Contents lists available at ScienceDirect

Journal of Engineering Research

journal homepage: www.journals.elsevier.com/journal-of-engineering-research

Power plant induced-draft fan fault prediction using machine learning stacking ensemble

Tlamelo Emmanuel [a,*], Dimane Mpoeleng [b], Thabiso Maupong [b]

[a] *Computing and Information Systems, Botswana Accountancy College, Francistown, Botswana*
[b] *Department of Computer Science and Information Systems, Botswana International University of Science and Technology, Palapye, Botswana*

ARTICLE INFO

Keywords:
Fault prediction
Induced draft fans
Stacking ensemble

ABSTRACT

The improvement of fault prediction and diagnosis in industrial systems is crucial to minimize unscheduled shutdowns. However, the predictive performance of current models for thermal power plants is limited due to their reliance on single algorithm approaches. Furthermore, there is a shortage of experiments on thermal fired power plant equipment, as most research focuses on nuclear power plants. In this study, we propose a fault predictive stacking approach for a thermal power plant induced draft fan and evaluate the performance of base learners, including Support Vector Machines (SVM), K Nearest Neighbors (KNN), and Random Forests (RF). Our proposed stacking ensemble approach achieved a prediction accuracy of 99.89 % which demostrated superior prediction performance compared to the base methods.

Introduction

Fault prediction models in industrial systems have gained popularity due to their ability to provide predictive results based solely on previously collected data [1]. This development has resulted in their extensive use in day-to-day management and predictive maintenance of modern industrial systems [2]. There are three main strategies that deal with fault prediction which are: statistical, machine learning and model-based approaches. Model- based techniques frequently need familiarity with the monitored equipment, while statistical methods need extensive mathematical background, making machine learning approaches preferable and leading in the data driven fault prediction experiments [3,4]. To date vast research has been conducted on fault prediction, particularly on industrial systems equipment using ma- chine algorithms [5–8]. One area that has been gaining more popularity in industrial systems fault prediction is the power plant industry [9]. Power plant fault prediction and prognosis gained popularity because modern power plants have safety critical systems that contain many feed- back control loops, making identification of faults in these interconnected systems highly complex [10].

Modern power plants also rely on a network of sensors that relay data to supervisory computers to monitor power plant systems and emit huge amounts of data in the process. The large volumes of routinely collected data have led to an increased interest in data-based analysis and prediction for power plants. Numerous machine learning methods have been proposed for fault prediction and diagnosis of power plants. One study [11] experimented with a fault diagnosis method based on lightweight conditional generative adversarial networks (CGAN). Three case studies were performed and indicated that the method could generate high quality multi-class fault samples in small sample scenarios. The experimental results also demonstrated that the method had good diagnostic performance for both rotating machinery and nuclear power plant systems. An SVM based model to diagnose fault of a conventional island pump equipment in a nuclear power plant was proposed [12]. The method was reported to have a high generalization ability, which enabled it to be used as an auxiliary means for fault diagnosis in a conventional island of a nuclear power plant. In another study a fault diagnosis approach for applications in nuclear power plants was

[*] Corresponding author.
E-mail address: tlameloe@bac.ac.bw (T. Emmanuel).

https://doi.org/10.1016/j.jer.2023.10.001
Received 2 May 2023; Received in revised form 24 September 2023; Accepted 1 October 2023
Available online 4 October 2023

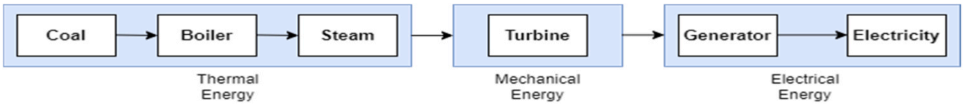

Fig. 1. Thermal power plant schematic.

Table 1
ID fan operating ranges.

Variables	Location	Unit	Operating range
MNDE bearing temperature	2	Â C	17–105
MDE bearing temperature	4	Â C	22.5–80
FNDE bearing temperature	8	Â C	28.7–80
FDE bearing temperature	6	Â C	20.4–80
FNDE vertical vibration	7	mm	0.3–3.5
FDE vertical vibration	5	mm	0.2–3.5
MNDE bearing vibration	1	mm	0.2–3.5
MDE bearing vibration	3	mm	0.2–3.5

introduced [13]. The research introduced a solution where measurements from the power plant were combined with data observed under fault conditions to train the semi supervised classification models. The trained models were applied to new data for fault diagnosis and tested using various fault scenarios on a desktop nuclear power plant simulator as well as on a physical nuclear power plant simulator using a graph-based semi supervised learning algorithm and all faults were effectively diagnosed.

Most of the preceding fault diagnosis methods are based on different predictive models but share one common feature, that they are all based on a single predictive approach. Though some combine algorithms, they are mostly used for data pre-processing and do not directly participate in the diagnosis process [14]. To overcome the limitations of single fault prediction algorithms, an ensemble learning method is adopted in this study. Compared to single prediction algorithms, ensemble methods trains several base models and combines them resulting in higher accuracy and lower variance. The ensemble learning approach has been used to diagnose faults in a variety of fields with good performance [15–17]. Though ensemble methods have been applied to various fields of fault diagnosis, their effectiveness may depend on the specific context of the problem and the characteristics of the data being used. Therefore, this paper presents a comprehensive investigation into the application of ensemble to the domain of power plant maintenance, with a specific focus on induced-draft fan fault prediction. We provide insights into the data collection process and ensemble construction. Additionally, we offer an empirical evaluation of the proposed approach using real-world operational data from a representative power plant.

