6-2023

Utilizing Machine Learning Models to Estimate Energy Savings from an Industrial Energy System

Eva McLaughlin

Jun-Ki Choi

Resources, Environment and Sustainability 12 (2023) 100103

Contents lists available at ScienceDirect

Resources, Environment and Sustainability

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

Utilizing machine learning models to estimate energy savings from an industrial energy system

Eva McLaughlin, Jun-Ki Choi *

Renewable and Clean Energy Program, Department of Mechanical Engineering University of Dayton, 300 College Park, Dayton, OH, USA

ARTICLE INFO

Keywords:
Compressed air system
Energy efficiency
Supervised machine learning
Compressed air leaks
Industrial energy audits

ABSTRACT

Energy audits are an important part of reducing energy usage, costs, and carbon emissions, but there have been discrepancies in the quality of audits depending upon the auditor, which can negatively affect the impacts and credibility of the energy assessment. In this paper, historical energy auditing data from a U.S. Department of Energy sponsored research program was gathered and analyzed with a machine-learning algorithm to predict demand savings from a compressed air system assessment recommendation in industrial manufacturing facilities. Different energy auditors calculate savings for repairing leaks in compressed air systems in various ways, so the energy demand savings have been calculated differently throughout the historical assessment recommendations. Machine learning models are utilized in order to enhance the accuracy of the existing practice and reduce variations resulting from the abovementioned discrepancies. A large set of historical assessment recommendation data was used to train five unique machine learning models. Four base learner models and one metalearner model were devised and compared. Results showed that the distributed random forest model best predicted compressed air energy demand savings against the new scenarios within an error of 17%. This indicates that the distributed random forest model can more accurately quantify savings from repairing leaks in compressed air systems. In addition, the results from this study provide insight into the important factors contributing to leaks in the compressed air systems and why it is crucial to repair those leaks regularly to save money and energy while decreasing emissions.

1. Introduction

Improving energy efficiency is a common concern among the world, especially manufacturers. As climate change continues to worsen, it is more important than ever to reduce energy consumption and decrease emissions (Zaidi et al., 2021; Mungai et al., 2022). Industrial energy accounts about 30 percent of total world energy consumption (Choi et al., 2018a). Energy audits are an important first step in reducing industries' energy usage by analyzing energy consumption and recommending more energy-efficient practices (Vivek Jadhav et al., 2017; Johansson et al., 2022). Energy audits optimize energy usage without affecting the facility's output, productivity, or comfort (Gokul et al., 2017; Errigo et al., 2022). Historically, industrial energy provided opportunity to save energy consumption around 13% under the U.S. Department of Energy's Industrial Assessment Center program (Kapp et al., 2023). Though energy audits are crucial in reducing energy consumption and emissions, issues can arise in the lack of methodology or training of auditors (Shen et al., 2012; Kapp et al., 2022). Low-quality energy audits often result from unstandardized audit procedures and recommendations, which negatively impact the adoption rate of energy efficiency recommendations (Tobias Fleiter and Ravivanpong, 2012; Carlander and Thollander, 2022). These issues can lead to discrepancies and errors in energy audits. Studies have shown that standardized assessment recommendations (ARs) assisted with the machine learning approach could assist in developing effective and inexpensive auditing methods since energy audits tend to be costly and time-consuming (Marasco and Kontokosta, 2016; Glick et al., 2021; Goss and King, 2020).

Machine learning utilizes data-driven models focusing on the relationships in the raw data and variables instead of equations and calculations (Montáns et al., 2019). Auditing organizations and manufacturing facilities have significant historical energy audit data that can be used in the machine learning model and helps identify energy and cost savings more accurately. Manufacturers typically record energy usage data for observational purposes, not for process improvement. However, this underutilized valuable data can be employed to discover and identify the most important control variables in manufacturing processes (Sadati et al., 2018).

This paper utilized the historical data from the University of Dayton Industrial Assessment Center, that has performed over 1050 energy audits on manufacturers over the past forty years (University of Dayton Industrial Assessment Center, 2022). This paper analyzes the data from the compressed air system (CAS) of all the energy systems focused on

* Corresponding author.
E-mail address: jchoi1@udayton.edu (J.-K. Choi).

https://doi.org/10.1016/j.resenv.2022.100103
Received 5 August 2022; Received in revised form 22 November 2022; Accepted 12 December 2022
Available online 16 December 2022

during these audits. Specifically, it looks at the recommendation of repairing leaks in CASs. Compressed air usage accounts for about 10% of total industrial energy usage (Saidur et al., 2010). Leaks in CASs can waste as much as 20 to 50% of compressed air output (Scaron et al., 2011; Kaya et al., 2002). New cases are introduced and developed to show the efficacy of the methodology developed in this work. By utilizing machine learning methodology for CASs in manufacturing facilities, the savings for repairing industrial compressed air leaks can be more accurately quantified, and the most important variables can be identified.

CAS is highly energy intensive and one of the more expensive equipment in industrial facilities (Dindorf, 2012). CAS is equipment that is often inefficiently utilized in industrial applications such as equipment cooling, blow-off, and liquid agitation, among other misuses (Nagarkar and Pal, 2021). Some common appropriate uses of compressed air include controls, actuators, air tools, and vacuum (McLaughlin et al., 2021). With the wide range of industrial uses, leaks in CASs are unavoidable. They can contribute to a significant amount of energy wasted, especially through joints, hoses, and other connections in the compressed air lines. Though large compressed air leaks can be heard, most leaks are small and difficult to find, which requires leak detection equipment like an ultrasound or an infrared detector to identify and help quantify the leaks (Dudić et al., 2012). Necessary information for quantifying leaks and their potential savings in CASs can be obtained through energy audits (Paucar et al., 2017).

Energy auditors require various information about the compressor and plant operations to accurately calculate savings from repairing leaks (Choi et al., 2018b). Compressors can have different flow controls, which influence how much power they are using depending on how loaded the compressors are. The common types of flow control are start/stop, load/unload, inlet-modulation, auto-dual, and variable speed drive (Challent, 2002). Though all of these modes can be beneficial in various circumstances, Variable speed drive (VSD) is typically recommended in manufacturers due to its capability. Variable speed drive (VSD) controls allow the motor to vary its speed depending upon the compressed air demand (Kissock, 2005; Schmidt and Kissock, 2003; Shinde and Jadhav, 2017; Mousavi et al., 2014). The other types of flow do not have such a wide range of variability so the compressors in typical manufacturers would be unable to save the same amount of energy.

Machine learning could help calculate savings from repairing leaks by utilizing historical industrial energy consumption data to estimate energy and cost savings more accurately with reducing the carbon dioxide emission (Naji et al., 2021). There are three main types of machine learning: supervised, unsupervised, and reinforcement learning (Deepika et al., 2018). This paper will focus on supervised learning. Supervised learning uses historical data to predict the desired outputs, so the model learns through examples (Alzubi et al., 2018). The gathered data for this research is more suited to supervised learning because the data spans the past forty years and already has input and output data labeled which the supervised model can learn and apply to unlabeled future datasets.

Machine learning is frequently used in manufacturing industries to manage and process data from the facility and products to support manufacturing decisions, estimate product costs, and improve manufacturing facility operations (Sharp et al., 2018). It was found that utilizing machine learning capabilities reduced error and uncertainty by 50% in measurement and verification applications in the industry (Gallagher et al., 2018). Similarly, smart manufacturing and predictive manufacturing combine machine learning, big data, artificial intelligence, and advanced technology to optimize efficiency and productivity while reducing costs and production time (Haricha et al., 2021; Nikolic et al., 2017). Another study was able to accurately predict maintenance system failures using a random forest machine learning model with a high R-squared (R^2) value based on the real-time data from the production line equipment (Ayvaz and K, 2021).

Machine learning techniques have been frequently used in manufacturing facilities (Shook and Choi, 2022) and some CASs. Previous research has shown how intelligent systems can decrease energy usage in CASs by using past data to anticipate future needs (Thabet et al., 2021). An ensemble learning framework, known as stochastic gradient boosting, was developed to model the properties of compressed air, which would minimize laboratory testing to acquire these specific properties (Dehaghani et al., 2022). Bendetti et al. created maturity models to optimize efficiency in compressed air systems using autonomous assessment tools to evaluate the system from potential managerial opportunities instead of focusing on technological concerns (Benedetti et al., 2019). It has been suggested that deep reinforced learning and recurrent neural networks could be possible approaches to improve overall compressed air system efficiencies (Thabet et al., 2020). One concern with utilizing these models is that they use historical data, so if there are any severe changes to the compressed air system or technology, the models would no longer be accurate (Metthee, 2021). These examples from the literature indicate there has been some research into utilizing machine learning in compressed air systems and lay the groundwork for further research.

