Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Probabilistic occupancy forecasting for risk-aware optimal ventilation through autoencoder Bayesian deep neural networks

Chaoqun Zhuang\textsuperscript{a,b,*}, Ruchi Choudhary\textsuperscript{a,b}, Anna Mavrogianni\textsuperscript{c}

\textsuperscript{a} Data-centric Engineering, The Alan Turing Institute, London, United Kingdom
\textsuperscript{b} Energy Efficient Cities Initiative, Department of Engineering, University of Cambridge, Cambridge, United Kingdom
\textsuperscript{c} Institute for Environmental Design and Engineering, Bartlett Faculty of the Built Environment, University College London, London, United Kingdom

\textbf{A B S T R A C T}

Ventilation plays a noteworthy role in maintaining a healthy, comfortable and energy-efficient indoor environment and mitigating the risk of aerosol transmission and disease infection (e.g., SARS-COV-2). In most commercial and office buildings, demand-controlled ventilation (DCV) systems are widely utilized to conserve energy based on occupancy. However, as the presence of occupants is often inherently stochastic, accurate occupancy prediction is challenging. This study, therefore, proposes an autoencoder Bayesian Long Short-term Memory neural network (LSTM) model for probabilistic occupancy prediction, taking account of model misspecification, epistemic uncertainty, and aleatoric uncertainty. Performances of the proposed models are evaluated using real data in an educational building at the University of Cambridge, UK. The models trained on data of one open-plan space are used to predict occupant numbers for other spaces (with similar layout and function) in the same building. The probabilistic occupant profiles are then used for estimating optimal ventilation rates for two scenarios (i.e., normal DCV mode for energy conservation and anti-infection mode for virus transmission prevention). Results show that, during the test period, for the 1-h ahead prediction, the proposed model achieved better performance with up to 5.8\% mean absolute percentage error reduction than the traditional LSTM model. More flexible alternatives for ventilation can be offered by the proposed risk-aware decision-making schemes serving different purposes under real operation. The findings from this study provide new occupancy forecasting solutions and explore the potential of probabilistic decision making for building ventilation optimization.

\section{1. Introduction}

\subsection{1.1. Background}

Ventilation plays an essential role in building energy conservation, improvement of indoor air quality as well as prevention of respiratory disease (such as syndrome caused by a SARS-COV-2). To achieve acceptable indoor environmental control with minimal operation costs, existing standards such as ASHRAE 62.1–2016 \cite{1} and CEN EN15251-2007 \cite{2} prescribe the effective minimum ventilation rates to safeguard acceptable indoor air quality based on the occupancy estimation. Demand-controlled ventilation (DCV) \cite{3} has been introduced to conserve energy by adjusting ventilation rates in response to changes in occupancy or indoor pollutant concentration. In the continuing COVID-19 pandemic, infection control as a new ventilation control objective should be considered to reduce the transmission of the virus.

As the prescribed outdoor airflow rate of public buildings by current standards is insufficient to control the infection risk at a low level \cite{4}, it is essential to develop ventilation control schemes so as to reduce the probability of infection based on occupancy levels.

Currently, various industry associations and public health organizations have issued ventilation guidelines in response to COVID-19, as listed in Table 1. In summary, to reduce the cross-infection of airborne transmitted diseases, the currently accepted approach is to increase the ventilation rate using large amounts of fresh air and extend the ventilation periods to dilute/remove contaminants based on occupancy levels. Simple rule-based ventilation control strategies are widely used in practical applications. For instance, the ventilation systems are suggested to operate at least 2 h with the maximum volume before and after occupied times suggested by most guidelines. Furthermore, researchers have developed real-time ventilation control strategies to reduce the risk of respiratory infection. Sha et al. \cite{5} determined the optimal ventilation rates based on a modified Wells-Riley model to fulfill the requirements of...
both dilution ventilation and ventilative cooling. Wang et al. [6] proposed a metabolism-based ventilation control method based on occupants’ activity intensity estimation, reducing the infection probability down to 4.3–6.3% while saving 13% of energy compared with fixed-fresh-air ventilation. These control methods determine the ventilation rates based on the real-time collection of occupancy data. However, since the fresh airflows supplied to occupied spaces are coupled with the ventilation duct network, in practice, it usually takes a relatively long time to make airflows achieve their desired values, resulting in a delay in pollutant dilution and removal [7]. Predicting future occupancy profiles could be a promising approach for achieving proactive ventilation operation. However, due to the stochastic nature of occupant presence and the inherent variability in behaviour amongst individuals, accurate occupancy prediction is challenging.

Over the past decade, forecast algorithms have significantly improved the accuracy of occupancy predictions. Generally, forecast methods can be broadly classified into four groups: conventional statistical approaches, unsupervised machine learning approaches, supervised machine learning approaches, and hybrid approaches [14]. The mainstream statistical approaches are Markov chain-based and recursive models. For instance, Candanedo et al. [15] utilized the hidden Markov models to define occupancy schedules based on temperature, humidity, light, and occupancy measurements. Salimi et al. [16] developed an inhomogeneous Markov chain model to extract occupancy information with varying time steps from collected occupancy data. Burak et al. [17] adopted a recursive occupancy learning algorithm based on maximum likelihood estimation to update the estimate of occupant profiles.

More recently, machine learning-based models began to dominate the literature, replacing traditional statistical models [18]. As typical unsupervised models, the clustering algorithms, such as k-means, k-nearest neighbour techniques and support vector clustering, were introduced to identify occupancy patterns in both residential [19, 20] and commercial buildings [21, 22]. The cluster results could further represent different occupancy patterns, distinguishing weekdays from weekends. For supervised models, studies have analysed the forecast performance of gradient boosting [23], support vector regression [24], decision tree [25], random forest [26], and deep neural networks [27]. Particularly, Long Short-term Memory (LSTM), one of the Recurrent Neural Network architectures, is well designed for time-series prediction due to its ease of incorporating exogenous variables, and automatic feature extraction abilities. Kim et al. [28] developed an LSTM deep learning approach for predicting real-time occupancy in a large exhibition hall with a prediction accuracy of over 84%. Qolomany et al. [29] found that the LSTM-based occupancy forecasting method significantly reduced the root mean square error (RMSE) by 80.9%–93.4% in a commercial building compared with Auto Regression Integrating Moving Average models. Normally, most existing occupancy forecasting models derive point forecasts based on the input of continuous historical occupancy data and contextual information, such as hours, weekdays,
temperature and CO$_2$ environmental monitoring variables, such as sound, relative humidity, air velocity, electric loads, and outside temperature, have been shown to indicate the occupant density level [32]. For instance, electrical loads, energy consumption, and environmental data are considered as an alternative to building usage time and switch to a lower speed 2 h after the building usage time. Keep the ventilation on 24/7, with lowered ventilation rates when people are absent.

