow void fraction estimation based on bubble image segmentation using Randomized Hough Transform with Neural Network (RHTN). Prog Nucl Energy 2020;118:103133. https://doi.org/10.1016/j.pnucene.2019.103133.

[60] Santamaría-Bonfil G, Santoyo E, Díaz-González L, Arroyo-Figueroa G. Equivalent imputation methodology for handling missing data in compositional geochemical databases of geothermal fluids. Geothermics 2022;104:102440. https://doi.org/10.1016/j.geothermics.2022.102440.

[61] Barrios Castellanos M, Serpa AL, Biazussi JL, Monte Verde W, do Socorro Dias Arrifano Sassim N. Fault identification using a chain of decision trees in an electrical submersible pump operating in a liquid-gas flow. J Pet Sci Eng 2020;184:106490. https://doi.org/10.1016/j.petrol.2019.106490.

[62] Kayfeci M, Yabanova İ, Keçebaş A. The use of artificial neural network to evaluate insulation thickness and life cycle costs: pipe insulation application. Appl Therm Eng 2014;63:370–8. https://doi.org/10.1016/j.applthermaleng.2013.11.017.

[63] Zhi L-H, Hu P, Chen L-X, Zhao G. Multiple parametric analysis, optimization and efficiency prediction of transcritical organic Rankine cycle using trans-1,3,3,3-tetrafluoropropene (R1234ze(E)) for low grade waste heat recovery. Energy Convers Manag 2019;180:44–59. https://doi.org/10.1016/j.enconman.2018.10.086.

[64] Peng Y, Lin X, Liu J, Su W, Zhou N. Machine learning prediction of ORC performance based on properties of working fluid. Appl Therm Eng 2021;195:117184. https://doi.org/10.1016/j.applthermaleng.2021.117184.

[65] Senturk Acar M. Multi-stage artificial neural network structure-based optimization of geothermal energy powered Kalina cycle. J Therm Anal Calorim 2021;145:829–49. https://doi.org/10.1007/s10973-020-10125-y.

[66] Jiang A, Qin Z, Cladouhos TT, Faulder D, Jafarpour B. Recurrent neural networks for prediction of geothermal reservoir performance. In: Proceedings, 46th workshop on geothermal reservoir engineering, vol. 46; 2021.

[67] Harry M, Situmorang J, Prabata W. A new machine learning algorithm for production well analysis. In: Proceedings, 46th workshop on geothermal reservoir engineering, vol. 46; 2021.

[68] Jiang A, Qin Z, Cladouhos TT, Faulder D, Jafarpour B. A multiscale recurrent neural network model for long-term prediction of geothermal energy production. In: Proceedings, 47th workshop on geothermal reservoir engineering; 2022.

[69] Harry M, Aditya Wahyudi MF, Midat Al Islam MP, Sabrina NA, Situmorang J. Comparative study of decline curve prediction in geothermal injection well using machine learning and wellbore simulator. In: Proceedings, 46th workshop on geothermal reservoir engineering, vol. 46. Stanford: Stanford University; 2021.

[70] Ling W, Liu Y, Young R. Deep learning models for prediction and optimization of air-cooled binary cycle geothermal operation. In: Proceedings, 47th workshop on geothermal reservoir engineering. Stanford, California: Stanford University; 2022.

[71] Arslan O. Power generation from medium temperature geothermal resources: ANN-based optimization of Kalina cycle system-34. Energy 2011;36:2528–34. https://doi.org/10.1016/j.energy.2011.01.045.

[72] He R, Wang Z, Fu W. Application of PID control based on BP neural network in the expansion machine of organic rankine cycle system. GRC Transactions 2018;42.

[73] Baser A, Kucuk S, Saracoglu O, Senturk E, Akın S. Optimization of production and injection of geothermal fields: a machine learning approach. Proceedings World Geothermal Congress 2020+1; 2021.

[74] Gheysari AF, Holländer HM, Maghoul P, Shalaby A. Sustainability, climate resiliency, and mitigation capacity of geothermal heat pump systems in cold regions. Geothermics 2021;91:101979. https://doi.org/10.1016/j.geothermics.2020.101979.

[75] Yilmaz C, Koyuncu I. Thermoeconomic modeling and artificial neural network optimization of Afyon geothermal power plant. Renew Energy 2021;163:1166–81. https://doi.org/10.1016/j.renene.2020.09.024.

[76] Lalot S, Pálsson H. Detection of fouling in a cross-flow heat exchanger using a neural network based technique. Int J Therm Sci 2010;49:675–9. https://doi.org/10.1016/j.ijthermalsci.2009.10.011.

[77] Álvarez del Castillo A, Santoyo E, García-Valladares O. A new void fraction correlation inferred from artificial neural networks for modeling two-phase flows in geothermal wells. Comput Geosci 2012;41:25–39. https://doi.org/10.1016/j.cageo.2011.08.001.