Although a large amount of research has been done focusing on machine learning in manufacturers, there is little research on how energy auditors can adopt and utilize similar machine learning practices to perform better audits based on historical data. This article explores how to estimate energy and cost savings from compressed air leaks in industrial compressed air systems using various machine learning techniques based on manufacturers' past audit data.

2. Methodology

The overall research steps are shown in Fig. 1. The first step is to review the energy auditing process. This process has three main steps: a pre-assessment, a facility tour, and a post-assessment. Specific assessment recommendations (ARs) are suggested during each energy audit to decrease energy demand based on data and calculations. In our 40 years' historical data, there were three unique ways of estimating the compressed air energy demand savings by repairing leaks. We classified all data based on these three different methodologies. Secondly, the data from all compressed air leak recommendations were collected and organized prior to creating models using machine learning algorithms. Thirdly, five specific models were developed from the data utilizing various machine learning techniques. These models vary in strengths to capture different patterns in the data. Finally, results from several models compared to identify the highest performing model that is to be used in estimating savings from repairing leaks in CASs. Three validation metrics used to compare the models include the R-squared values, the mean absolute error values, and the root mean square error values. Three different scenarios were also used against the five models to determine which models were able to predict the energy demand savings more accurately. More details of each step are described in this section.

3. Case study

3.1. Industrial energy/resource auditing

The research framework starts with the data collection from the industrial energy audits performed by a U.S. Department of Energy sponsored Industrial Assessment Center program. UD-IAC has performed energy and resource audits (Choi et al., 2019) for small and medium enterprises in the Ohio region for forty years. Every audit had three main parts: the baseline analysis, the audit, and the final report (University of Dayton Industrial Assessment Center, 2022). The baseline analysis used the specific facility's utility, production, and weather data to break down their utility charges and identify how much of their energy consumption could be attributed to weather or production dependency.

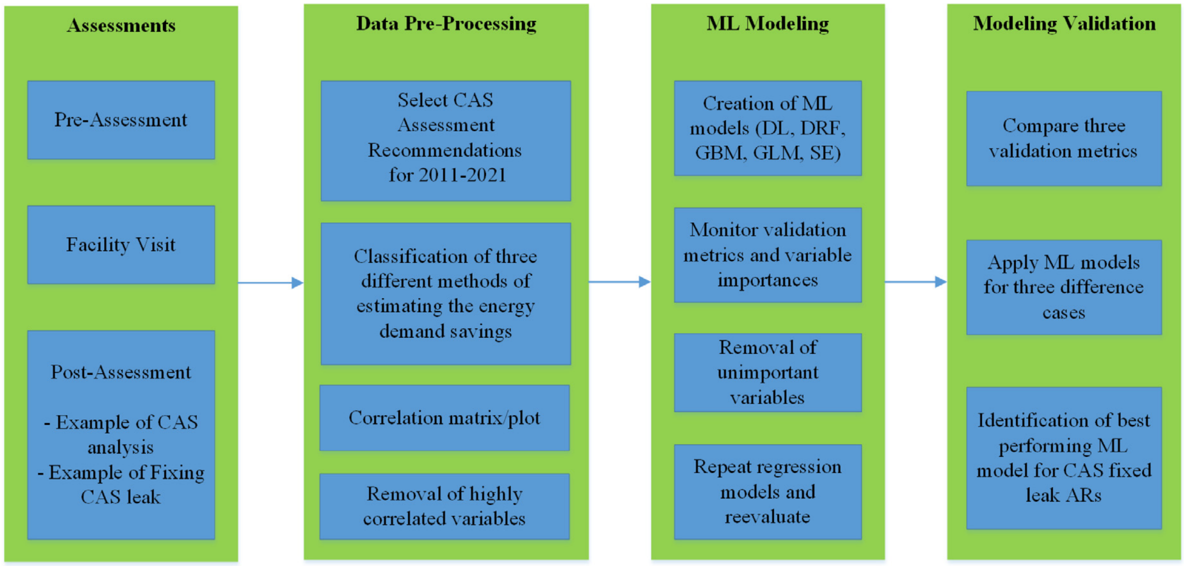

Fig. 1. Process framework for creating machine learning models to predict energy demand savings from repairing leaks in CASs.

This initial analysis also includes gathering the facility's general data like the products, operating hours, shift arrangement, number of employees, floor area, and the annual utility usage. This information helps the energy auditors understand the utility rate structures and how the facility operates throughout the day, week, and year before attending the audit. During the audit, the energy auditors tour the facility and gather further information about the equipment and practices throughout the plant. Based on the information gathered throughout the tour, the auditors will recommend energy-efficient practices that the facility could implement to save energy and money. The last stage of the audit process is the final report. The final report contains all details from the audit, including the baseline analysis, recommendations, and potential implementation strategies. The specific recommendations describe what is being proposed, show the calculations for the cost and emission savings, and discuss implementation cost and internal rate of return. The final report summarizes all the recommendations clearly and concisely for the facility's benefit.

3.1.1. *Example of CAS analysis*

Repairing leaks in the CAS is one of the most common recommendations for facilities with compressed air. Compressed air is expensive and energy-intensive, so if leaks are not being regularly checked and repaired, then leaks could contribute to a large amount of the facility's compressed air consumption, as discussed in the background section.

This section offers one example of the compressed system analysis, which mainly focuses on finding total compressed output, economic savings, and CO_2 savings. Fig. 2 shows the current compressed air system in a plant the audit team visited. It consisted of a 150-hp compressor (CAS 1), a 100-hp compressor (CAS 2) and a 75-hp compressor (CAS 3). The 100-hp compressor was located in a separate building, while the other two were in the boiler room. All compressors operate continuously throughout the year. These air compressors were all under modulation control with auto shut-off. Each of the 100-hp and 150-hp compressors was served by a heated desiccant dryer. The 75-hp compressor was served by an unheated desiccant dryer. The current operating set point was 105 psig.

The audit team logged the current draw of all three compressors. According to management, the voltage to the air compressors is 480 V. Compressor was monitored continuously using a data logger. The current draw was measured at the main electrical feed to the air compressor. The power factor is assumed to be 0.86. Assuming this current draw profile represents typical manufacturing conditions, the average power when operating the air compressor can be calculated by Eq. (1) (Alkadi, 2011).

Table 1
Compressed air systems information and calculated parameters.

Compressor information	CAS 1	CAS 2	CAS 3
Horsepower (hp)	150	100	75
Power Factor*	0.86	0.86	0.86
Voltage (V)	480	480	480
Average Current Draw (Amps)	206.6	120.9	82.6
Fan Size (hp)	5	5	5
P_{avg} (kW)	129.2	74.0	49.4
P_{max} (kW)	120.2	78.7	58.0
Fraction Power (FP)	1.1	0.9	0.9

$$P_{avg} = Voltage\,(\mathrm{V}) \times Current\,(\mathrm{A}) \times \sqrt{Phase} \times PF\left(\frac{\mathrm{kW}}{\mathrm{kVA}}\right) \times \frac{1\ \mathrm{kW}}{1,000\ \mathrm{W}} - P_{fan} \quad (1)$$

Based on the size of the air compressors, we estimated the fan motors to be 5 hp. The average fraction power of the air compressor, FP, can be calculated by Eq. (2). Table 1 shows the parameters associated with the three CAS.

$$Fraction\ Power\ (FP) = \frac{Average\ Power\ Draw}{Max.Power\ Draw} \quad (2)$$

The fraction capacity can be found in Eq. (3).

$$Fraction\ Capacity\ (FC) = \frac{(FP - FP_0)}{(1 - FP_0)} \quad (3)$$

As seen in Fig. 3, modulating the intake air is one of the least efficient control strategies for an air compressor. Modulation compressors are estimated to be 0.70 fraction power at zero fraction capacity.

Typical compressed air output from older rotary screw compressors is about 4.2 scfm/hp. Therefore, the average compressed air output for the air compressor can be solved with Eq. (4) (Abels and Kissock, 2011).

$$Compressed\ Air\ Output\ (C) = \frac{4.2\,\mathrm{scfm}}{\mathrm{hp}} \times \frac{1\ \mathrm{hp}}{0.746\ \mathrm{kW}} \times P_{max} \times FC \quad (4)$$

Table 2 shows the calculated values, and the total average compressed air output of the CAS was 1200 scfm.