### Table 1: Building ventilation system guidelines to reduce the risk of COVID-19.

| Source                  | Ventilation control requirements                                                                 |
|-------------------------|---------------------------------------------------------------------------------------------------|
| CIBSE [3]               | o Maintain CO$_2$ concentrations below 800 ppm (10–15 l/s/person)                                  |
| REHVA [9]               | ✓ Start ventilation at the nominal speed at least 2 h before the building usage time and switch to a lower speed 2 h after the building usage time. |
| WHO [12]                | ✓ Start ventilation at the minimum mechanical ventilation rate higher than 10 l/s/person          |
| CDC [13]                | ✓ Increase the introduction of outdoor air by opening outdoor air dampers beyond minimum settings  |
| CCI AQ [11]             | ✓ Operation of ventilation systems should be beyond the hours of occupancy (2 h before and after) |
| ASHRAE [10]             | ✓ Provide more outdoor air than the current maximum (i.e., preferably 100% outdoor air and exhaust) |
| CIBSE/REHVA/ASHRAE/WHO   | ✓ Rebalance or adjust HVAC systems to increase total airflow to occupied spaces when possible  |
|                         | ✓ Turn off any DCV controls                                                                       |
|                         | ✓ Run the HVAC system at maximum outside airflow for 2 h before and after the building is occupied |

and holidays. However, deterministic models often neglect uncertainty inherent to the non-deterministic occupant behaviour and thus fail to derive variances of occupancy predictions, especially when using an incomplete dataset for model training. Forecasting occupancy in a probabilistic manner with uncertainty bound or confidence interval enables the decision-makers to make risk-informed building energy management and ventilation operations [30].

Another challenge encountered is that some buildings or zones do not install sensors to detect occupancy but need energy-efficient ventilation control. In such buildings, the ventilation rates are calculated based on maximum occupancy conditions [31], resulting in over-ventilation and energy waste during low-occupancy periods. Since data on occupant behaviour are hard to collect and privacy considerations may preclude making the data widely available, occupant-related energy and environmental data are considered as an alternative to indicate the occupant density level [32]. For instance, electrical loads, such as plug-in equipment and lighting loads, have been shown to exhibit a strong relationship with building occupancy. It is reported that the R-squared values between plug loads and the occupant counts by linear regressions could achieve 0.65–0.78 [33,34]. In addition, environmental monitoring variables, such as sound, relative humidity, air temperature, and CO$_2$ can also be used as the exogenous variables to train the forecasting model [35]. Due to the increasing penetration of smart meters in the residential sector [36] and mandatory requirements for smart sub-metering systems in the commercial sector regulated by building codes [37], occupant-related energy data are more feasible to obtain. In the context of occupancy prediction, it is possible to conduct the occupancy forecasting for the zones without explicit occupant counts, by taking advantage of the rich data/knowledge obtained from other zones that have similar activities and applying transfer learning [38] or model transferrable methods [39].

### 1.2. Problem definition and research objectives

Based on the literature review, the following observations were made: (1) As an important feature of occupancy, building electrical load (i.e., plug and lighting electrical loads) should not be ignored in occupancy prediction. (2) Predicting the occupancy in a probabilistic manner contributes to the risk-aware ventilation decision making, due to the random nature of occupant behaviour. (3) Different operation decision schemes are required in the face of different ventilation purposes (e.g., fulfilling the requirement of thermal comfort and indoor air quality under non-pandemic periods or reducing the risk of infection under pandemic periods). The key research questions of this study are:

- Which factors noticeably affect the occupancy prediction in buildings? Which features should be chosen to predict occupancy?
- How are the different uncertain sources taken into account in model development? Could the developed models be transferrable to other zones with similar functions and activities?
- How are the predicted probabilistic occupancy profiles utilized for risk-aware ventilation decision making?

To address the above-mentioned research gaps, this study develops an autoencoder Bayesian Long Short-Term Memory Network (ABLSTM) model for probabilistic occupancy prediction, enabling the risk-aware decision-making schemes for optimal ventilation under pandemic and non-pandemic periods based on quantified occupant profiles. To verify the proposed strategy, case studies were conducted in an educational building located at the University of Cambridge, UK. The performance and the transferability of the proposed models were investigated and compared with conventional deep neural networks.

The main contributions of this study are briefly summarized as follows: 1) An autoencoder Bayesian deep neural network model is developed for probabilistic occupancy-based ventilation operation, which can take account of different uncertain sources (i.e., model misspecification, aleatoric and epistemic uncertainties). 2) The transferability of the probabilistic model framework is tested and validated for zones (similar activities) with and without explicit occupancy counts. 3) Risk-aware decision-making schemes (i.e., normal DCV and anti-infection modes) are provided for different ventilation requirements under pandemic and non-pandemic periods.

### 2. Methodology

#### 2.1. Outline of probabilistic occupancy forecasting method for optimal ventilation

The outline of the proposed probabilistic occupancy forecasting method for optimal ventilation is shown in Fig. 1, which consists of four major tasks to be addressed. The first task is to pre-process the datasets (e.g., occupant counts, plug loads and lighting loads), including data cleaning, data transformation and normalization. The second task is to build an autoencoder Bayesian LSTM (ABLSTM) framework, using the information currently available (at timestamps $t$, $t-1$, $t-2$ ... ) to forecast the occupant number for the future timestamps ($t+1$, $t+2$ ... ) in a probabilistic manner. Two ABLSTM models, suitable for zones with and without explicit occupancy counts respectively, are built to forecast the occupant number at different sequential input features. The third task involves the test and validation of the model transferability. The transferability test is performed, in which the models trained by one floor of a building are applied to new data from the other two floors of the same building. The fourth task is to make the risk-aware decision-making for ventilation operations based on the forecasted occupant profiles. The probabilistic decision-making schemes could be applied in two ventilation operation modes: 1) normal DCV mode under non-pandemic periods and 2) anti-infection mode under pandemic periods. The details of each task are illustrated in Sections 2.2 to 2.5, respectively.
2.2. Data collection and pre-processing

2.2.1. Descriptions of test building and datasets

An educational building located at the University of Cambridge, UK was selected as the testbed for this study. The data were collected from the first to third floors with similar layouts, in which the primary use is open-plan offices. The view of the building and its first-floor plan is presented in Fig. 2. The design occupant capacity was 50 for each floor. Recessed LED modular luminaires with daylight-linked dimming, presence detection and manual override were controlled by an infrared handheld controller. The data for hourly plug-in power, lighting power, CO\textsubscript{2} concentration, indoor temperature, and fan operating status from June 2018 to July 2019 were collected. During the monitoring period, all the windows were closed, and the dedicated outdoor air ventilation system with constant-speed fans served the occupied spaces.

Methods for occupant detection found in the literature include CO\textsubscript{2}-based models [40, 41], camera-based sensors [31, 42], radio-frequency identification systems [43], and mobile sensing-based detection [43]. In this study, occupant counts are indirectly collected based on measured indoor CO\textsubscript{2} concentration while the plug-in and lighting load data were directly collected from the building’s submetering system. The CO\textsubscript{2}-based models, which treat the indoor CO\textsubscript{2} concentration as uniform, are selected to detect occupant counts, and the indoor occupant number (O) can be solved from the mass balance equation:

\[
O_t = \frac{V_r \frac{dc_t}{dt} + Q_{sup} (c_t - c^{sup})}{s}
\]

where \(V_r\) is the volume of the room (l), \(Q_{sup}\) is the supply airflow rate (l/s), obtained from the design flow rate and on/off status of constant-speed fans. \(s\) is the CO\textsubscript{2} generation rate per person (5 mg/s) following the settings in ASHRAE Standard 62.1–2016 [1]. \(O_t\) is the number of occupants, \(c_t\) and \(c^{sup}\) are the CO\textsubscript{2} concentration (ppm) of indoor air and supply air at the time instant \(t\) (s), respectively.