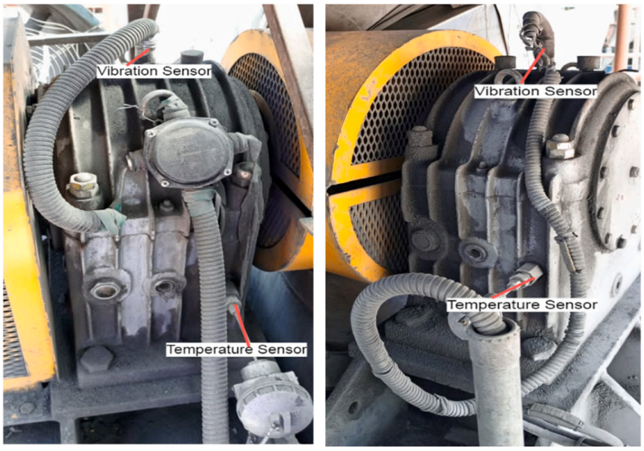

Fig. 2. Fan drive end and fan non drive end.

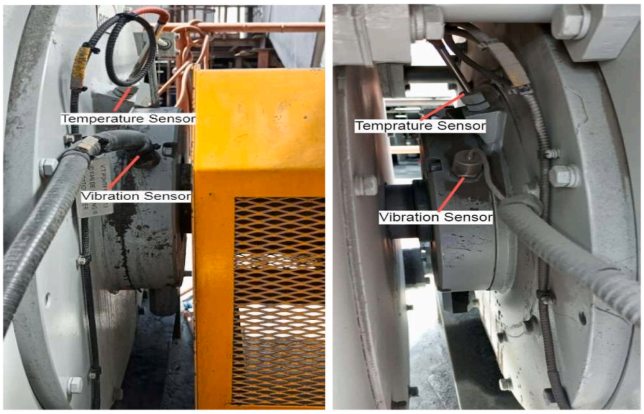

Fig. 3. Motor drive end and motor non drive end.

The paper is structured as follows: Section 2 provides a brief introduction to coal-fired power plants and explains the significance of an induced fan to coal-powered power plants. Section 3 presents an overview of the base predictors used for this study. In Section 4, we discuss the dataset and give a description of the performance measures used. Methods and steps used to address the problem identified by this study are discussed in Section 5. We then present the results of the experiment in Section 6, and we summarize the main findings of the study in Section 7.

Significance of the induced fan in a coal power plant

In this section, we give a brief introduction of a coal fired power plant, and then explain the significance of an induced fan to a coal power plant. Fig. 1 shows the main processes leading to the generation of electricity on a coal plant. The boiler converts coal energy to heat energy, thus turning water into steam and the turbine utilizes the pressure steam to produce electricity.

Induced draft fan

Induced draft fans are auxiliary equipment that emit flue gas generated in the boiler and maintain furnace negative pressure in balanced draft boilers [18]. After passing through the air filter bags, the boiler flue gas goes to the induced draft fan and passed through the discharge duct up to the chimney to the atmosphere. Induced draft fans are mostly used in environments, that include high temperatures, acidic air streams, and other extreme exhaust gases. These conditions typically make Induced draft fans (ID fans) a critical component of operations and the same conditions present risks to the fan's health. For instance vibrations, fly ash or flue gas at high temperatures can damage the fan blades [19,20]. Fig. 4 gives a schematic of a coal-powered plant ID fan.

The induced draft fans are obligated to function safely, soundly and efficiently to enable emission of dispose gases. However, as other industrial systems, they may experience fault such as bearing failure, the bearings that support the fan blades can wear out over time. Eventually, the bearings may seize or break, causing the fan to stop working. The blades of the fans can also become damaged due to wear, corrosion, or impact from foreign objects. This can cause the fan to become unbalanced, leading to increased vibration and noise. The motor that drives the fan can also fail due to electrical or mechanical issues. The impeller, which is the rotating component of the fan that generates the airflow, can also become imbalanced due to wear, corrosion, or damage. Excessive vibration in a fan can lead to failure [21]. As fault incidents are unavoidable, accurate predictive detection is crucial to enhance the safety of day-to-day operations and reduce manufacturing costs. Faults can also result in low system efficiency, unscheduled shutdowns, process interruptions, power generation losses, and high maintenance costs. Therefore, predictive detection of ID fan faults is important since unscheduled shut- downs of the boiler can be caused by faults in the ID fans. Table 1 gives the normal ranges that a coal fired power plant ID fan operates at and Figs. 2 and 3 demonstrate the position of the sensors on the fan bearings and motors on the ID fan.

Background

This section explains the missing value approach and details on base predictors; Random Forest (RF), K Nearest Neighbours (KNN), Support Vector Machines (SVM) and a meta predictor Linear Regression(LN) which are used in this study. Further discussion is made on ensemble approaches and their application in industrial fault detection and prognosis is also discussed.

Imputation

Traditional techniques to imputation include unconditional mean and mode imputation, which handle missing data by using the mode, mean, or median of the available values [22]. A quantitative characteristic or a qualitative attribute of all the non-missing values is used in the imputation procedures to replace each missing values [23]. Due to its ease of use and ability to be used as a quick reference tool, the median method is the one that most research choose to employ [23],[24]. The formula for median imputation can be represented as:

$$\widetilde{x}i = median(x_1, x_2, \ldots, xi - 1, x_{i+1}, \ldots, x_n \quad) \tag{1}$$

where $\widetilde{x_i}$ is the imputed value for the missing value of x_i, and x_1, x_2,., x_n are the observed values of the variable.