3.1.2. *CAS leak energy efficiency recommendation example*

This section shows one example of calculating savings from CAS by fixing the leaking. As shown in the previous section, the average

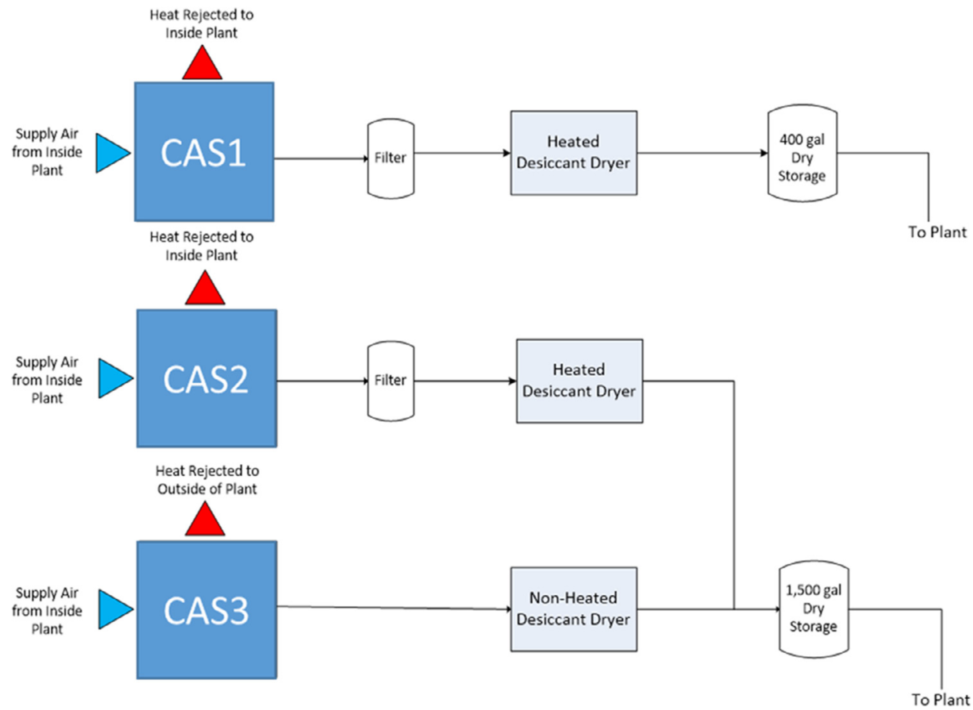

Fig. 2. Schematic of compressed air systems in a facility.

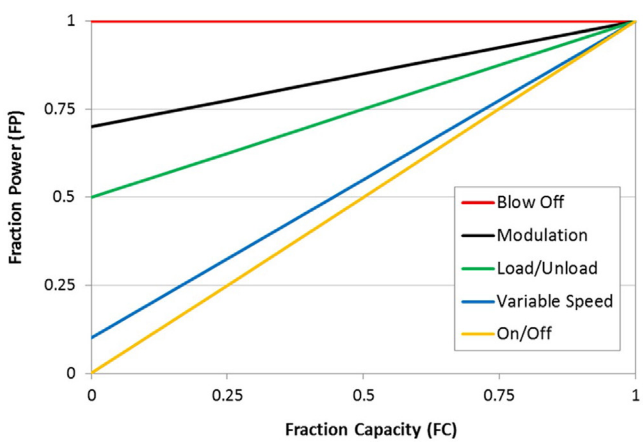

Fig. 3. General power/output relationship by CAS control type.

Table 2
Parameters utilized for calculating the average compressed air output of each CAS.

Compressor information	CAS 1	CAS 2	CAS 3
Fraction Power at No Load (FP_0)	0.7	0.7	0.7
Average Fraction Power (FP)	1.07	0.94	0.85
Fraction Capacity (FC)	125%	80%	51%
P_{max} (kW)	120.2	78.7	58.0
Avg. Compressed Air Output (scfm)	845.4	355.0	165.5

compressed air consumption for CAS 1 was found to be 845 scfm and 355 scfm for CAS 2, and 165 scfm for CAS 3. Leaks in the compressed air system increase the load on the air compressors resulting in excess energy consumption. According to the US Department of Energy, leaks typically comprise 20%–30% of the total compressed air load in plants without maintenance programs. According to management, the facility process calls for about 95 psig compressed air, and the air compressors are set at 105 psig to overcome distribution loss. When the audit was performed, the facility already had an ultrasonic leak detector, which was not frequently used. It can greatly aid in detecting compressed air leaks. An effective way to minimize compressed air leaks is to establish a bi-weekly or monthly preventative maintenance program to identify and fix compressed air leaks with the detector. Therefore, we recommended instituting a monthly leak check program to identify and repair compressed air leaks. Table 3 shows one example of calculating the compressed air savings for CAS 1.

The same calculation was performed for two other compressors (i.e., CAS 2 and CAS3). Table 4 shows total cost savings and CO_2 emission savings for three compressors.

We estimated that it would take about an hour per month to identify and fix leaks in the compressed air system at a labor cost of $50/hr. Therefore, the annual labor cost would be about $600/yr. The net total cost savings would be about $3736/yr–$600/yr = $3136. Since there is already an ultrasonic leak detector used in the facility, there would not be any implementation cost. Simple payback would be immediate.

3.1.3. Variations in calculating the demand savings from the fix leak recommendation

Throughout the historical data, we found that the calculations for repairing leaks have differed depending upon the audit and auditor. There have been three main methods for calculating savings for repairing leaks recommendation. There is no reason why different methods have been used, except different auditors may prefer one method over another. Also, the methods require different input information, which may influence which method is used if there is no access to the necessary information. Common input variables for these methods include compressed air reduction, electricity savings, fraction power at no production, and carbon dioxide emission savings. Compressed air reduction refers to the amount of compressed air in standard cubic feet per minute that could be reduced for a specific facility if they implemented a program to repair leaks.

As shown in Eq. (5), Compressed air reduction can be calculated as a product of average compressed air consumption, percent of compressed air consumption contributed to leaks, and the realistic percentage of leaks that could be regularly fixed.

$$CAR = (C_{avg}) \times (PL) \times (PCAR) \quad (5)$$

Table 3
Technical information of CAS 1 and facility for savings calculations.

Term	Value	Units
Average Compressed Air Consumption (C_{avg})	845	scfm
Percentage of total due to leaks (PL)	20%	–
Nominal Compressed Air Output (CAO)	4.2	scfm/hp
Fraction Power at No Production (FP_0)	70%	–
Motor Efficiency (Em)	90%	–
Percentage Compressed Air Leaks Reduction (PCAR)	50%	–
Average Operating Hours per Year (HPY)	8,760	h/year
Electricity Demand Cost (CD)	$16.11	/kW
Electricity Energy Cost (CE)	$0.0307	/kWh
Electrical Utility Carbon Intensity Factor (CIF)	1.56	lb-CO_2/kWh
Compressed Air due to Leaks (CAL) = C_{avg} × PL	169.0	scfm
Compressed Air Reduction (CAR) = CAL × PCAR	84.5	scfm
Demand Savings (DS) = (CAR/CAO × [0.746 kW/hp] × (1−FP_0)/Em	5.0	kW
Electricity Savings (ES) = DS × HPY	43,826	kWh/year
Demand Cost Savings (DCS) = DS × CD × [12 months/year]	$967	/year
Electricity Cost Savings (ECS) = ES × CE	$1,345	/year
CO_2 Emission Savings (CES) = ES × CIF/[2,205 lb/tonne]	31	tonnes
Total cost savings (TCS) = DCS + ECS	$2,313	/year

Table 4
Total economic and environmental savings from all three CAS.