[78] Jin G, Xing H, Zhang R, Guo Z, Liu J. Data-driven discovery of governing equations for transient heat transfer analysis. Comput Geosci 2022;26:613–31. https://doi.org/10.1007/s10596-022-10145-7.

[79] Arslan O, Yetik O. ANN based optimization of supercritical ORC-Binary geothermal power plant: Simav case study. Appl Therm Eng 2011;31:3922–8. https://doi.org/10.1016/j.applthermaleng.2011.07.041.

[80] Fast M, Palmé T. Application of artificial neural networks to the condition monitoring and diagnosis of a combined heat and power plant. Energy 2010;35:1114–20. https://doi.org/10.1016/j.energy.2009.06.005.

[81] Kim M, Yoon SH, Payne WV, Domanski PA. Development of the reference model for a residential heat pump system for cooling mode fault detection and diagnosis. J Mech Sci Technol 2010;24:1481–9. https://doi.org/10.1007/s12206-010-0408-2.

[82] Keçebaş A, Yabanova İ, Yumurtacı M. Artificial neural network modeling of geothermal district heating system thought exergy analysis. Energy Convers Manag 2012;64:206–12. https://doi.org/10.1016/j.enconman.2012.06.002.

[83] Keçebaş A, Yabanova İ. Thermal monitoring and optimization of geothermal district heating systems using artificial neural network: a case study. Energy Build 2012;50:339–46. https://doi.org/10.1016/j.enbuild.2012.04.002.

[84] Şencan Şahin A, Yazıcı H. Thermodynamic evaluation of the Afyon geothermal district heating system by using neural network and neuro-fuzzy. J Volcanol Geoth Res 2012;233(234):65–71. https://doi.org/10.1016/j.jvolgeores.2012.04.020.

[85] Fannou J-LC, Rousseau C, Lamarche L, Kajl S. Modeling of a direct expansion geothermal heat pump using artificial neural networks. Energy Build 2014;81:381–90. https://doi.org/10.1016/j.enbuild.2014.06.040.

[86] Sun W, Hu P, Lei F, Zhu N, Jiang Z. Case study of performance evaluation of ground source heat pump system based on ANN and ANFIS models. Appl Therm Eng 2015;87:586–94. https://doi.org/10.1016/j.applthermaleng.2015.04.082.

[87] Ruliandi D. Geothermal power plant system performance prediction using artificial neural networks. In: IEEE conference on technologies for sustainability (SusTech). IEEE; 2015. p. 216–23. https://doi.org/10.1109/SusTech.2015.7314349. 2015.

[88] Yılmaz F, Selbaş R, Şahin AŞ. Efficiency analysis of organic Rankine cycle with internal heat exchanger using neural network. Heat Mass Tran 2016;52:351–9. https://doi.org/10.1007/s00231-015-1564-9.

[89] Priyangga HY, Ruliandi D, Pertamina PT, Merdeka J, No T. Application of pattern recognition and classification using artificial neural network in geothermal operation. In: PROCEEDINGS, 43rd workshop on geothermal reservoir engineering; 2018.

[90] Ruliandi D, Dwi Susanto A, Djanarto. Application of artificial neural network to exergy performance analysis of geothermal power plant. In: Proceedings world geothermal congress 2020+1; 2021.

[91] Ping X, Yang F, Zhang H, Xing C, Yao B, Wang Y. An outlier removal and feature dimensionality reduction framework with unsupervised learning and information theory intervention for organic Rankine cycle (ORC). Energy 2022;254:124268. https://doi.org/10.1016/j.energy.2022.124268.

[92] Ping X, Yang F, Zhang H, Xing C, Zhang W, Wang Y. Evaluation of hybrid forecasting methods for organic Rankine cycle: unsupervised learning-based outlier removal and partial mutual information-based feature selection. Appl Energy 2022;311:118682. https://doi.org/10.1016/j.apenergy.2022.118682.

[93] Jiang A, Qin Z, Faulder D, Cladouhos TT, Jafarpour B. Recurrent neural networks for short-term and long-term prediction of geothermal reservoirs. Geothermics 2022;104:102439. https://doi.org/10.1016/j.geothermics.2022.102439.

[94] Yan L, Hu P, Li C, Yao Y, Xing L, Lei F, et al. The performance prediction of ground source heat pump system based on monitoring data and data mining technology. Energy Build 2016;127:1085–95. https://doi.org/10.1016/j.enbuild.2016.06.055.

[95] Park SK, Moon HJ, Min KC, Hwang C, Kim S. Application of a multiple linear regression and an artificial neural network model for the heating performance analysis and hourly prediction of a large-scale ground source heat pump system. Energy Build 2018;165:206–15. https://doi.org/10.1016/j.enbuild.2018.01.029.

[96] Wibowo SN, Aji P, Taufiq Fathaddin M, Oetomo HK, Pudyastuti K, Dalimunthe YK, et al. A robust prediction method based on artificial neural network for power output of organic rankine cycle in lahendong geothermal field. In: Proceedings world geothermal congress; 2021. 2020+1.