Fig. 3 illustrates the occupant-based power use and occupancy data collected from the test building. The plug load data was collected from the building’s submetering system, while the occupant counts were calculated by Eq. 1, based on the fan specification, measured CO\textsubscript{2} data.

---

Fig. 1. Research outline of probabilistic occupancy forecasting and ventilation decision-making schemes.

Fig. 2. (A) The view of the building and (B) its first-floor plan.
Fig. 4 presents the linear regression results for occupant counts with plug and lighting loads during the test period. The R-square value varied from 0.47 to 0.55 for the linear regression between plug-in loads and occupant counts and from 0.43 to 0.50 for the linear regression between lighting loads and occupant counts, which shows a strong correlation between occupant-based internal loads and occupant counts. It can be found that the lighting power density appeared to reach a plateau due to a maximum number of lighting appliances used, whereas plug-in power density continued to increase with the number of occupants. Although the lighting power density more likely reached maximum values at the high occupancy level, there was also a probability that light loads at the peak value under low occupancy level as office users tended to leave the lighting on when they temporarily left their offices. In contrast, the level of plug-in loads certainly increased with the increase of the occupancy level. For instance, the action of turning on plug-in appliances (i.e., desktop PCs) could be a strong signal that occupants were entering their offices.

2.2.2. Feature selection, data transformation and normalization

The input selection of a forecast model is key to its performance because it reflects deep correlations between the inner physical features and characteristics of occupancy. The inputs of forecast models can be divided into three categories: (1) historical occupancy data, (2) building electricity usage and (3) calendar information. Historical values of occupancy counts (or the estimation of them) are traditionally used as
inputs to occupancy prediction models, due to their usual high correlation with future values. Sub-metered electricity data (i.e., plug loads and lighting loads) are selected as they are highly correlated to building occupant numbers as presented in Fig. 4. Furthermore, for open-plan office spaces, since most occupants follow a daily work schedule, calendar information such as the hour of day and day of week (i.e., weekdays versus weekends/holidays), can capture the cyclic daily behaviour of occupants.

In this study, two sets of sequential input features with the same time lag of 24 h are selected to build predictive models for different purposes, as shown in Table 2. Option A includes all three input categories, which selects the 24 lagged plug-in power variables (P1, 2, ..., P24), 24 lagged lighting power variables (L1, 2, ..., L24), 24 lagged occupant counts (O1, 2, ..., O24) and calendar variables (e.g., hour of day [0,1] × 24, day of week [0,1] × 7). Option B includes the last two input categories, which remove inputs of the 24 lagged occupant counts to test whether it is possible to use the information of occupant-based electricity usage for occupancy prediction. Option A and Option B are suitable for zones with and without explicit occupancy counts, respectively, as the building energy data are most available for zones regulated by the building codes, but the occupancy sensors are not always installed. Here, for the zones without occupancy sensors, Option B can be potentially applied by transferring the knowledge learned from zones with occupancy information.

The purpose of data transformation is to provide a suitable dataset for input selection and model development. Data transformation methods are used to covert the categorical variable (e.g., calendar information) into dummy/indicator variables. Data normalization aims to eliminate the influence of the scales of data/measurement. All the candidate inputs, including the plug loads, lighting loads, and occupant counts, are normalized using max-min normalization as shown in Eq. 2, where $X = (x_1, x_2, ..., x_d)$ and $\tilde{X}$ is the normalized data of $X$.

$$\tilde{X} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

(2)

The data are further split into three sets: the training, validation, and testing sets. The testing sets are selected from the last month of the data, whereas the training and validation sets are split randomly using the rest data with a proportion of 7:3. The training and validation sets are used for hyperparameter optimization and model training, while the testing sets are used for the performance evaluation of models.

### 2.3. Probabilistic deep learning framework

The complete architecture of the developed ABLSTM framework for occupancy forecasting contains two layers as shown in Fig. 5: (i) an unsupervised learning layer, autoencoder, that captures the inherent pattern in the time series and is learned during pre-training, and (ii) a supervised prediction network that receives input both from the learned representative embeddings within the autoencoder framework as well as external features (e.g., calendar information) to predict the occupancy.

Three types of prediction uncertainty are considered in this deep learning framework [44]: model misspecification, epistemic uncertainty, and aleatoric uncertainty. Model misspecification refers to the assumed model class used in a learning/reasoning algorithm that represents an imperfect approximation to reality. Aleatoric uncertainty represents the variability of true random or uncontrollable processes, which is irreducible because there will always be variability in the underlying variables. Epistemic uncertainty represents deficiencies by a lack of knowledge, information or expertise, which is reducible at least theoretically by searching for knowledge that allows predicting its outcome with greater accuracy.

#### 2.3.1. Autoencoder

An autoencoder layer is introduced to eliminate model misspecification when predicting unseen samples with very different patterns from the training data sets. The autoencoder layer consists of an encoder and a decoder. The encoder extracts representative features from a time series from the training data set, and the decoder reconstructs the original data back to the original input space [45]. This network is trained by minimising reconstruction errors with a loss function, which measures the differences between the original input and the consequent reconstruction. Let the encoder be given by $f(X)$ and the decoder be given by $g(X)$, the autoencoder then minimizes the loss function $\xi$ (e.g., mean square error) as represented by Eq. 3, where input is given by $\tilde{X} \in \mathbb{R}^d$ and $d$ is the dimension of the input.

$$\text{minimize } \xi(X, g(f(X)))$$

(3)

In this study, both the encoder and decoder adopt LSTM Networks. LSTM is a special subtype of deep learning neural network that has the capability to handle sequential data, such as forecasting time-series building energy demand [45], indoor air temperature [46] and PV power generation [47]. It learns long- and short-term impacts from past time series and outperforms traditional non-linear machine learning algorithms by introducing three special gates, namely forget gate, input gate and output gate. Forget gate decides what is relevant to keep from historical information, given by Eq.4. The input gate decides what new information in $\tilde{X}$ should enter the cell state, accomplished by Eqs. 5-6.

The old state was multiplied by $f^{t}$ to forget what is unnecessary and to update from $C^{t-1}$ to $C^{t}$, given by Eq.7. Eventually, based on $C^{t}$, $h^{t-1}$ and $\tilde{X}$, the output gate provides the value of the next hidden state $h^{t}$ by Eqs.8-9.

$$f^{t} = \sigma(W_{f}[\tilde{X}, h^{t-1}] + b_{f})$$

(4)

$$u^{t} = \sigma(W_{u}[\tilde{X}, h^{t-1}] + b_{u})$$

$$\tilde{C}^{t} = \tanh(W_{c}[\tilde{X}, h^{t-1}] + b_{c})$$

(5)

$$C^{t} = f^{t} \odot C^{t-1} + u^{t} \odot \tilde{C}^{t}$$

$$C^{t} = f^{t} \odot \tanh(C^{t})$$

where $f^{t}$, $u^{t}$, and $\sigma$ denote the update gate, forget gate, and output gate at time $t$, respectively. $\tilde{X}$ denotes the inputs at time $t$. $h^{t-1}$ and $h^{t}$ denote the outputs of the current layer of the at time $t-1$ and $t$, respectively. $t$ denotes the time step with an hourly interval in this study. $\sigma$ denotes the sigmoid function. $\tanh()$ denotes the hyperbolic tangent function. $C^{t}$ denotes the memory candidate at time $t$. $C^{t}$ denotes the memory value at time $t$. $W_{o}$, $W_{f}$, $W_{i}$, and $W_{u}$ denote weight matrices to calculate the update gate, forget gate, output gate, and candidate gate, respectively.