Term	Value	Units
Demand Savings (DS)	8.1	kW
Electricity Savings (ES)	70,795	kWh/year
Demand Cost Savings (DCS) = DS × CD × [12 months/year]	$1,562	/year
Electricity Cost Savings (ECS) = ES × CE	$2,173	/year
CO_2 Emission Savings (CES) = ES × CIF/[2,205 lb/tonne]	50	tonnes
Total cost savings (TCS) = DCS + ECS	$3,736	/year

where *CAR* is the amount of compressed air reduced in standard cubic feet per minute, C_{avg} is the average compressed air consumption, *PL* is the estimated percent of leaks, and *PCAR* is the realistic percent of leaks that could be regularly repaired. C_{avg} and *PL* can be retrieved from data loggers if applicable, or else plant personnel are able to estimate the values. Eq. (6) calculates the electricity savings from the product of energy demand savings and annual operation hours.

$$ES = (DS) \times (HPY) \tag{6}$$

where *ES* is the total electricity savings, *DS* is the energy demand savings in kilowatts, and *HPY* is the annual operating hours. Eq. (7) calculates the carbon dioxide emission savings from the electricity savings utilizing a carbon intensity factor and pound-to-tonne conversion factor.

$$CES = (ES) \times (CIF) \ / \ [2,205 \text{ lb/tonne}] \tag{7}$$

where *CES* is the carbon dioxide emission savings, *ES* is the electricity savings, and *CIF* is the carbon intensity factor. In the U.S. Midwest, we typically use 1.95 pounds of CO_2 per kilowatt hour (eGrid, 2016). Compressor flow controls utilize varying degrees of power at differing loads. Compressor fraction power is important when calculating savings in CASs and understanding a facility's compressed air consumption. Eq. (8) represents a generalized linear relationship between associated engineering parameters (Kissock, 2005; Murphy and Kelly Kissock, 2015).

$$FP = (FP_0) + (1 - FP_0) \times FC \tag{8}$$

where FP is full-load power, FC is fraction rated capacity and FP_0 is fraction power at no production which can vary from zero to one with start/stop controls at zero, followed by VSD at about 0.1, load/unload at about 0.5, and modulation at 0.75 (Murphy and Kelly Kissock, 2015).

The first type of estimating the electricity demand savings is shown in Eq. (9) (Schmidt et al., 2005). This was the most common calculation methodology found in our historical data and the one used in the case study shown in the previous section.

$$DS = ((CAR/CAO) \times [0.746 \text{ kW/hp}]) \times (1 - FP_0) \tag{9}$$

where, *CAO* is the nominal compressed air output. *CAO* and FP_0 can be found on the compressor's data sheets

The second most frequently used type for estimating the compressed electricity demand savings from leak detection is shown in Eqs. (10)–(12) (Schmidt and Kissock, 2003).

$$RFC = (C_{avg} - CAR)/(C_{avg}/AFC) \tag{10}$$

$$RFP = RFC \times (1 - FP_0) + FP_0 \tag{11}$$

$$DS = MCPD \times (AFP - RFP) \tag{12}$$

where, *RFC* is the reduced fraction capacity of the compressor, *AFC* is the average fraction power which can be calculated from the data loggers, *RFP* is the reduced fraction power, and *MCPD* is the maximum compressor power draw which can also be calculated from the data loggers or retrieved from the compressor's data sheets, *AFP* is the average fraction power.

The last type to calculate the compressed air electricity demand savings utilizes the voltage and amperage of the compressor, as shown in Eq. (13) (Schmidt and Kissock, 2004). This method is appropriate if the compressors can be turned off when there is no production and leaks are eliminated. Often, compressors run during hours of non-production due to leaks in the system. If the non-production amperage of the compressor was obtained through data loggers at the facility, then potential savings can be estimated using the voltage, amperage, power factor, efficiency, and annual operating hours. This method has not been frequently used, though.

$$DS = V \times A \times sqrt(3) \times (PCAR/1000) \tag{13}$$

where *V* is the voltage needed for the compressor, *A* is the amperage of the compressor during non-production hours. The voltage and amperage can be obtained from the compressor nameplate and data logger. Table 5 shows the summary of three different ways of calculating the electricity demand savings. Other calculations such as CAR, ES, CES are sam.

Table 5
Summary of three different Compressed Air Demand Saving.

Term	Variable	Calculation	Units
Compressed Air Reduction	CAR	C_{avg} × PL × PCAR	scfm
Demand Savings (Type I)	DS	(CAR/CAO × [0.746 kW/hp] × (1−FP_0)	kW
Demand Savings (Type II)	DS	MCPD × (AFP − RFP)	kW
Demand Savings (Type III)	DS	V × A × sqrt(3) × PCAR/1000	kW
Electricity Savings	ES	DS × HPY	kWh/year
CO_2 Emission Savings	CES	ES × CIF/[2,205 lb/tonne]	tonnes

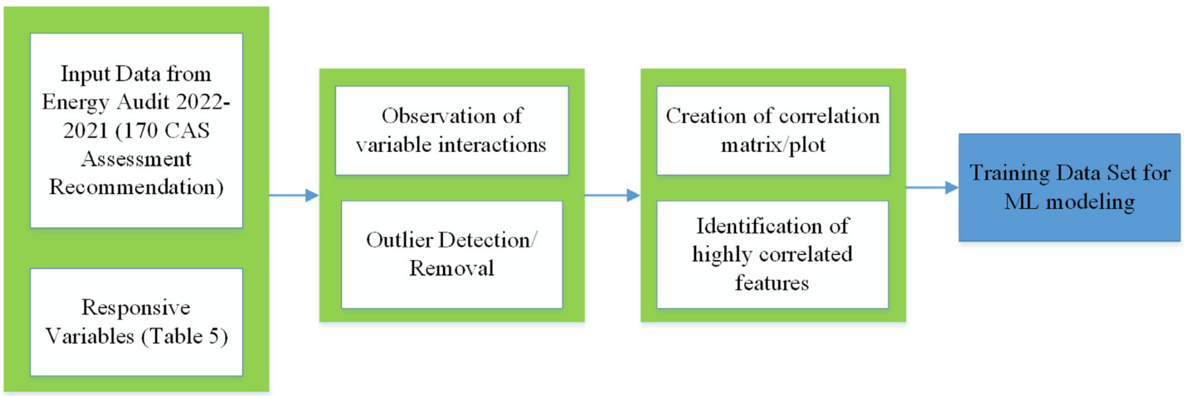

Fig. 4. Dataflow management steps for the data pre-processing.

Table 6
Input data and description.

Category	Variables
Operating parameters	Employed Average compressed air consumption Percent of leaks Annual operating hours
Economic/environmental	Type of industry/production Carbon dioxide emission Carbon intensity factor
Equipment parameters	Full CAS load power Fraction rated capacity Fraction power at no production Nominal compressed air output Reduced fraction capacity of compressor Average fraction power
Target variable	Electrical Energy Demand

3.2. Data pre-processing

After the data was collected, the data needed to be cleaned and pre-processed before creating a supervised machine learning model from the data. Fig. 4 shows the dataflow management utilized for the pre-processing step.

The data is made of numeric input variables and a response variable. Table 6 lists the variables used in the different ML algorithm. There are three categories of data that contribute to the prediction. The operating parameters provide insight into how much energy is consumed based on the amount of time that the facility is in use. More employees, larger compressed air consumption, percentage leaks, and longer occupied hours contribute to increased energy usage. Economic and environmental parameters are indicative of production energy usage and associated carbon dioxide emissions. The third category deals with major process equipment—compressed air power, output that provide large and unique independent loads. These parameters are of particular interest because they are unique to manufacturing facilities and compressed air systems, making them essential to training models that accurately capture compressed air energy demand.

Supervised machine learning models learn patterns and functions from the input data and response variable. By observing and identifying how the variables in the data interact and influence each other, the model attempts to learn correlations within the data. After creating the connections between variables, the model can utilize the historical and the new input data to predict the response variable output. Recommendations where information was missing, were removed if the necessary information was unable to be found. Also, scatter plots were created between some variables to identify if there were any major outliers. Some outliers were found, but after investigation, it was discovered that values were mistyped in those instances. The values were therefore corrected before other pre-processing could continue.

The next pre-processing step was creating a correlation matrix from the data. At this point in the process, there was data from about 170 ARs for repairing leaks in CASs. Each recommendation was from a unique facility with different general facility data and specific compressor data. The correlation matrix was created to identify the highly correlated features. Multiple highly correlated features could cause unnecessary complexity and errors in the algorithm.

For this reason, a correlation matrix was created, and highly correlated features were removed. Fig. 5 shows a correlation chart. The vertical axis on the right side of the figure depicts varying colors from dark red to dark blue. The dark blue indicates a high positive correlation between the variables, while the dark red indicates a high negative correlation. The lighter colors are closer to zero and indicate little correlation. If two features are highly correlated, then one of the features is removed because they have about the same impact on the data. Highly correlated features were identified with a correlation value of .90 or higher. For example, the compressor horsepower and average compressed air consumption have a correlation of 0.8, while fraction power at no production has a negative 0.2 correlation with kilowatt savings.