Random forest base predictor

Random forest is an ensemble of tree predictors that have a high classification accuracy and strong generalization [25]. The tree structures of the random forest are defined as:

$$\{h(\boldsymbol{x}, \boldsymbol{\Theta}_n)n = 1, \ldots\} \tag{2}$$

where each tree votes for the class at input x that is the most popular, and$\boldsymbol{\Theta}_n$ are independent random vectors. To develop the tree using a training set and $\boldsymbol{\Theta}_n$, a random vector $\boldsymbol{\Theta}_n$that is independent of the preceding random vectors $\boldsymbol{\Theta}_n$,., $\boldsymbol{\Theta}_{n-1}$ is created for the *n*th tree.

Research on machine fault diagnosis using random forest has been emerging due to its fast execution and high accuracy in fault diagnosis. The random forests algorithm was experimented for machine fault diagnosis by combining it with the genetic algorithm to improve the classification accuracy, the approach was compared to other methods and resulted on a higher accuracy [26]. Random forest fault prediction was also experimented by [27] where the authors presented a methodology for predicting fault prone mod- ules. The approach was then compared to logistic regression, discriminant analysis and algorithms in WEKA which is tool that contains a collection of machine learning algorithms for data mining tasks. The random forests proved to be more accurate than the other methods in large data sets. In another research [28], carried out a random forest prediction approach for a nuclear power plant coolant tower.

K nearest neighbors

The KNN algorithm works by identifying the nearest neighbours and use those neighbours for classification. Identification of the nearest neighbors is done by finding the value of K. The test label is selected depending on the labels of these samples. The KNN then computes the distance between samples using a distance measure. The most used is the Euclidean distance as follows:

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \tag{3}$$

Where: The Euclidian distance is $Dist_{xy} \quad =$, data attributes are k, data dimensions $j = 1, 2, 3,., k$, and the value for the property with incomplete data is (X_{ik}), whereas the value for the attribute with complete data is (X_{jk}). The KNN algorithm has proven to provide simple and intuitive approaches for solving a great variety of real-world classification problems [29]. There are approaches that appear in literature that used the KNN for industrial fault prediction. In a study by [30], a fault prediction approach for a high-pressure feed water heater system on a coal-fired power plant simulator was investigated. Firstly, particle swarm optimization algorithm was implemented to generate a minimal set of prototypes, then the KNN algorithm was utilized for classification by using the set of prototypes as reference.

Support vector machines

The SVM algorithm is a widely used classification approach. The SVM searches for the best separating hyper-plane for a labelled training sample, maximizing the distance from the hyper-plane to the closest data points [31]. The SVM algorithm for a classification problem can also be defined as:

$$y(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{w}^T\boldsymbol{x} + b) \tag{4}$$

where $\boldsymbol{x}$ is the input vector, $\boldsymbol{w}$ is the weight vector, b is the bias term, and the sign function, *sgn*, returns 1 if the input is positive and -1 if it is negative. The goal of SVM is to find the optimal values of $\boldsymbol{w}$ and b that maximize the margin between the two classes in the feature space. The optimal hyperplane that separates the two classes is found by solving the optimization problem:

$$\underset{w,b}{armin} \quad \frac{1}{2}\| \boldsymbol{w} \|^2 \text{subject to} y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1\forall_i \tag{5}$$

where y_i is the class label of the *i*-th training example, and $\boldsymbol{x}_i$ is the corresponding feature vector.

Research on the use of the SVM algorithm for prediction is very popular in literature. In a study by, [32], a support vector machine and optimized particle swarm optimization approach was proposed to predict fault on a coolant loop of a nuclear power plant. The enhanced particle swarm optimization was used to improve diagnostic accuracy of the SVM algorithm by optimizing the kernel function and punishment factor. In another research, an SVM based approach for fault prediction on thermal power plant turbines was proposed for prediction of turbine faults. The study produced high accuracy results, however, was not trained when evaluating data mining models [33]. Another fault prediction methodology based on the fusion of the SVM algorithm and adaptive neuro-fuzzy inference system classifiers was proposed, on a

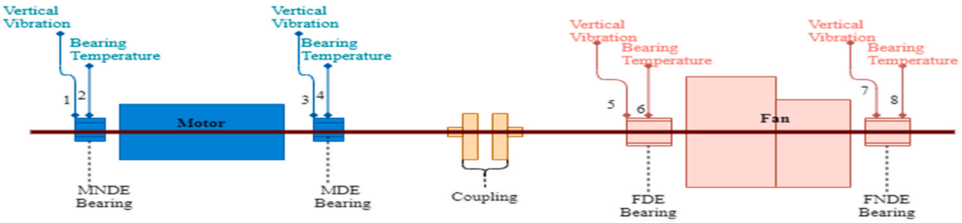

Fig. 4. Induced draft fan schematic.

steam turbine system [34]. he authors in another study [35], proposed an improved approach based on semi-supervised learning algorithm that employed labelled and unlabeled input data for classification using unsupervised optimized self- organizing map and supervised SVM.

Logistic regression

Logistic regression is a method that makes prediction by analyzing dependent and independent variables by building multi-nominal regression model [36]. The estimator is flexible and less prone to over-fitting [37]. Logistic regression can be defined by a linear equation as in Eq. (6) [38]:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2+X_2+...+\beta_n+X_n)}} \tag{6}$$

Where, P $(y = 1|x)$ represents the probability of the dependent variable y taking the value of 1 given the independent variables x_0, x_1, x_2, $...x_n$, r$\beta_0, \beta_1, \beta_2, ...\beta_n$ are the coefficients of the independent variables in the model. The logistic function is represented by $e^{-(\beta_0+\beta_1 X_1+\beta_2+X_2+...+\beta_n+X_n)}$which converts the linear combination of the independent variables to a probability between 0 and 1.