The reconstructed representatives together with the external features are then used as the inputs for the Bayesian LSTM networks. It is worth noticing that, for training the proposed model with Option A, the

---

### Table 2

Comparison of model feature options.

| Feature option | Model inputs | Model outputs |
|----------------|--------------|---------------|
| Option A      | Historical data (24 lagged time step): occupant counts (O1, O2, ..., O24), plug loads (P1, P2, ..., P24), lighting loads (L1, L2, ..., L24), Calendar information: information: hour of day (0,1] × 24, day of week (0,1] | Occupant counts (O1, O2, ..., O24), lighting loads (L1, L2, ..., L24) |
| Option B      | Historical data (24 lagged time step): plug loads (P1, P2, ..., P24), lighting loads (L1, L2, ..., L24), Calendar information: information: hour of day (0,1] × 24, day of week (0,1] | Plug and lighting loads (P1, P2, ..., P24), lighting loads (L1, L2, ..., L24) |
autoencoder’s inputs are the lagged occupant counts, lighting loads and plug loads, while for the proposed model with Option B, the autoencoder’s inputs are lagged lighting loads and plug loads. The external features are the calendar information for both models.

2.3.2. Bayesian deep neural networks

A Bayesian LSTM will be further applied for the prediction of time-series data, taking account of epistemic uncertainty and aleatoric uncertainty. In this study, a Monte Carlo (MC) dropout approach [48] is used for approximate Bayesian inference, to capture the ignorance of the model parameters and reduce the impacts of epistemic uncertainty. This method applies stochastic dropout in each weight layer during model training and also performs dropout at the testing process to provide an ensemble of predictions from the approximate posterior [49]. Specifically, a diagonal matrix diag(d) is applied to the weight matrix (W) and bias vector (B) in each hidden layer, as given by Eq. 10. All diagonal elements of the diag(d) are Bernoulli random variables.

\[
[W', B'] = \text{diag}(d)[W, B]
\] (10)

To consider the impacts of aleatoric uncertainty, different from that (typically mean square error) of deterministic regression models, the loss function is adopted given by Eq. 11 during the model training, which captures noise inherent in the observations. It consists of two parts: the residual regression obtained with stochastic samples and an uncertainty regularization term. Therefore, the training process can be conducted by supervised learning of the regression task without "uncertainty labels" [50].

\[
\text{Loss} = \frac{1}{2\delta^2} + \frac{1}{2} \log \delta^2
\] (11)

where the model predicts a mean \(\hat{y}\) and variance \(\sigma^2\), and \(y\) denotes the actual output (i.e., occupant number).

To optimize the hyperparameters of the developed ABLSTM models, the Bayesian optimization [51] in the Keras Tuner [52] is implemented. In this study, the hyperparameters studied include the number of hidden neurons in each hidden layer, number of hidden layers, learning rate, and dropout rate.

2.3.3. Quantification of model misspecification, aleatoric uncertainty and epistemic uncertainty

As aforementioned by Eq. 11, both the means (\(y_1, y_2, y_3, ..., y_M\)) and variances (\(\sigma_1^2, \sigma_2^2, ..., \sigma_M^2\)) of the occupancy forecasts can be directly derived from the ABLSTM model. As the encoder reconstructs the inputs in embedding spaces (accounting for potential model misspecification) and gets further propagated through the prediction network with random dropout, the variances are therefore the aggregated uncertainties taking account of both model misspecification and aleatoric uncertainty, as given by Eq. 12. The epistemic variance of the occupancy forecasts is obtained by calculating the variance of sampled outputs using Monte Carlo, as given by Eq. 13.

\[
\delta_{\text{mis+ale}} = \frac{1}{M} \sum_{j=1}^{M} \delta_j
\] (12)

\[
\delta_{\text{ep}} = \frac{1}{M} \sum_{j=1}^{M} \left( \hat{y}_j - \frac{1}{M} \sum_{j=1}^{M} \hat{y}_j \right)^2
\] (13)

where, \(M\) is the number of Monte Carlo samples, taking 200 in this study. \(j\) is a certain positive integer number.

2.4. Model transferability test

The three open-plan floors in the test building have similar functions and floor areas, mainly the offices used by research students and staff. As curriculum schedules and teaching activities varied on each floor, the occupant pattern and behaviours may differ. For instance, although the design occupant capacity was 50 for each floor, the maximum occupant counts were different as presented in Fig. 4. To transfer representations of occupancy from one floor to another floor, the models trained by data of the first floor are used or fine-tuned to predict occupant numbers for the second and third floors to explore the model transferability.

For Option A, the test process consists of the following parts. (1) The source datasets from the first floor are used to pre-train a base ABLSTM model. (2) The target training and validation dataset from the second and third floors are used to fine-tune base ABLSTM and thus we can get
the target ABLSTM. (3) The target test datasets from the second and third floors are used to validate the target ABLSTM.

For Option B, as the occupant counts are assumed unknown, the test process consists of the following parts: (1) The source dataset from the first floor is used to pre-train an ABLSTM model. (2) The structure and weights of the ABLSTM model will be directly applied to target datasets, while the predicted normalized occupant counts ($\hat{O}$) from ABLSTM will be scaled back via inverse transform function to the original representation ($\bar{O}$) based on the design occupant number ($d_{oc}$) and baseline occupant number ($b_{oc}$) of the target spaces, as presented by Eq. 14. In this study, the baseline occupant number is set to zero based on the site spot observations. The flowchart of the model transferability test can be found in Fig. 6.

$$\hat{O} = \bar{O} \cdot (d_{oc} - b_{oc}) + b_{oc}$$

(14)

2.5. Performance evaluation

2.5.1. Baseline model

The conventional LSTM neural network is chosen as the baseline model to demonstrate the performance improvement of developed ABLSTM models. The hyperparameter tuning, model training, and feature selection are carried out in the same way as for those in the ABLSTM model.