3.3. Machine learning modeling

The third step was to create a machine learning model that best represents the compressed air analysis. The historical data were utilized to create several machine learning models so the validation metrics could be compared and the best model could be identified. Five ML models were created. Descriptions of these five machine learning models are provided below:

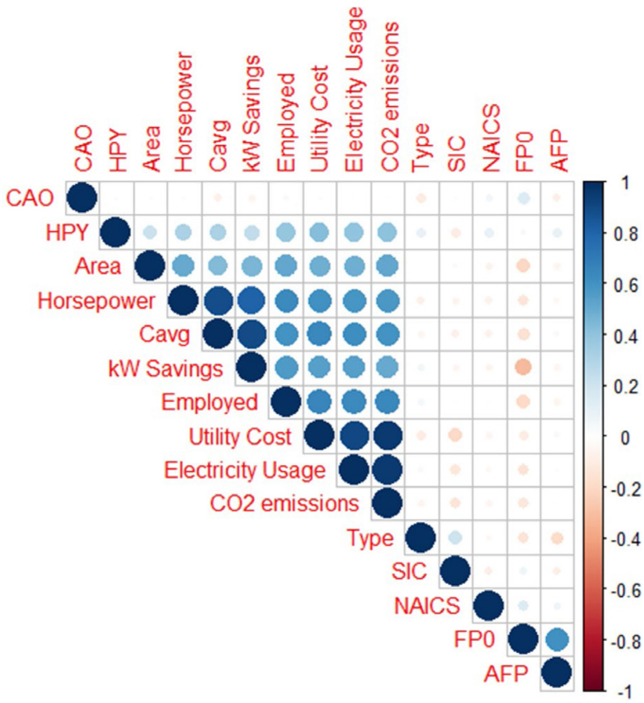

Fig. 5. Correlation plot between variables.

1. Deep Learning (DL) is an artificial neural network that consists of neuron layers. There is the input layer, the hidden layers, and the output layer. Supervised deep learning models use back-propagation and are based on stochastic gradient descent. (Schmidhuber, 2015)
2. Distributed Random Forest (DRF) selects groups of features at random and favorably splits the data based on the subsets of predictors. The number of trees wanted depends on the user, but the process is continuously repeated for a specified number of iterations. A random subset is created with each new iteration. (Liaw and Wiener, 2002)
3. Gradient Boosting Machine (GBM) also utilizes decision trees, but adds a new tree at each iteration to focus on the weaknesses in the model and minimize the mean squared error. (Touzani et al., 2018)
4. Generalized Linear Modeling (GLM) describes various general distributions, like Gaussian and Poisson, that can estimate the response variable. (Xia et al., 2021)
5. Stacked Ensemble (SE) is a metalearner that uses multiple base learners. A metalearner model creates a model based on the best models from the other four models. A stacked ensemble combines the information from all the chosen base learners (i.e., DL, DRF, GBM, GLM) to create a more experienced model to create a single stronger model. A stacked ensemble uses cross-validation to avoid overfitting the model, a common issue in GBM and GLM. Cross-validation occurs when the data is split into a number of groups, and the majority of those groups are used to train the model, and the others are used to validate the model. This subgrouping, training, and validation will occur for as many iterations as the user specifies. (Xia et al., 2021)

In addition to the validation metrics, each trial output variable assigns numeric values from 0 to 1 to each variable. 0 indicates the variable had little effect on the model, while a value closer to 1 indicates a greater effect on the model. Variables with variable importance of 0.02 or below were removed from the model. Variables with little importance can add unnecessary information to the model making it more complicated.

3.4. Model validation

Validating the model provides evidence that the model is performing up to the expected standards. Models can be easily created, but the evidence is needed to show that it is accurate and reliable, so it is important to validate the model by testing it against historical data. The model can be evaluated through different validation metrics. Validation metrics are a way to compare models and understand how well the model could predict values. Three key validation metrics were used to compare the various models' quality. These metrics are the R-squared (R^2), mean absolute error (MAE), and root mean squared error (RMSE). As shown in the Eq. (14), the R^2 value is the coefficient of determination that tells how close the regressor, or model, is to the actual data. R^2 values indicate how much of the variation in the data can be explained by the model. If the model can explain all the variations, then the R^2 would be one, while it would be zero if the model did not explain any of the data. As shown in Eq. (15), the MAE value measures the average difference between the predicted and actual values in the regressor. The smaller the MAE value is, the smaller the error amount. As shown in Eq. (16), the RMSE is similar to MAE, but the square root is taken from the squared difference between the predicted and actual values. As with the MAE value, RMSE values indicate better regressors when the value is closer to zero.

$$R^2 = \frac{\sum_{i=1}^{n}(y_{predict,i} - \bar{y}_{data})^2}{\sum_{i=1}^{n}(y_{data,i} - \bar{y}_{data})^2} \tag{14}$$

$$MAE = \frac{\sum_{i=1}^{n} \lfloor y_{predict,i} - y_{data} \rfloor}{n} \tag{15}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{predict,i} - y_{data})^2}{n}} \tag{16}$$

where $y_{predict,i}$ is the predicted values i, y_{data} is the historical values achieved from the training dataset, n is the number of ARs in the dataset, and the $\bar{y}_{data}$ is the average value.

4. Results

The different machine learning models can be compared through the validation metrics and new cases introduced to the models. The models with superior validation metrics indicate the model should perform well, but introducing new cases to the models allows the models' predictive accuracy to be tested. The accuracy of the models can therefore be compared to identify which model can be adapted to predict energy demand savings for repairing leaks in CASs accurately. Utilizing the historical data, multiple models were created to predict savings in CASs. With each consecutive model, variables were removed if they had less significance to the model since unimportant variables can create poor models with irrelevant data. The best models and their results are selected. Table 7 compares the validation metrics for five different models. R^2 values closer to one indicate the model can fit the training data well. For example, the gradient boosting machine model has the highest R^2 value, so its model can account for almost all variations in the data. The distributed random forest has the lowest R^2 value, so its model can account for the majority but not all variation in the data. From applying various machine learning models to the collected data, it can be seen what variables are most important to the estimated savings from repairing leaks in CASs. Variables were removed if they had a less significance to the models which left the data with seven variables: C_{avg}, employee, horsepower of the CAS, electricity usage variables, area, utility cost, FP_0 In each model, the variables are weighted differently which impact how each model uses the variable data to predict.

Fig. 6 provides insight into what variables influence energy demand savings for the ARs of repairing leaks in CASs. Only the most important variables and their scaled importance are included. Since stacked ensemble (SE) combines the best models, it does not have variable

Table 7
Comparison of the validation metrics for five machine learning models applied to predict energy savings in CASs.

Machine learning model	Validation metrics			Cross-validation
	R^2	MAE	RMSE	R^2
Deep learning	.852	2.910	5.22	.815
Distributed random forest	.786	4.187	6.272	.779
Gradient boosting machine	.999	0.144	0.220	.803
Generalized linear modeling	.825	3.739	5.667	.787
Stacked ensemble	.974	1.431	2.168	.834

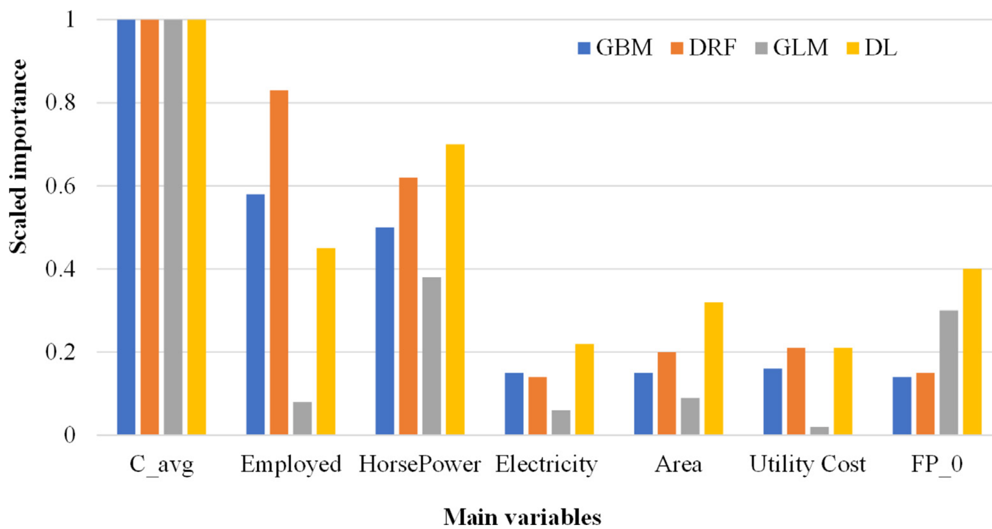

Fig. 6. The relative variable importance for each model.

importance, but one for the other four applicable models (GBM, DRF, GLM, DL) is included. For all the models, the average compressed air consumption is the most important variable, which makes sense because the quantity of compressed air has a clear correlation to the amount of compressed air lost through leaks. The horsepower of the compressors is also not a surprising variable since the horsepower is another indicator of the capacity of compressed air that the compressor can produce. Fraction power at no production is another variable that is not surprising since it is frequently used in calculating savings. Compressors have different control modes that affect the fraction power at no production, which was previously discussed. The other four variables identified as having importance to savings are more surprising. These four variables are indicators of the size of the facility. Facilities frequently have tools powered by compressed air that individual employees operate. Thus, the number of employees could relate to the number of compressed air tools. Similarly, the annual electricity and utility costs indicate the plant's energy usage. The energy usage coupled with the size of the compressor can help the model identify how much energy usage is going towards the compressors and, therefore leak.