There are studies in literature where logic regression was used for fault pre- diction in power plants, [39] merged the logistic regression with the SVM algorithm for fault prediction in a nuclear power plant. In particular, the regression algorithm is direct and fast while it also needs to handle the abnormal value.

Stacking

Stacking is an ensemble strategy that utilize multiple models to produce a single, improved outcome. Typically, this approach yields results that are more accurate than those of a single algorithm. This has been the case in several machine learning competitions, where the triumphant models used ensemble techniques [40]. Ensemble approaches like stacking work best when the maximum level of accuracy is needed [41]. To develop an ensemble that is as varied as feasible, a plan must be in place before an ensemble is formed. This is so that the optimal ensemble approach heavily relies on the problem being handled [42]. When stacking, a meta-level classifier is used to compute the predictions from several models as input, and the result of this meta classifier is the final prediction [43]. The meta-level algorithm and the best features make up the bulk of stacking [44]. Stacking algorithm can also be defined as:

$$\widehat{y}stack = y\sum j = 1^M w_j \widehat{y}_j \tag{7}$$

where $\hat{y}stack$ is the stacked prediction, $y\hat{j}$ is the prediction of the j^{th} base model, M is the total number of base models, w_j is the weight assigned to the j^{th} base model, and γ is the scaling factor used to ensure that the weights sum to one.

Stacking has been shown to perform similarly to selecting the ensemble's best classifier by cross-validation. Using the base classifier outputs as inputs in the new space, stacking combines generalizations rather than selecting one out of several generalizations. Then, using this new space, predictions are made. In the context of base-level classifiers produced for various learning processes, stacking is seen as an ensemble for future investigation [45]. This ensemble learning methods also ensures diversity among the base models, improving prediction accuracy [46].

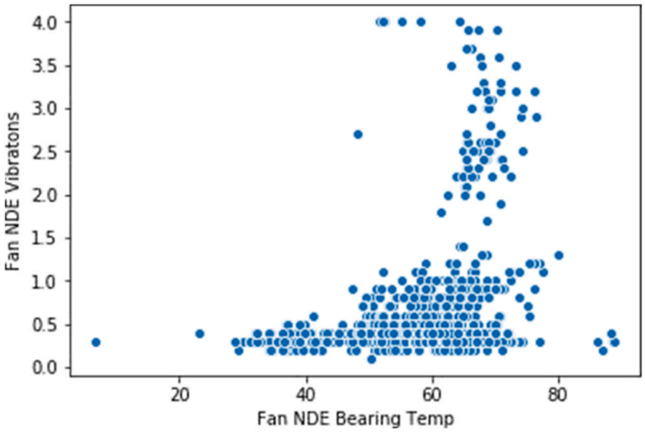

Fig. 5. Actual sample of FNDE vibrations and temperature.

Table 2
ID fan data 10/06/2021.

Variables	Unit	0:00	04:00	08:00	12:00	16:00	20:00
MNDE bearing temperature	Â C	23.1	21.3	19.7	29.9	34.1	29
MDE bearing temperature	Â C	39.7	38.4	35.5	46.2	50.9	47.1
FNDE bearing temperature	Â C	53.9	51.9	50.6	57.6	61.6	58.1
FDE bearing temperature	Â C	47.3	47.2	47.6	57.4	56.1	52.8
FNDE vertical vibration	mm	0.4	0.4	0.4	0.2	0.4	0.4
FDE vertical vibration	mm	0.4	0.4	0.4	0.5	0.4	0.4
MNDE bearing vibration	mm	0.6	0.4	0.6	0.5	0.5	0.7
MDE bearing vibration	mm	0.6	0.5	0.3	0.5	0.5	0.6

Experimental setup

The experiment was done in a local power plant where sensors on the induced draft fan captured day to day readings. The ID fan which is part of a coal powered plant that produces a target output of 117 W runs at an air flow rate of 5365 Am^3/min, with a density of 0.85 kg/m^3, diameter of

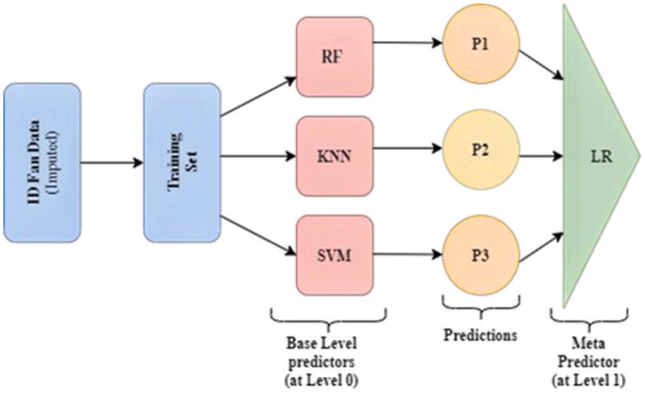
ID Fan Data (Imputed)
Training Set
RF
KNN
SVM
P1
P2
P3
LR
Base Level predictors (at Level 0)
Predictions
Meta Predictor (at Level 1)

Fig. 6. Fault prediction model.