2.5.2. Performance metrics

Five predictive performance metrics are used for evaluating the model performance. Three of them are for deterministic accuracy, i.e., Root Mean Square Error (RMSE), the Mean Absolute Percentage Error (MAPE), and the Coefficient of Variance of the Root Mean Square Error (CVRMSE) as given by Eqs. 15-17. These indices are used to evaluate the deviations between the predictive mean and the real occupant number. The other two are for uncertainty accuracy, $\chi$-accuracy [53] and prediction interval coverage probability (PICP), as given by Eq. 18 and Eq. 19. These two indices are used to evaluate the performances of the predictive variances due to model uncertainty. The $\chi$-accuracy can be calculated by adjusting the error tolerance threshold $\chi$ of correct prediction. For instance, the 0-accuracy ($\chi = 0$) means a prediction is regarded as correct only when this prediction perfectly matches actual occupancy. PICP is defined as the proportion of observed occupant numbers located at the predicted occupant number confidence interval. The larger the PICP is, the better the model performs. It is worth noticing that PICP is only applicable for the proposed ABLSTM, as the baseline LSTM only derives deterministic values.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (O_i - \bar{O}_i)^2}$$

(15)

$$\text{MAPE} = 100\% \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{|O_i - \hat{O}_i|}{O_i}$$

(16)

$$\text{CVRMSE} = 100\% \cdot \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (O_i - \bar{O}_i)^2}}{\bar{O}_i}$$

(17)

$$\chi - \text{accuracy} = \frac{\sum_{i=1}^{N} \Omega(|O_i - \bar{O}_i|, \chi)}{N}, \quad \text{while } \Omega(|O_i - \bar{O}_i|, \chi) = \begin{cases} 1, & |O_i - \bar{O}_i| \\ 0, & \text{otherwise} \end{cases}$$

(18)

$$\text{PICP} = \frac{\sum_{i=1}^{N} k_i}{N}, \quad \text{while } k_i = \begin{cases} 1, & \text{if } l_i \leq O_i \leq u_i \\ 0, & \text{otherwise} \end{cases}$$

(19)

where $O_i$ and $\hat{O}_i$ are observed and predicted occupant number at the hourly resolution, respectively. $\bar{O}_i$ is the mean value of the $N$ actual occupant number. $\bar{O}_i$ is a predictive mean at the hourly resolution. $l_i$ and $u_i$ is a certain positive integer. $N$ is the size of the test data set. $\chi$ is the tolerance level. $l_i$ and $u_i$ are the lower and upper bounds of the predictive occupant number at the ith hour.

It is worth noticing that, for actual practice, if the occupant level inside a zone or room is relatively low, it is not necessary to turn on the ventilation systems or open the windows because the indoor infiltration can provide sufficient fresh air. Therefore, to assess the performance of ventilation systems or open the windows when the indoor infiltration level is relatively low, it is not necessary to turn on the ventilation systems. Ventilation rates for zones. Two operation modes, normal DCV mode and anti-infection mode, are

A) Option A

- Source dataset
- Knowledge transfer
- Base ABLSTM

Input: Lagged occupant counts, plug-in load, lighting loads, indoor temperature, calendar information

Output: Future occupant number

B) Option B

- Source dataset
- Input: Lagged plug-in load, lighting loads, indoor temperature, calendar information

Output: Future occupant number

Fig. 6. Test of model transferability for zones with (A) and without (B) occupant sensors.

C. Zhuang et al.

Building and Environment 219 (2022) 109207
proposed and can be integrated into the building management systems for different ventilation requirements under pandemic and non-pandemic periods, respectively. The normal DCV mode is mainly intended to save energy with the satisfaction of indoor air quality in contaminant based on occupancy density. Generally, the airflow rate required per person by anti-infection mode will be higher than that of normal DCV mode. For instance, it is reported that to control the infection risk in offices below 2%, the ventilation rate should be improved from 30 m$^3$/h-person (a designed rate of normal DCV mode) to 225 m$^3$/h-person under the design occupant capacity with a 2-h exposure time [4].

In this study, the ventilation decision is made by evaluating the probability of ventilation inadequate ($R_d$) under normal DCV mode and individual risk of infection ($R_i$) under anti-infection mode respectively.

### 2.6.1. Normal DCV mode

The probability of ventilation inadequacy ($R_d$) is used as an index to quantify the performance of the ventilation system under normal DCV mode during the non-pandemic period. $R_d$ is defined as the probability that the demand ventilation rate is lower than the supply ventilation rate, as shown in Eq. (20).

$$ R_d = \Phi(V_{dem} < V_{sup}) $$

where, $\Phi$ is the cumulative density function. $V_{dem} = [v_{dem,1}, v_{dem,2}, \ldots, v_{dem,n}]$ and $V_{sup} = [v_{sup,1}, v_{sup,2}, \ldots, v_{sup,n}]$ are the probabilistic demand and supply ventilation rates respectively (l/s). In the study, we select the scenario where the mechanical ventilation systems are turned off, and all windows of the zone are closed. Therefore, the supply ventilation rate $v_{sup}$ approximates the infiltration rate ($v_{inf}$).

The demand ventilation rate ($v_{dem}$) is adjusted based on detected occupancy and occupied area by adopting the DCV strategy, while the infiltration ($v_{inf}$) is calculated using an empirical method described by Ref. [54], as per Eqs. (21-22):

$$ v_{inf} = P_{net}v_{ref} + A_{zone}v_{R} $$

where $P_{net}$ is the total occupancy, $A_{zone}$ is the served floor area (m$^2$), $v_{ref}$ and $R_a$ are the fresh air requirement per person (l/s-person) and per unit area (l/s-m$^2$) respectively, specified by Ref. [1]. $v_{inf,d}$ is the design infiltration rate (l/s), set as 0.00137 l/s-m$^2$ of building envelope surface area [55]; $|\Delta T|$ is absolute indoor-outdoor temperature difference (°C); $U_{H}$ is wind speed (m/s); $\phi_0, \phi_1, \phi_2, \phi_3$ are regression coefficients. We used $\phi_0 = 0.606, \phi_1 = 0.036, \phi_2 = 0.1177, \phi_3 = 0$ for the regression coefficients, the coefficient scheme recommended by the EnergyPlus predecessor BLAST [56]. Taking account of the measurement uncertainty of the anemometer as well as the design infiltration rate uncertainty, the distributions of local wind speed and design infiltration rate are set following the relative normal distribution (1, 10%) [57-58]. The distribution of air infiltration rates can therefore be obtained by propagating these two uncertainties.

### 2.6.2. Anti-infection mode

The individual risk of infection ($R_i$) is used as an index to quantify the performance of ventilation systems under an anti-infection mode during the pandemic period. We consider that individuals are far apart from each other (the recommended safe social distance, i.e., ≥2 m [59]), and there is no risk of short-range transmission either by large or small droplets/aerosols, nor by contact with contaminated surfaces. Based on the perfect-mixing-based Wells-Riley model, $R_i$ can be calculated by Eq. (22), which is the function of air changes rate per hour ($ACH$) of supply air, the proportion of infected people ($\alpha$) and the number of occupants ($O$), where $k_p = 4.1 \times 10^2$, and $N_{virus}$ is the total number of viral particles (copies) inhaled over some time. $N_{virus}$ can be calculated based on knowledge of the mask type, occupant activity, ventilation filter efficiency, etc., given by Eqs. (23-25):

$$ R_i = k_p \cdot \exp \left( -\frac{N_{virus}(ACH, \alpha, O)}{k_p} \right) $$

$$ N_{virus}(t) = N_{virus}(t_0) + (1 - \eta) \int_{t_0}^{t} C(t) \cdot Q_{inf} dt $$

$$ C(t) = \alpha O (1 - \eta) \frac{N_{virus}(t)}{V(\lambda + ACH)} + \left( \frac{C(t_0) - \alpha O (1 - \eta) N_{virus}(t_0)}{V(\lambda + ACH)} \right) \exp \left( -(\lambda + ACH)(t - t_0) \right) $$

where $N_{virus}(t_0)$ is the total amount of virus inhaled until $t = t_0$, $\eta$ is mask efficiency ($\eta = 0$ if without face masks). $C_{avg}$ is the temporal evolution of the particle concentration (PFU/m$^3$). $Q_{inf}$ is an inhalation rate (0.521 l/s for an average person [60]). $N_{virus}$ is viral particles per unit of time emitted by continuous exhalation (PFU/s). $\lambda$ is decay coefficient, 1.026 h$^{-1}$, considering viral decay and the settling of the aerosol droplets. The proportion of infected people ($\alpha$) is chosen as 8%, selecting the peak population rate testing positive for COVID-19 in the UK [61]. More details can be found in Ref. [62].