The five models were tested against data not included in the training or cross-validation data. Data from three separate industrial energy audits were used in cases to compare the predicted energy demand savings for each model. Only the previously discussed seven variables were used to predict the kilowatt savings for three compressed air leak recommendations.

In the first case, the facility undergoing the energy audit had ninety employees, 98,000 square feet of area, an annual utility cost of $149,272, and used 1,974,155 kWh of electricity per year. The actual energy demand savings calculated during the energy audit were 3.9 kW. Fig. 7 shows each model's predicted energy demand savings as columns with a line indicating the percent error from the predicted savings and the originally calculated savings. The predicted energy demand savings range from 4.25 kW to 6.21 kW, with most predictions within one kW of the actual energy demand savings. The outlier is 6.21 kW from the GBM, which tends to overfit the data. The DL model predicted energy demand savings closest to the calculated savings.

The facility was much larger in the second case than in the first case. There were 655 employees, square footage of 418,855, an annual utility cost of $ 3,296,258, and used 46,960,650 kWh of electricity per year. The actual savings were 55.3 kW. Fig. 8 shows the predicted values from each model compared to the actual energy demand savings. Since the facility was large, the predicted and actual energy demand savings are higher than in the first case, and the predicted energy demand savings range is much larger. Most manufacturers use compressed air for various usages, so it would be understandable that the compressed air usage and leaks would increase with facility size. The DRF model underestimates the energy demand savings while the other models overestimate, with the stacked ensemble model predicting the highest savings. In this case, the GBM model has the least percent error, followed by the distributed random forest model.

In the last case, the facility had 180 employees, square footage of 170,000, an annual utility cost of $1,321,244, and used 11,741,214 kWh of electricity per year. This facility has a size between the first and second cases. The actual savings were 12.4 kW. Fig. 9 shows the predicted values from each model compared to the actual savings. The distributed random forest model had the least percentage error compared to the original savings.

From the previous three example cases, the percentage of error was calculated for each combination of model and case. Fig. 10 shows each case's percent error and average error for each model. The overall accuracy and precision of the models can be seen as well as how each model performs depending upon the facility size. Case 1 has the smallest facility size, case 2 has the largest, and case 3 has a size between the first two. Depending upon the facility size, different models could be used to predict energy demand savings accurately.

5. Discussion

Five machine learning models were created and then the models were compared to each other through validation metrics, variable

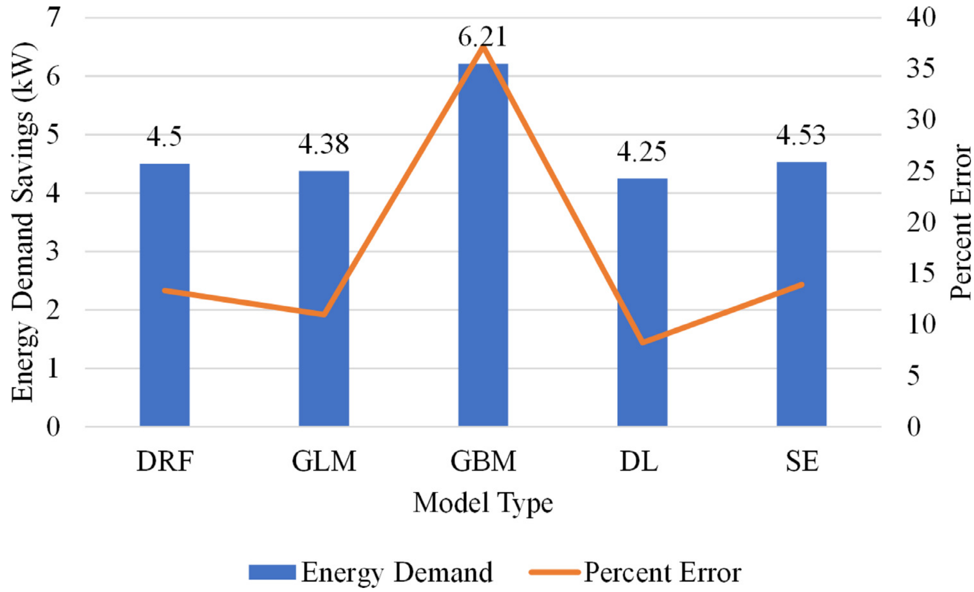

Fig. 7. Comparison of predicted energy demand savings for case 1.

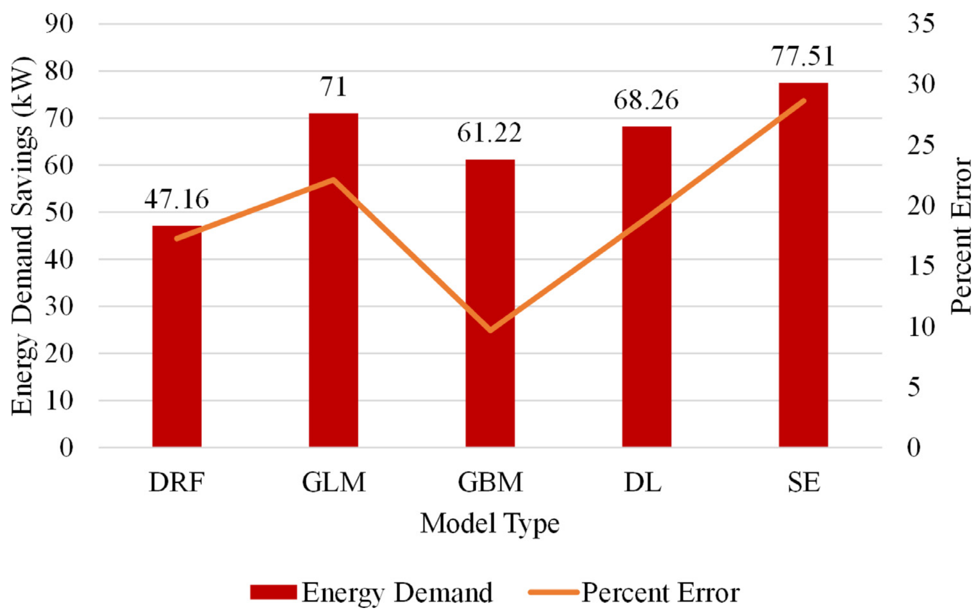

Fig. 8. Comparison of predicted energy demand savings for case 2.

importance, and cases applying each model to a new scenario. None of the models consistently performed well in the validation metrics and all three trials. Based on the validation metrics, the gradient boosting machine has superior metrics and should fit the data best. The cross-validation metric for the stacked ensemble has an R-squared value closest to one, indicating the stacked ensemble model should also perform very well against new data. By applying the five different models to three industrial cases, it was observed that the gradient boosting machine model and the stacked ensemble model did not predict savings accurately, with average percent errors of about 30%. The stacked ensemble model had over a 50% error in the third case. Though the stacked ensemble had good validation and cross-validation metrics which should indicate a well-performing model, it is clear from the three cases that the model is severely lacking in accurately predicting savings for scenarios unfamiliar with the input data. Conversely, the distributed random forest model had relatively poor validation metrics but performed well in different cases. The R-squared values for the validation and cross-validation metrics were the lowest of all the models, and the MAE and RMSE values were the highest.