2440 *mm*, with a total pressure of 505 *mmAq* and a speed of 989 *rpm*. The fan impeller is driven by a 650 kW electric induction motor (6600 V 50 Hz 3 Phase) with a speed of 1000 rpm. A schematic of the induced draft fan used in this study is illustrated in Fig. 4 and the normal operating ranges in Table 1. The fan's multiple sensors located on the ID fan non-drive end (FNDE) and fan drive end (FDE), fan motor were used in the experiment data acquisition.

Data description

The dataset contains readings from March 2020 until June 2021, of a single unit of the power plant consisting of 2422 samples. The dataset primarily centers around two key operational parameters critical for assessing the health and performance of the induced-draft fan: temperature and vibrations. These parameters are measured at strategic points within the fan-motor assembly, each offering valuable insights into the equipment's condition. Bearing Vibrations: Vibrational measurements at both the fan's non-drive end (FNDE) and drive end (FDE) bearings are recorded. These measurements are fundamental indicators of bearing wear and potential defects, providing early warning signs of mechanical issues that could compromise fan performance. Bearing Temperatures: Temperature measurements at the FNDE and FDE bearings are included. Bearing temperature is a critical parameter for assessing lubrication effectiveness and detecting anomalies related to overheating or inadequate cooling. Motor Vibrations: Vibrational data is collected at both the motor's non-drive end (MNDE) and drive end (MDE). Motor vibrations are indicative of imbalances, misalignments, or structural issues within the motor assembly, all of which can adversely affect fan operation. Motor Temperatures: Temperature measurements at the MNDE and MDE provide insights into the motor's thermal performance and can signal potential issues such as overheating or insulation degradation. The ID fan temperature and vibrations are measured by sensors and recorded by the technicians every 4 h when the plant was running. Figs. 2 and 3 demonstrate the position of the sensors on the fan bearings and motors which the values are recorded. These values are captured every 4 hrs daily by recording them into files that will enable technicians to be able to identify any anomalies that may occur in the data see Fig. 2. We also give a visual representation of the data used on the experiment in Fig. 5 and (Table 2).

Performance measures

Different performance measures including accuracy, precision, recall and F1-score were used for analysis. Accuracy measure was used to rate the error of correct (or incorrect) predictions made by the model. The method tested how close predictions are to the target. While recall score is the number of correctly diagnosed samples in the actual samples, precision score is the percentage of correctly diagnosed samples in the predicted samples. A classifier's precision and recall are combined into a single parameter known as the F1-score [47]. The precision score, recall score, and F1-score all have a range of 0.0–1.0. The higher the F1-score demonstrates a better over- all performance for the model. Another method called the train-test split was used to ensure that the test and train datasets are representative of the overall dataset. The 60/40 split was used to evaluate the model. The split was chosen in particular because it achieves a balance between having enough data to effectively train the model and enough data to reliably assess the model's performance. The model has more data to learn from the 60 % training set and can potentially achieve better performance. Meanwhile, the test set 40 % is large enough to provide a good estimate. The definitions of accuracy, precision, recall and F1score are represented in Eqs. (8, 9, 10 and 11):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{11}$$

Where TN is a true negative indicating that the algorithm correctly classified the negative class as negative, TP true positive indicates that the algorithm correctly predicted the positive class, False positives are denoted by FP and false negatives by FN, respectively, when the algorithm predicts the wrong class as positive or negative.

Proposed model

The employed model could be divided into steps described as below on Fig. 6 and Algorithm 1. The first step of our model was to identify the monitoring parameters which are related to ID fans and explored the dataset to understand its structure, size, content, missing values and duplicate records. The parameters were identified and discussed on Table 1 and empty cells were identified. The empty cells were replaced with NaN, which stands for ”Not a Number” to indicate missing or undefined data. The missing values were then replaced by using an imputation method discussed in 3.1. The missing values were handled using the median approach because of its simplicity [48]. Step 7–12 outlines the stacking algorithm, three base learners namely, SVM, KNN, RF and a meta learner LRN were implemented. Finally, the effectiveness of the built stacking classifier was evaluated.

Fundamentally, the aim of using a stacking ensemble was to generate a more reliable and accurate prediction than that produced with single prediction-based methods. The base predictors were firstly trained on a training set, the meta-model was then trained on the results of the base predictors as features. A weighted combination of predictions from the base predictors was used as input features for the meta-learner. The base level algorithms included a set of various learning algorithms. The algorithms were trained using induced fan and a meta level algorithm logic regression was trained for the final prediction.

Algorithm 1. Staking method.

```
Input: Coal Power Plant Induced Draft Fan Data ($D$)
Output: Fault Prediction
Step 1: Handle missing values in $D$ using Median imputation.
Step 2: Split $D$ into $D_T$ for training and $D_S$ for testing using the train-testsplit of 60/40.
Step 3: For each fold $i$ from 1 to $k$:
    Train a base model SVM, KNN, RF, on all folds of $D_T$ except for fold $i$.
    Predict the target values for the test instances in fold $i$ using basemodels $SVM$, $KNN$, $RF$.
Step 5: Combine the predictions of all base models SVM, KNN, RF into a single matrix $X$, where each row corresponds to a test instance and each column corresponds to a base model.
Step 6: For each test instance, calculate the weighted average of the pre- dictions from all base models, where the weights are the performance of each base model on the validation set fold $i$.
Step 7: Apply the meta model LR to the entire training set $D_T$ and use it to predict the target values for $D_S$.
Step 8: Evaluate the performance of the stacked model using accuracy, precision, recall and F1-score.
```