### 3. Test results of model performance

In the case study, the performances of the developed models were evaluated for the 1-h ahead occupancy forecasting using 24-h lagged inputs.

#### 3.1. Forecasting performances of developed models

Based on well-tuned hyperparameters in each model, the generalization capacity was further evaluated by applying each model to the testing datasets. The calculated performance metrics of the developed models were shown in Table 3.

The developed ABLSSTM model outperformed the baseline model for both options. For Option A, CVRMSE and MAPE of the ABLSSTM were around 7.0% and 5.8% less than the baseline model, respectively. The 3-tolerance accuracy (when the error of a prediction is less than 3 occupants) of both models were shown in Table 3.

### Table 3

| Feature option | Metrics | Model |
|----------------|---------|-------|
| **Baseline LSTM** |          |       |
| **Proposed ABLSSTM** |          |       |
| **Option A**: lagged occupant count, plug loads, lighting loads, calendar information (hour of day, day of week) | RMSE 2.96 | 1.43 |
| | CVRMSE 23.6% | 16.6% |
| | MAPE 20.8% | 15.0% |
| | accuracy 91.5% | 93.8% |
| | PICP – | 83.3% |
| **Option B**: lagged plug loads, lighting loads, calendar information (hour of day, day of week) | RMSE 5.45 | 4.85 |
| | CVRMSE 65.1% | 55.2% |
| | MAPE 59.6% | 55.6% |
| | accuracy 72.1% | 75.4% |
| | PICP – | 75.1% |
the values and trends can be observed for the two models, the proposed ABLSTM was more suitable to use for the decision-making process, as its predictive confidence interval covered almost all observed values and occupant scenarios.

For Option B, the ABLSTM also outperformed the baseline model. However, as lagged occupant counts were not used in the training, the CVRMSE and MAPE of ABLSTM were increased from 16.6% (Option A) to 55.2% and 15.0% (Option A) to 55.2%. However, the 3-tolerance accuracy and PICP for the proposed ABLSTM were both higher than 75%, indicating that the model’s performance was acceptable most of the time in a day. The forecasting results of the models with Option B were compared in a typical two weeks, as shown in Fig. 8. The ABLSTM model had a satisfactory prediction performance at low occupant level, which can guide the ventilation fans to run at low speed under low occupant density. For instance, in existing studies [63] [64], for zones with dual speed fans but without occupant sensors, the selection of high

![Fig. 7. Prediction result of models with Option A.](image)

![Fig. 8. Prediction result of models with Option B.](image)
speed and low speed was entirely based on the anticipation of occupant flow or predefined operating schedule. Although the proposed ABLSTM model cannot accurately predict the occupant numbers, it approximated the occupant pattern (i.e., at a low level or high level). Therefore, it is suitable to be applied for dual-speed ventilation control for energy conservation. In addition, with the adoption of the Bayesian framework, most of the observed occupant values were located at the prediction confidence interval, as the model detected high uncertainty at a high occupant level.

From Figs. 7 and 8, it can be seen that the confidence interval was wider when occupancy levels were high, indicating the high prediction uncertainty at high occupancy. Different uncertainty sources of occupancy forecasting were further quantified as shown in Fig. 9. The model misspecification and aleatoric uncertainty variances altered in a small range, while the aleatoric variances fluctuated substantially. Compared with model misspecification and aleatoric uncertainty, epistemic uncertainty contributed to a large proportion of main uncertainty sources. The aggregated model misspecification and aleatoric uncertainty for Option A and Option B varied in the ranges of [0, 0.4] and [0, 2.0] respectively, while the epistemic uncertainty for Option A and Option B varied in the ranges of [0, 8.1] and [0, 26.9] respectively. Option B experienced a higher uncertainty compared with Option A, as it only treated the building energy demand and calendar information as inputs. Building energy demand may be affected not only by occupant number but also other factors such as occupant actions and activities (window opening/closing, temperature setpoint, usage patterns of electrical appliances). Under high building energy demand, occupants tend to have high spatial and temporal randomness. For instance, during lunchtime when people leave offices, building plug loads would not be changed much as occupants may keep their electrical appliances (e.g., desktop PCs) running. Taking account of past steps' occupancy information into feature selection (i.e., Option A) can effectively reduce the predictive uncertainties.

3.2. Transferability performance of developed models

The transferability of developed models was also tested using the datasets from the second and third floors in the same buildings. It is worth noticing that, for Option A, as fine-tuning the pre-trained model was needed, the target datasets were split into training, validation, and testing sets. For Option B, as the developed models by source target (1st floor) were directly applied to the datasets without pretraining, the whole target datasets were used as the testing sets. Table 4 lists the performance of the ABLSTM models when applied to the new datasets for both options.

For Option A, the CVRMSE and MAPE of ABLSTM were 17.4% and 14.0%, respectively, using second-floor datasets, and 19.1% and 15.7% for third-floor datasets. The ABLSTM using both two datasets showed a satisfactory performance with 3-accuracy higher than 98% and PICP higher than 85%. The transferability results of the models with Option A were compared in a typical two weeks, as shown in Fig. 10, it can be seen that predictive means from the developed models match well the observations.

Compared with Option A, the CVRMSE and MAPE of ABLSTM of Option B experienced a 31.2% and 30.1% increase respectively using second-floor datasets, and a 39.8% and 31.4% increase respectively for third-floor datasets. Though larger errors were observed, the ABLSTM using both two datasets for Option B still showed a good performance with 3-accuracy higher than 74% and PICP higher than 71%. The transferability results of the models with Option B were compared in a

Table 4

| Feature option | Metrics | Dataset (second floor) | Dataset (third floor) |
|----------------|---------|------------------------|----------------------|
| Option A       | RMSE    | 0.92                   | 0.70                 |
|                | CVRMSE  | 17.4%                  | 19.1%                |
|                | MAPE    | 14.0%                  | 15.7%                |
|                | 3-accuracy | 98.2%                  | 99.5%                |
|                | PICP    | 85.4%                  | 86.6%                |
| Option B       | RMSE    | 4.85                   | 5.31                 |
|                | CVRMSE  | 48.6%                  | 58.9%                |
|                | MAPE    | 44.1%                  | 47.1%                |
|                | 3-accuracy | 74.1%                  | 74.3%                |
|                | PICP    | 73.9%                  | 71.8%                |

Fig. 9. Quantified forecast uncertainties with different input features.
typical two weeks, as shown in Fig. 11. Similar to that of the first-floor datasets, a higher prediction confidence interval can be observed during periods of higher occupancy.