During the application of the ML models to three different testing cases, the distributed random forest model had the least amount of average error. The distributed random forest model still had an error in its predictions, but since the model does not overfit the data. The error is more contained and smaller than in the other models because it does not extrapolate as much as the other types of machine learning models discussed in this paper. Though the distributed random forest model performed most accurately overall, other models performed better for specific cases. The deep learning model predicted more accurate savings for the first case, while the gradient boosting machine model predicted most accurately for the second case. So the deep learning model may be used for smaller facilities, while the gradient boosting machine model may be used to predict energy demand savings for larger facilities. The distributed random forest model has the lowest average error for all three cases so this model is more generally applicable to all situations.

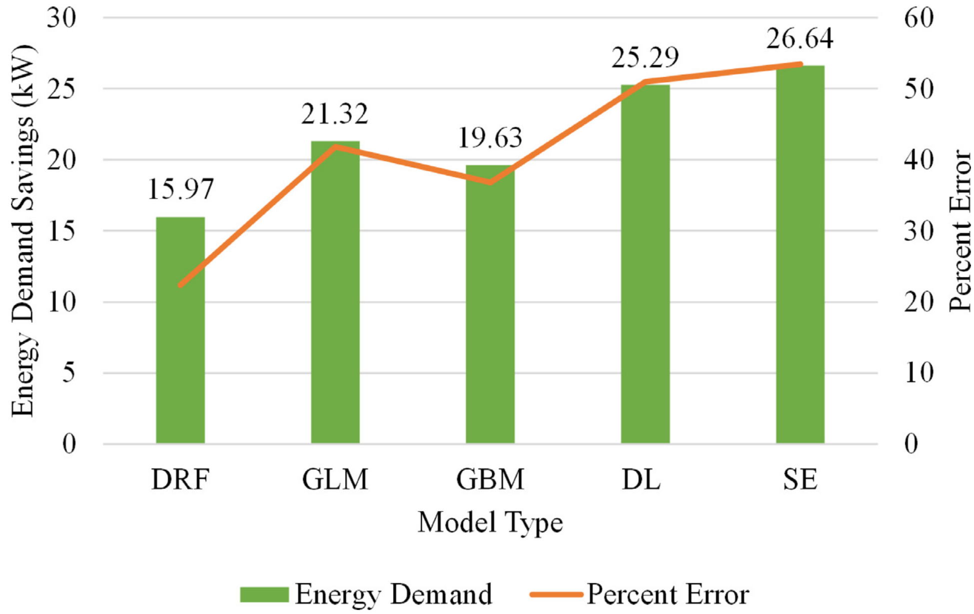

Fig. 9. Comparison of predicted energy demand savings for case 3.

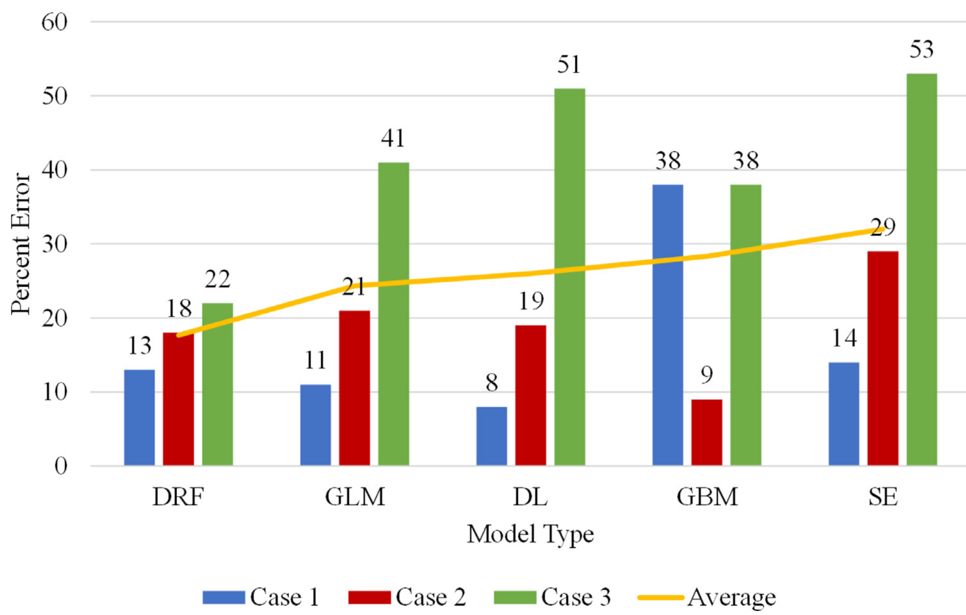

Fig. 10. Summary of percent error for each case.

Though the distributed random forest model performs the best out of all the models, more data may be needed to create a better model. Only 170 data points were used, and it is clear that the models are still lacking in accuracy. The distributed random forest model can predict kilowatt savings for repairing leaks within an average error of 17%. More historical data would improve this average percent error, but the model can be used in energy audits to approximate kilowatt savings within an error of 17%. One of the limitation of this study is that there is no data feedback from the manufacturers to determine whether these savings are accurate. Obtaining more data and feedback from the manufacturers would be recommended in future studies. Though the model is not a perfect predictor and can have errors in the historical data, it is a start to better understanding how machine learning can work with various systems in manufacturers since data collection and analysis is becoming so crucial in recent years.

6. Conclusion

Compressed air is frequently used by manufacturers for control valves and other pneumatic tools. Air compressors are large energy consumers, so it is critical to minimize compressed air usage to minimize compressed air costs. Compressed air leaks can consume much of a facility's total compressed air consumption. Repairing leaks is a good practice, but facility personnel often do not repair them since it can be time-consuming. This paper discusses the potential savings from repairing compressed air leaks based on historical data from an energy auditing research program. Energy audits have helped calculate energy demand savings in several ways over the years in this program. Since the calculations have been completed differently, a machine learning model was created to estimate the energy demand savings. The distributed random forest model can help energy auditors better estimate savings for repairing leaks in CASs. The most important variables were also identified. These variables include general facility information and specific compressor information, but the variables should not provide much difficulty in obtaining during a physical or virtual energy audit. This paper provides insight into the important factors contributing to leaks in CASs and why it is crucial to repair them regularly to save money and energy while decreasing emissions.

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.

Acknowledgments

We would like to express our gratitude to the US Department of Energy for supporting this work through its funding of the Industrial Assessment Center program (DE-EE0009721). We thank previous and current UD-IAC students for their contributions to this continuing effort and our industrial partners for their significant contributions.

References

Abels, Brian James, Kissock, Kelly, 2011. Optimizing compressed air storage for energy efficiency. SAE Internat. J. Mater. Manufact. 4 (1), 477–485.

Alkadi, K, 2011. Improving compressed air energy efficiency in automotive plants, In: SAE 2011 World Congress, Detroit, MI, April (2011) 12-14.

Alzubi, J., Nayyar, A., Kumar, A., 2018. Machine learning from theory to algorithms: an overview. In: Journal of Physics: Conference Series. IOP, Moscow, Russia.

Ayvaz, S.A., K, 2021. Predictive maintenance system for production lines in manufacturing: a machine learning approach using IoT data in real-time. Expert Syst. Appl. 173.

Benedetti, M., Bonfà, F., Bertini, I., Introna, V., Salvatori, S., Ubertini, S., Paradiso, R., 2019. Maturity-based approach for the improvement of energy efficiency in industrial compressed air production and use systems. Energy 186.

Carlander, J., Thollander, P., 2022. Drivers for implementation of energy-efficient technologies in building construction projects — Results from a Swedish case study. Resour. Environ. Sustain. 10, 100078.

Challent, C.A., 2002. Fundamentals of Compressed Air Systems. US Department of Energy.

Choi, J.-K., Eom, J., McClory, E., 2018a. Economic and environmental impacts of local utility-delivered industrial energy-efficiency rebate programs. Energy Policy 123, 289–298.

Choi, J.-K., Schuessler, R., Ising, M., Kelley, D., Kissock, K., 2018b. A pathway towards sustainable manufacturing for mid-size manufacturers. Procedia CIRP 69, 230–235.

Choi, J.-K., Thangamani, D., Kissock, K., 2019. Resour. Conserv. Recy. 147, 19–27.

University of Dayton Industrial Assessment Center, 2022. Energy audit for small and medium sized industry. https://go.udayton.edu/iac.

Deepika, J., Senthil, T., Rajan, C., Surendar, A., 2018. Machine learning algorithms: a background artifact. Int. J. Eng. Technol. 7 (1), 143–149.