Results and discussion

To address the potential negative impacts of the missing values in the ID fan dataset, a median approach was used to prepare the data by replacing the missing values before prediction. This step was taken mainly to reduce bias in analysis and prevent the loss of valuable information that may be caused by discarding the missing data. Afterward, base algorithms were implemented individually since their diagnostic results may differ significantly, and it was important to compare them with the stacking approach. The base algorithms experiment started with the KNN algorithm, firstly the consideration of the value of K was implemented, this was based solely on experimental results starting with $K = 1$ and stopped at $K = 5$, where the best results were achieved. Then the SVM algorithm using the linear kernel function was then implemented followed by the RF where we focused on the number of decision trees in the forest for hyper-parameters for the induction of the forests. The Gini index which demonstrates a degree of inequality in a distribution, was also used as an evaluation index for a single tree, and the results showed that the diagnostic results increased when the Gini index was used. Finally, the ensemble approach was implemented after optimizing the base models and evaluating them. The predictive results of the ensemble and the base algorithms are summarized on Figs. 7, 8, 9, 10

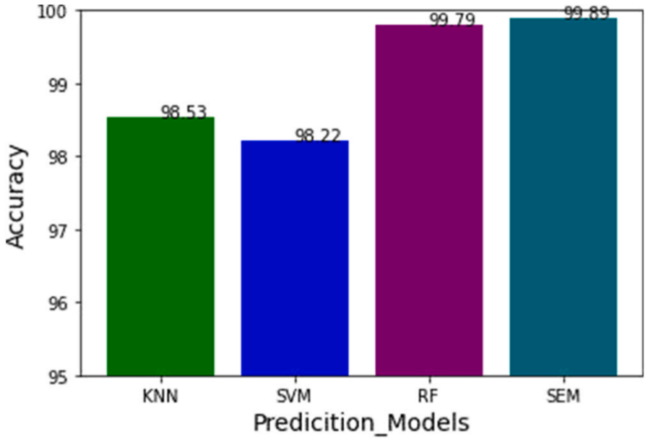

Fig. 7. Comparison of accuracy between the used algorithms.

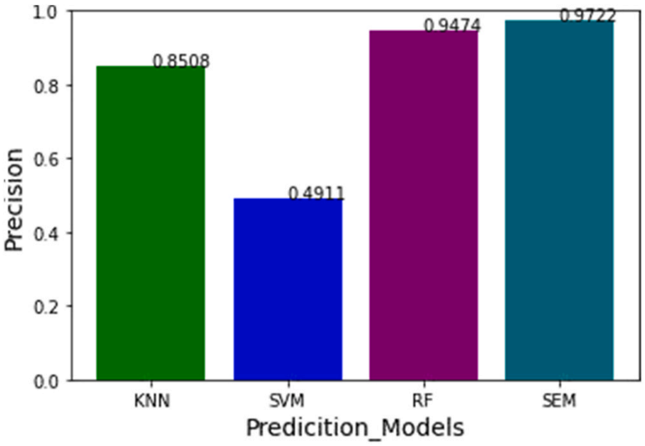

Fig. 8. Comparison of precision between the prediction models.

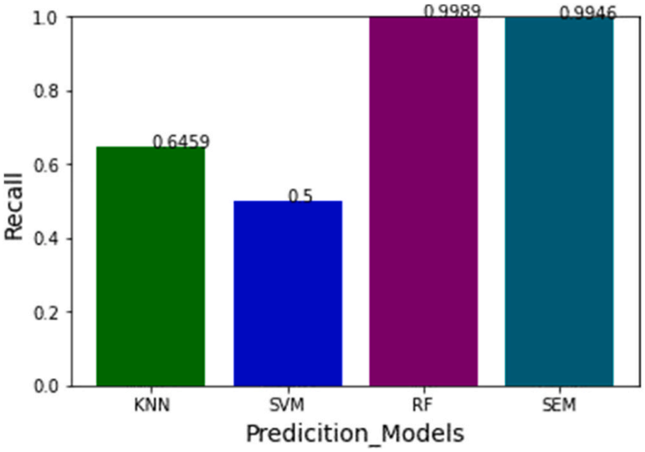

Fig. 9. Comparison of recall between the prediction models.

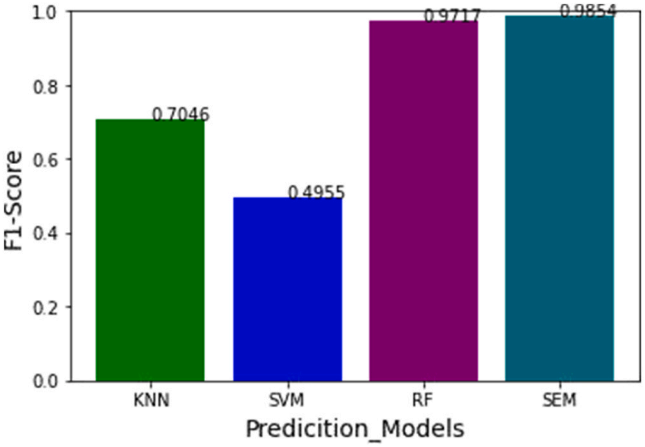

Fig. 10. Comparison of F1-Score between the prediction models.

and Table 3. The results of the experiment demonstrates that the accuracy's of the predictive algorithms were all greater than 98.2 % and the stacking method achieved an accuracy of 99.89 which was higher than that of the individual learners. The precision and F1-Score for the base learners also demonstrated to be lower than that of the stacked algorithm. However, the RF and demonstrated to have a good overall performance with the RF performing better than all the other algorithms including the stacking algorithm in recall. RF exceeded stacking by in

Table 3
ID fan data 10/06/2021.