In this study, the main purpose of the model transfer is to save time for constructing and training models. The target models (i.e., second and third floors) had similar or even better performance compared with the source model (i.e., first floor) for the following two reasons: 1) All three floors have similar functions, area, occupant activity and pattern. The number of datasets was the same for each floor (from Jun 2018 to July 2019 with 1-h intervals). Compared with that of the source model solely using the first floor’s datasets for training, the target models were first pre-trained by datasets of the source model and then fine-tuned using their own datasets. Therefore, the datasets for training or tuning the target models doubled that of the source model. 2) During the test
periods, the mean occupant numbers for the first, second and third floors are 7.5, 5.4 and 5.9, respectively. As it can be seen in Figs. 7 and 9, high occupancy yields high model prediction uncertainty. Due to the smaller magnitudes and ranges of actual occupancy on the second and third floors, the target models show even better performance compared with the source model.

Through the transferability test, we find that developed models for the information-rich floor (i.e., with occupant sensors) could be further applied for information-poor floors (i.e., without occupant sensors) in the same building. For real applications, the ventilation control in information-poor offices could be beneficial from this approach.

4. Application of probabilistic deep learning models for optimal ventilation control

As observed from Sections 3 and 4, the BLSTM with Option A shows satisfactory performance for building occupancy forecasting. Therefore, the BLSTM with Option A can be used for real-time ventilation control to reduce energy use under normal DCV mode and to prevent infection under the anti-infection mode.

4.1. Normal DCV mode

Fig. 12 shows the probability of ventilation inadequacy ($R_n$) under normal DCV mode. With the assumption that all the windows and ventilation fans were closed (only infiltration involved), the probability of ventilation inadequacy was calculated by comparing the required supply airflow rates and air infiltration rates. The meteorological data (i.e., wind speed and outdoor temperature) from a nearby weather station and indoor temperature data from the building management system were also collected to calculate the air infiltration rates by Eq. 22. It can be found that, due to the low occupant levels in buildings, there was no need for mechanical invention or window opening on 4, 5, 11 and 12 Aug ($R_n = 0$). Besides, the quantification of $R_n$ also provided alternatives for ventilation operation by setting the risk thresholds ($\gamma$). For instance, for comfort-oriented alternative, decision-makers (e.g., facilities managers) can set the low value of $\gamma$ (e.g., 10%). Therefore, except for the above-mentioned four days, mechanical ventilation or window opening was required for the rest of the days to guarantee indoor thermal comfort or air quality. For the energy-oriented alternative, facilities managers can raise the value of $\gamma$ (e.g., 30%). Therefore, mechanical ventilation or window opening was not required for an additional three days such as 30 July, and 8–9 Aug for energy-saving purposes.

4.2. Anti-infection mode

Fig. 13 shows the individual infection risk 1-h exposure under anti-infection mode. The number of infected people was proportional to the forecasted occupancy. With the assumption that all the windows and ventilation fans were closed (only infiltration involved), the individual infection risk was calculated by using the infiltration air flowrate (blue curve in Fig. 12) as inputs. The individual risks were also quantified in a probabilistic manner to guide the ventilation operation. For instance, facilities managers can compare the individual risks at different quantiles with the targeted infection rate to determine the operation of ventilation systems. If we set the targeted infection rate to be 2% following the suggestion by Ref. [4] and compared it with the mean infectious risk, it can be found that, apart from 6 days (30 Jul, 4, 5, 9, 11 and 12 Aug), the mechanical ventilation or window opening were required at the rest of days to achieve the targeted infection rate.

5. Discussion

In Sections 3 and 4, it is found that the proposed ABLSTM models with Option A had a satisfactory performance, which could potentially be used for real-time ventilation control. The performances of the developed models were first evaluated for the 1-h ahead load forecasting based on previous 24-h data. Subsequently, a representative model was selected to further investigate the impacts of the input time lags and forecasting horizons on model performances. The limitations and future work are further presented in this section.

5.1. Impact of input time lags

Based on the above results, we selected the ABLSTM model as the representative to investigate the impact of time lags of input features (Option A) on model performances. The studied range of time lags (lagged hours) was set as 1–72 h. Fig. 14 shows the forecasting performances of the developed ABLSTM model under different input time lags. Similar trends were seen between the two metrics. For the first 10 time lags, both CVRMSE and MAPE decreased, while the errors fluctuate at the rest of the time lags. This indicates that the forecasting performances of the ABLSTM model were considerably affected by the input time lag when it was short, e.g., less than 10 in this study. After the 10th time lag, most of these two metrics stabilized in the ranges of [16%, 18%] and [15%, 17%], respectively.

5.2. Impact of forecasting horizons

To facilitate predictive building energy management, multiple-hour ahead occupancy forecasting is usually required. In the 1-h ahead load forecasting, it was identified that the performance of the model was not stable when the input time lags were longer than 10. Therefore, to explore the impacts of forecasting horizons, we also used the 24-h
historical trajectories to forecast the occupancy. Fig. 15 shows the CVRMSE and MAPE of the ABLSTM model under different forecasting horizons (1–24 h). The overall trends in these two metrics were similar. When the forecasting horizons were higher than 3, the CVRMSE and MAPE values were high than 30%, which indicates the ABLSTM model would not be suitable for occupant detection due to the high randomness of the occupancy. By increasing the forecasting horizon from 1 to 8 h, both CVRMSE and MAPE values increased. This indicates that the accuracy of the developed ABLSTM model decreased with the increase of the forecasting horizon before reaching 8 h. From the 8–24 h forecasting horizon, most of the CVRMSE and MAPE values stabilized.
5.3. Limitations and future work

The proposed probabilistic occupancy forecasting model integrated with risk-aware ventilation decision-making schemes could offer indoor spaces with low infection probability and high energy efficiency based on quantified risks. This is especially crucial for real applications, where different sources of uncertainty may affect the operation and performance of ventilation systems. Based on the performance tests of the developed method and a review of the existing literature, several issues still need to be addressed in further studies for the successful implementation in practice:

1) The occupant counts were estimated based on a CO2-based model in this study. However, the latency of the CO2 sensor responses (i.e., time delay) due to the slow gas mixture rate, the improper place locations and numbers reduce the precision in occupancy estimation [65]. Existing studies [33] [34] showed that the R-squared values could achieve 0.65–0.78 for the linear regression between plug-in loads and occupant counts in offices, while the values were only 0.47–0.55 in this study. This is partially attributed to the deficiencies of CO2-based occupancy detection methods for occupancy estimation. Although we proposed a Bayesian deep learning approach that forecasts occupancy in a probabilistic manner rather than a deterministic way to compensate for variations in occupancy estimation, the model prediction accuracy with Option B is expected to be further improved with advanced occupancy detection methods. More accurate occupancy detection methods, such as radio frequency-based, video camera-based, and sensor fusion methods [66] should be applied to obtain the occupant counts with higher accuracy.

2) The transferability test can be also conducted in different types of buildings to further validate the performance of the proposed models and the effectiveness of the model framework. In addition, the developed occupancy prediction model can be further expanded or improved by expanding the scale and scope of the building types, locations, and design alternatives. The proposed ventilation schemes could be implemented in real building energy management systems with different ventilation types (e.g., variable air volume systems) to improve energy efficiency and ventilation effectiveness.

3) To determine the appropriate supply airflow rate, the building design infiltration rate can be calibrated for various building types (i.e., different surface area to volume ratios, ventilation systems, and layouts). The uncertainty of the design infiltration rate can also be estimated. Possible approaches involve the use of tracer gases tests [67], CFD-assisted air-velocity method [68] and Bayesian calibration [69].