[97] Zulkarnain Surjandari I, Bramasta RR, Laoh E. Fault detection system using machine learning on geothermal power plant. In: 2019 16th international conference on service systems and service management (ICSSSM). IEEE; 2019. p. 1–5. https://doi.org/10.1109/ICSSSM.2019.8887710.

[98] Siratovich P, Buster G, Taverna N, Rossol M, Weers J, Blair A, et al. GOOML-finding optimization opportunities for geothermal operations. In: Proceedings, 47th workshop on geothermal reservoir engineering. Stanford, Caifornia: Stanford University; 2022.

[99] Keçebaş A, Yabanova İ. Economic analysis of exergy efficiency based control strategy for geothermal district heating system. Energy Convers Manag 2013;73:1–9. https://doi.org/10.1016/j.enconman.2013.03.036.

[100] Keçebaş A, Alkan MA, Yabanova İ, Yumurtacı M. Energetic and economic evaluations of geothermal district heating systems by using ANN. Energy Pol 2013;56:558–67. https://doi.org/10.1016/j.enpol.2013.01.039.

[101] Lin Y, Wang H, Hu P, Yang W, Hu Q, Zhu N, et al. A study on the optimal air, load and source side temperature combination for a variable air and water volume ground source heat pump system. Appl Therm Eng 2020;178:115595. https://doi.org/10.1016/j.applthermaleng.2020.115595.

[102] Liu Y, Ling W, Young R, Hsieh M. Deep learning for prediction and fault detection in geothermal operations. In: Proceedings, 46th workshop on geothermal reservoir engineering, vol. 46. Stanford, California: Stanford University; 2021.

[103] Lemmon EW, Bell IH, Huber ML, McLinden MO. NIST standard reference database 23: reference fluid thermodynamic and transport properties-REFPROP,

version 10.0. National Institute of Standards and Technology; 2018. https://doi.org/10.18434/T4/1502528.
[104] Bell IH, Wronski J, Quoilin S, Lemort V. Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library CoolProp. Ind Eng Chem Res 2014;53:2498–508. https://doi.org/10.1021/ie4033999.
[105] Loewenberg MF, Laurien E, Class A, Schulenberg T. Supercritical water heat transfer in vertical tubes: a look-up table. Prog Nucl Energy 2008;50:532–8. https://doi.org/10.1016/J.PNUCENE.2007.11.037.
[106] Fulcher BD. Feature-based time-series analysis. Feature Engineering for Machine Learning and Data Analytics 2017;87–116. https://doi.org/10.48550/arxiv.1709.08055.
[107] Murty MN, Devi VS. Introduction to pattern recognition and machine learning, vol. 5. Bangalore, India: Co-Published with Indian Institute of Science (IISc); 2015. https://doi.org/10.1142/8037.
[108] Kalogirou SA. Applications of artificial neural networks in energy systems. Energy Convers Manag 1999;40:1073–87. https://doi.org/10.1016/S0196-8904(99)00012-6.
[109] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Network 1994;5:157–66. https://doi.org/10.1109/72.279181.
[110] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory - colt '92. New York, New York, USA: ACM Press; 1992. p. 144–52. https://doi.org/10.1145/130385.130401.
[111] Draper NR. Applied regression analysis. third ed. New York: Wiley; 1998.
[112] Groß J. Linear regression, vol. 175. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. https://doi.org/10.1007/978-3-642-55864-1.
[113] Cook D. Practical machine learning with H2O. first ed. Incorporated: O'Reilly Media; 2016.
[114] Dahan H, Cohen S, Rokach L, Maimon O. Proactive data mining with decision trees. New York, NY: Springer New York; 2014. https://doi.org/10.1007/978-1-4939-0539-3.
[115] Zhang C, Cao L, Romagnoli A. On the feature engineering of building energy data mining. Sustain Cities Soc 2018;39:508–18. https://doi.org/10.1016/J.SCS.2018.02.016.
[116] Ahmadi A, El Haj Assad M, Jamali DH, Kumar R, Li ZX, Salameh T, et al. Applications of geothermal organic Rankine Cycle for electricity production. J Clean Prod 2020;274:122950. https://doi.org/10.1016/J.JCLEPRO.2020.122950.
[117] Kempa-Liehr AW, Oram J, Wong A, Finch M, Besier T. Feature engineering workflow for activity recognition from synchronized inertial measurement units. 2019. https://doi.org/10.1007/978-981-15-3651-9_20.
[118] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. third ed. Melbourne, Australia: OTexts; 2021.
[119] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Cham: Springer International Publishing; 2018. https://doi.org/10.1007/978-3-319-98074-4.
[120] Dempsey DE, Cronin SJ, Mei S, Kempa-Liehr AW. Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand. Nat Commun 2020;11:3562. https://doi.org/10.1038/s41467-020-17375-2.