Methods	Accuracy	F1 Score	Precision	Recall
KNN	98.53	0.7046	0.8508	0.6459
SVM	98.22	0.4953	0.4911	0.5
RF	99.79	0.9217	0.9474	0.9989
SEM	99.89	0.9854	0.922	0.9946

recall by 00.0043 and stacking exceeded KNN and RF by 1.36 % and 0.1 % in accuracy, and 0.1214 and 0.0248 in recall, 0.2808 and 0.0137 respectively, which demonstrates that the predictive performance was improved by using the stacking ensemble.

Conclusions

In this study, a novel data-driven fault prediction methodology was pro- posed using a stacking ensemble method applied to induced draft fan sensor data. The results showed that the stacked approach outperformed the SVM, KNN, and RF algorithms in terms of accuracy, precision, recall, and F1-Score for fault prediction. While the proposed method was specifically designed for induced draft fan faults, future research could explore applying the stacking approach to predict faults in other power plant machinery. Additionally, since the median method was used to handle missing values, further research could investigate novel approaches for handling missing data in the ID fan datasets.

Funding

This work is supported by Botswana International University of Science and Technology, Palapye, Botswana Grant ID: S00116.

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.

References

[1] R.-E. Precup, P. Angelov, B.S.J. Costa, M. Sayed-Mouchaweh, An overview on fault diagnosis and nature-inspired optimal control of industrial process applications, Comput. Ind. 74 (2015) 75–94.

[2] H. Luo, S. Yin, T. Liu, A.Q. Khan, A data-driven realization of the control-performance-oriented process monitoring system, IEEE Trans. Ind. Electron. 67 (2019) 521–530.

[3] N. Amruthnath, T. Gupta, A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance, in: Proceedings of the Fifth International Conference on Industrial Engineering and Applications (ICIEA), IEEE, 2018, 355–361.

[4] M. Fernandes, A. Canito, V. Bol´on-Canedo, L. Concei¸c˜ao, I. Pra¸ca, G. Marreiros, , Data analysis and feature selection for predictive maintenance: a case-study in the metallurgic industry, Int. J. Inf. Manag. 46 (2019) 252–262.

[5] S. Biswal, G. Sabareesh, Design and development of a wind turbine test rig for condition monitoring studies. in: Proceedings of the International Conference on Industrial Instrumentation and Control (ICIC), IEEE, 2015, pp. 891–896.

[6] R.S. Peres, A.D. Rocha, P. Leitao, J. Barata, Idarts–towards intelligent data analysis and real-time supervision for industry 4.0, Comput. Ind. 101 (2018) 138–146.

[7] N. Kolokas, T. Vafeiadis, D. Ioannidis, D. Tzovaras, Forecasting faults of industrial equipment using machine learning classifiers. in: Proceedings of the Innovations in Intelligent Systems and Applications (INISTA), IEEE, 2018, pp. 1–6.

[8] R. Liu, B. Yang, E. Zio, X. Chen, Artificial intelligence for fault diagnosis of rotating machinery: a review, Mech. Syst. Signal Process. 108 (2018) 33–47.

[9] J. Ma, J. Jiang, Applications of fault detection and diagnosis methods in nuclear power plants: a review, Prog. Nucl. Energy 53 (2011) 255–266.

[10] B. Upadhyaya, K. Zhao, B. Lu, Fault monitoring of nuclear power plant sensors and field devices, Prog. Nucl. Energy 43 (2003) 337–342.

[11] G. Qian, J. Liu, Fault diagnosis based on conditional generative adversarial networks in nuclear power plants, Ann. Nucl. Energy 176 (2022), 109267.

[12] H. Li, N.-W. Lan, X.-n. Huang, Application research of fault diagnosis in conventional island of nuclear power plant based on support vector machine, in: in: Proceedings of the Nuclear Power Plants: Innovative Technologies for Instrumentation and Control Systems: The Fourth International Symposium on Software Reliability, Industrial Safety, Cyber Security and Physical Protection of Nuclear Power Plant (ISNPP), Springer, 2020, pp. 304–312.

[13] J. Ma, J. Jiang, Semisupervised classification for fault diagnosis in nu- clear power plants, Nucl. Eng. Technol. 47 (2015) 176–186.

[14] J. Li, M. Lin, Ensemble learning with diversified base models for fault diagnosis in nuclear power plants, Ann. Nucl. Energy 158 (2021), 108265.

[15] J. Yang, G. Xie, Y. Yang, An improved ensemble fusion autoencoder model for fault diagnosis from imbalanced and incomplete data, Control Eng. Pract. 98 (2020), 104358.

[16] C. Kapucu, M. Cubukcu, A supervised ensemble learning method for fault diagnosis in photovoltaic strings, Energy 227 (2021), 120463.

[17] H. Han, Z. Zhang, X. Cui, Q. Meng, Ensemble learning with member optimization for fault diagnosis of a building energy system, Energy Build. 226 (2020), 110351.

[18] M. Bhowmick, S. Bera, Study the performances of induced fans and design of new induced fan for the efficiency improvement of a thermal power plant. in: Proceedings of the IEEE Region 10 and the Third international Conference on Industrial and Information Systems, IEEE, 2008, pp. 1–5.

[19] Y. Wang, H. Tan, K. Dong, H. Liu, J. Xiao, J. Zhang, Study of ash fouling on the blade of induced fan in a 330 mw coal-fired power plant with ultra-low pollutant emission, Appl. Therm. Eng. 118 (2017) 283–291.