Dehaghani, A.H.S., Soleimani, R., Mohammadi, A.H., 2022. Screening of important parameters in optimal design of compressed air energy storage system using an ensemble learning method. J. Energy Storage 48.

Dindorf, R., 2012. Estimating potential energy savings in compressed air systems. Procedia Eng. 39, 204–211.

Dudić, S., Ignjatović, I., Šešlija, D., Blagojević, V., Stojiljković, M., 2012. Leakage quantification of compressed air using ultrasound and infrared thermography. Measurement 45 (7), 1689–1694.

eGrid, E., 2016. Power profiler. Available from: https://www.epa.gov/egrid/power-profiler#/.

Errigo, A., Choi, J.-K., Kissock, K., 2022. Techno-economic-environmental impacts of industrial energy assessment: Sustainable industrial motor systems of small and medium-sized enterprises. Sustain. Energy Technol. Assess. 49, 101694.

Gallagher, C.V., Bruton, K., Leahy, K., O'Sullivan, D.T., 2018. The suitability of machine learning to minimise uncertainty in the measurement and verification of energy savings. Energy Build. 158, 647–655.

Glick, M.B., Peppard, E., Meguro, W., 2021. Analysis of methodology for scaling up building retrofits: Is there a role for virtual energy audits?—A first step in Hawai'i, USA. Energies 14 (18), 5914.

Gokul, M., Ponraj, S., Rohan, S., Venkatesh, M., Prince Winston, D., Praveen Kumar, B., 2017. Energy audit for indutries and institution. Int. J. Adv. Res. Manag. Architec. Technol. Eng. 3 (2), 199–205.

Goss, M., King, T., 2020. Changing how energy audits are done: The rapid energy modeling tool quickly and comprehensively analyzes campus energy efficiency, targets inefficient building, benchmarks energy use, and provides insights. Consult. Specif. Eng. 57 (3), 26–29.

Haricha, K., Khiat, A., Issaoui, Y., Bahnasse, A., Ouajji, H., 2021. Towards smart manufacturing: implementation and benefits. J. Ubiquitous Syst. Pervasive Networks 15 (2), 25–31.

Johansson, I., Johnsson, S., Thollander, P., 2022. Impact evaluation of an energy efficiency network policy programme for industrial SMEs in Sweden. Resour. Environ. Sustain. 9, 100065.

Kapp, S., Choi, J.-K., Hong, T., 2023. Predicting industrial building energy consumption with statistical and machine-learning models informed by physical system parameters. Renew. Sustain. Energy Rev. 172, 113045.

Kapp, S., Choi, J.-K., Kissock, K., 2022. Toward energy-efficient industrial thermal systems for regional manufacturing facilities. Energy Rep. 8, 1377–1387.

Kaya, D., Phelan, P., Chau, D., Ibrahim Sarac, H., 2002. Energy conservation in compressed air systems. Int. J. Energy Res. 26 (9), 837–849.

Kissock, K., 2005. Modeling and simulation of air compressor energy use. ACEEE Summer Study on Energy Effic. Ind. 1 (13), 131–142.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Marasco, D.E., Kontokosta, C.E., 2016. Applications of machine learning methods to identifying and predicting building retrofit opportunities. Energy Build. 128, 431–441.

McLaughlin, E., Choi, J.-K., Kissock, K., 2021. Techno-economic impact assessments of energy efficiency improvements in the industrial combustion systems. J. Energy Resour. Technol. 144 (8).

Metthee, G.J., 2021. Improved Control of Compressed Air Networks using Machine Learning. North-West University, South Africa.

Montáns, F.J., Chinesta, F., Gómez-Bombarelli, R., Kutz, J.N., 2019. Data-driven modeling and learning in science and engineering. C. R. Mécanique 347 (11), 845–855.

Mousavi, S., Kara, S., Kornfeld, B., 2014. Energy efficiency of compressed air systems. Procedia Cirp 15, 313–318.

Mungai, E.M., Ndiritu, S.W., Da Silva, I., 2022. Unlocking climate finance potential and policy barriers—A case of renewable energy and energy efficiency in Sub-Saharan Africa. Resour. Environ. Sustain. 7, 100043.

Murphy, Sean, Kelly Kissock, J., 2015. Simulating energy efficient control of multiple-compressor compressed air systems. In: Proceedings of the Industrial Energy Technology Conference. New Orleans, LA, pp. 1–13.

Nagarkar, P., Pal, P., 2021. Conserving Energy in Compressed Air System: Practical Case Studies from Indian Industry. Springer International Publishing, Cham.

Naji, A., Al Tarhuni, J.-K., Choi, B., Ajena, S., 2021. Toward cost-effective residential energy reduction and community impacts: A data-based machine learning approach. Energy and AI 4, 100068.

Nikolic, B., Ignjatic, J., Suzic, N., Stevanov, B., Rikalovic, A., 2017. Predictive manufacturing systems in industry 4.0: trends, benefits, and challenges. Annals of DAAAM & Proceedings 28.

Paucar, M.A., Amancha, P.I., Viera, E.F., San Antonio, T.D., Salazar, D.M., 2017. Implementation of a methodology to perform an energy audit with academic purpose. Int. J. Appl. Eng. Res. 12 (24), 14908–14913.

Sadati, N., Chinnam, R.B., Nezhad, M.Z., 2018. Observational data-driven modeling and optimization of manufacturing processes. Expert Syst. Appl. 93, 456–464.

Saidur, R., Rahim, N.A., Hasanuzzaman, M., 2010. A review on compressed-air energy use and energy savings. Renen. Sustain. Energy Rev. 14 (4), 1135–1153.

Scaron, D., Ignjatovi, I., Dudi, S., Lagod, B., 2011. Potential energy savings in compressed air systems in Serbia. Afr. J. Bus. Manag. 5 (14), 5637–5645.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Netw. 61, 85–117.

Schmidt, C., Kissock, K., 2003. Power characteristics of industrial air compressors, In: Industrial Energy Technology Conference, Houston, TX.

Schmidt, C., Kissock, K., 2004. Estimating energy savings in compressed air systems, In: Industrial Energy Technology Conference, Houston, TX, pp. 200.

Schmidt, C., Kissock, K., Eger, C., 2005. Improving performance of compressed air systems through calibrated simulation, In: ACEEE Summer Study on Energy in Industry, West Point, NY, July (2005) 19-22.

Sharp, M., Ak, R., Hedberg Jr., T., 2018. A survey of the advancing use and development of machine learning in smart manufacturing. J. Manuf. Syst. 48, 170–179.

Shen, B., Price, L., Lu, H., 2012. Energy audit practices in China: National and local experiences and issues. Energy Policy 46, 346–358.

Shinde, A., Jadhav, H.T., 2017. Energy saving through air compressor system automation. In: 2017 International Conference on Circuit, Power, and Computing Technologies. IEEE, Kollam, India, pp. 1–7.

Shook, P., Choi, J.-K., 2022. Predicting the impact of utility lighting rebate programs on promoting industrial energy efficiency: A machine learning approach. Environments 9 (8), 100.

Thabet, M., Sanders, D., Becerra, V., Tewkesbury, G., Haddad, M., Barker, T., 2020. Intelligent energy management of compressed air systems. 2020 IEEE 10th International Conference on Intelligent Systems. IEEE, Varna, Bulgaria, pp. 153–158.

Thabet, M., et al., 2021. Management of compressed air to reduce energy consumption using intelligent systems. In: Proceedings of SAI Intelligent Systems Conference. Springer, Cham, London, UK, pp. 206–217.

Tobias Fleiter, J.S., Ravivanpong, Ployplearn, 2012. Adoption of energy-efficiency measures in SMEs—An empirical analysis based on energy audit data from Germany. Energy Policy 51, 863–875.

Touzani, S., Granderson, J., Fernandes, S., 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build. 158, 1533–1543.

Vivek Jadhav, R.J., Magar, Pramod, Kharat, Sandip, Bagwan, S.U., 2017. Energy conservation through energy audit. In: International Confernce on Trends in Electronics and Informatics. Institute of Electrical and Electronics Engineers, Tirunelveli, India, pp. 481–485.

Xia, Y., Chen, K., Yang, Y., 2021. Multi-label classification with weighted classifier selection and stacked ensemble. Inform. Sci. 557, 421–442.

Zaidi, S.A.H., Hussain, M., Uz Zaman, Q., 2021. Dynamic linkages between financial inclusion and carbon emissions: Evidence from selected OECD countries. Resour. Environ. Sustain. 4, 100022.