[20] J. Du, J. Liang, L. Zhang, Research on the failure of the induced draft fan's shaft in a power boiler, Case Stud. Eng. Fail. Anal. 5 (2016) 51–58.

[21] B. Bhandari, Handbook of Industrial Drying, A.S. Mujumdar (Ed.), CRC Press, Boca Raton, FL, 2015. isbn: 978–1-4665–9665-8, 201.

[22] P.J. García-Laencina, J.-L. Sancho-Gomez, A.R. Figueiras-Vidal, M. Verleysen, K nearest neighbours with mutual information for simul- taneous classification and missing data imputation, Neurocomputing 72 (2009) 1483–1493.

[23] J.M. Jerez, I. Molina, P.J. García-Laencina, E. Alba, N. Ribelles, M. Martín, L. Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artif. Intell. Med. 50 (2010) 105–115.

[24] M. Adelantado-Renau, D. Moliner-Urdiales, I. Cavero-Redondo, M.R. Beltran-Valls, V. Martínez-Vizcaíno, C. Alvarez-Bueno, Association be- tween screen media use and academic performance among children and adolescents: a systematic review and meta-analysis, JAMA Pediatr. 173 (2019) 1058–1067.

[25] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[26] B.-S. Yang, X. Di, T. Han, Random forests classifier for machine fault diagnosis, J. Mech. Sci. Technol. 22 (2008) 1716–1725.

[27] L. Guo, Y. Ma, B. Cukic, H. Singh, Robust prediction of fault-proneness by random forests. in: Proceedings of the Fifteenth International Symposium on Software Relia- Bility Engineering, IEEE,, 2004, pp. 417–428.

[28] S. Sharanya, R. Venkataraman, Empirical analysis of machine learning algorithms in fault diagnosis of coolant tower in nuclear power plants, in: in: Proceedings of the International Conference On Computational Vision and Bio Inspired Computing, Springer, 2018, pp. 1325–1332.

[29] G. Chatzigeorgakidis, S. Karagiorgou, S. Athanasiou, S. Skiadopoulos, Fml-knn: scalable machine learning on big data using k-nearest neighbor joins, J. Big Data 5 (2018) 1–27.

[30] X.-X. Wang, L.-Y. Ma, A compact k nearest neighbor classification for power plant fault diagnosis, J. Inf. Hiding Multimed. Signal. Process. 5 (2014) 508–517.

[31] S. Yin, X. Gao, H.R. Karimi, X. Zhu, Study on support vector machine- based fault detection in tennessee eastman process, in: Abstract and Applied Analysis, 2014, Hindawi, 2014.

[32] Z.-Y. Wang, C. Lu, B. Zhou, Fault diagnosis for rotary machinery with selective ensemble neural networks, Mech. Syst. Signal Process. 113 (2018) 112–130.

[33] K.-Y. Chen, L.-S. Chen, M.-C. Chen, C.-L. Lee, Using svm based method for equipment fault detection in a thermal power plant, Comput. Ind. 62 (2011) 42–50.

[34] K. Salahshoor, M. Kordestani, M.S. Khoshro, Fault detection and diagnosis of an industrial steam turbine using fusion of svm (support vector machine) and anfis (adaptive neuro-fuzzy inference system) classifiers, Energy 35 (2010) 5472–5482.

[35] K. Moshkbar-Bakhshayesh, S. Mohtashami, Classification of npps transients using change of representation technique: a hybrid of unsupervised msom and supervised svm, Prog. Nucl. Energy 117 (2019), 103100.

[36] G. Hu, T. Zhou, Q. Liu, Data-driven machine learning for fault detection and diagnosis in nuclear power plants: a review, Front. Energy Res. 9 (2021) 185.

[37] M. Hill, P. Connolly, P. Reutemann, D. Fletcher, The use of data mining to assist crop protection decisions on kiwifruit in new zealand, Comput. Electron. Agric. 108 (2014) 250–257.

[38] P. Radhakrishnan, K. Ramaiyan, A. Vinayagam, V. Veerasamy, A stacking ensemble classification model for detection and classification of power quality disturbances in pv integrated power network, Measurement 175 (2021), 109025.

[39] A. Ayodeji, Y.-k Liu, H. Xia, Knowledge base operator support system for nuclear power plant fault diagnosis, Prog. Nucl. Energy 105 (2018) 42–50.

[40] C. Zhang, Y. Ma, Ensemble Machine Learning: Methods and Applications, Springer, 2012.

[41] M. Whitehead, L. Yaeger, Sentiment mining using ensemble classification models, in: in: Innovations and Advances in Computer Sciences and Engineering, Springer, 2010, pp. 509–514.

[42] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (2006) 21–45.

[43] Y. Chen, M.-L. Wong, H. Li, Applying ant colony optimization to configuring stacking ensembles for data mining, Expert Syst. Appl. 41 (2014) 2688–2702.

[44] C.C. Aggarwal, Data Classification: Algorithms and Applications, CRC Press, 2014.

[45] S. Dzeroski, B. Zenko, Is combining classifiers better than selecting the best one?, in: in: ICML 2002 Citeseer, 2002, 123e30.

[46] J.J. Rodriguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: a new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1619–1630.

[47] D.M. Powers, What the f-measure doesn't measure: features, flaws, fallacies and fixes, arXiv Prepr. arXiv 1503 (2015) 06410.

[48] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, J. Big Data 8 (2021) 1–37.