4) For quantifying the infectious risk in anti-infection mode, we selected the scenario in which the masks and filters were not used. The infectious risk could be further evaluated by considering various mask types used by occupants (e.g., N95, surgical and 3-ply cloth), ventilation air filter types (e.g., HEPA, ISO coarse) as well as occupant activities (e.g., sitting, standing, heavy exercise). Additional commissioning is also needed to determine the optimal risk thresholds (γ) to both fulfill user requirements and meet control objectives. Different risk thresholds should be provided for the models with different prediction accuracies under different operating modes. For instance, for the prediction model with lower accuracy, γ can be set lower and vice versa. For more conservative solutions, decision-makers can set the γ to zero to avoid potential risks.

5) The proposed strategy can be further implemented in the existing field networks of building automation systems. The ventilation supervisory controller can be developed to integrate into the central control systems, which consists of two parallel networks: the training network and the predictor network. The training network is used to learn the relationship between the indoor energy/environmental parameters and occupant numbers based on continuous online monitoring [70]. The training network weights are then passed to the predictor network where they are used for forecasting the occupant number at the next step. The predictor network subsequently finds optimal ventilation flowrate setpoints that can minimize the potential risks. Eventually, the local process controller (i.e., PID) ensures ventilation rates are controlled at the optimal setpoints by adjusting the supply air damper.

6. Conclusions

In this study, an autoencoder Bayesian long short-term memory neural networks (ABLSTM) were developed for probabilistic occupancy prediction, taking account of model misspecification, aleatoric and epistemic uncertainties. Two feature options with or without occupant counts (i.e., Options A and B), and their transferability for the developed models were investigated. Risk-aware decision-making schemes were proposed for optimal ventilation under pandemic and non-pandemic scenarios. The impacts of input time lags and forecasting horizons on model performances were explored. Based on the testing results, some detailed conclusions can be drawn:

- The developed ABLSTM models outperform the baseline LSTM model using both feature options. Option A is more suitable for real-time ventilation control due to its satisfactory accuracy. Option B is more suitable for zones without occupant sensors but with dual speed fans for two-level ventilation control, as it can accurately capture the low-level occupancy. During the test period (i.e., one month), the mean absolute percentage errors of the developed models with Option A and Option B were reduced by 5.8% and 4.0%, respectively, compared with that of the baseline model. However, due to the random nature of occupant behaviour, it is very difficult to forecast occupancy accurately. For the proposed ABLSTM, over 80% and 70% of actual occupant numbers were within the predicted confidence interval for Option A and Option B, respectively.

- The developed ABLSTM models can take three sources of uncertainty into account. Model misspecification is eliminated by the autoencoder layer, while epistemic uncertainty and aleatoric uncertainty are considered based on Monte Carlo dropout with a modified loss function. The developed models can also be transferable and used for other zones with similar activities without sacrificing the prediction accuracies.

- The proposed risk-aware decision-making schemes can offer flexible alternatives to ventilation operation for decision-makers under normal demand control ventilation mode and anti-infection modes. This enables the occupied zones with high energy efficiency under non-pandemic periods and low infection probability (i.e., below 2%) under pandemic periods.

CRediT authorship contribution statement

Chaoqun Zhuang: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. Ruchi Choudhary: Writing – review & editing, Supervision, Methodology, Funding acquisition. Anna Mavrogianni: Writing – review & editing, Supervision, Methodology.

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.

Acknowledgements

The research presented in this paper is financially supported by the AI for Science and Government (ASG), UKRI’s Strategic Priorities Fund.
awarded to the Alan Turing Institute, UK (EP/T001569/1 and EP/W00622/1).

Reference

[1] ASHRAE 62.1-2016. Ventilation for Acceptable Indoor Air Quality, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Atlanta, GA, 2016.

[2] European Committee for Standardization. En 15251, Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics, EUR Standard (2012), https://doi.org/10.1520/E2019-03R13; 2012;4:0-9.

[3] C. Zhuang, J. Shan, S. Wang, Coordinated demand-controlled ventilation strategy for energy-efficient operation in multi-zone cleanroom air-conditioning systems, Build. Environ. 191 (2021), 107988, https://doi.org/10.1016/j.buildenv.2021.107988.

[4] C. Sun, Z. Zhai, The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission, Sustain. Cities Soc. 62 (2020) 102390, https://doi.org/10.1016/j.scs.2020.102390.

[5] H. Sha, X. Zhang, D. Qi, Optimal control of high-rise building mechanical ventilation system for achieving low risk of COVID-19 transmission and ventilative cooling, Sustain. Cities Soc. 74 (2021) 103256, https://doi.org/10.1016/j.scs.2021.103256.

[6] J. Wang, J. Huang, Q. Fu, E. Gao, J. Chen, Metabolism-based ventilation monitoring and control method for COVID-19 risk mitigation in gymnasiums and alike places, Sustain. Cities Soc. 80 (2022) 103719, https://doi.org/10.1016/j.scs.2022.103719.

[7] B. Li, B. Wu, Y. Peng, W. Cai, Tube-based robust model predictive control of multi-zone demand-controlled ventilation systems for energy saving and indoor air quality, Appl. Energy 307 (2022), 118297, https://doi.org/10.1016/j.apenergy.2021.118297.

[8] Chartered Institution of Building Services Engineers (CIBSE). Coronavirus, SARS-CoV-2, 2020:1 virus – workplaces 2020:1 building services in order to prevent the spread of the coronavirus disease (COVID-19) in workplaces, www.cibse.org/coronavirus-covid-19/coronavirus-covid-19-and-hvac-systems. (Accessed 14 February 2022).

[9] C. Zhuang et al. – Proc 2011 – 11 - Proc 2011

[10] ASHRAE 62.1-2016. Ventilation for Acceptable Indoor Air Quality, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Atlanta, GA, 2016.

[11] World Health Organization (WHO), Roadmap to improve and ensure good indoor air quality, 2019. https://www.who.int/b built-environment/2019/HPCC-SmartCity-DSS.2017.09.

[12] Centers for Disease Control and Prevention (CDC), Coronavirus, SARS-CoV-2, 2019–COVID-19, 2020. https://www.cdc.gov/coronavirus/2019–2020/COVID-19/2019–2020.html. (Accessed 14 February 2022).

[13] World Health Organization (WHO), Roadmap to improve and ensure good indoor air quality, 2019. https://www.who.int/b built-environment/2019/HPCC-SmartCity-DSS.2017.09.

[14] C. Zhuang, S. Wang, Risk-based online robust optimal control of air-conditioning systems for buildings requiring strict humidity control considering measurement uncertainties, Appl. Energy 261 (2020), 114451, https://doi.org/10.1016/j.apenergy.2019.114451.

[15] J. Wang, J. Huang, Z. Feng, S.J. Cao, F. Haghighat, Occupant-density-detection based energy efficient ventilation system: prevention of infection transmission, Energy Build. 240 (2021), 110883, https://doi.org/10.1016/j.enbuild.2021.110883.

[16] B. Dong, K.P. Lam, Building energy and comfort management through occupant behaviour pattern detection based on a large-scale environmental sensor network, J. Build Perform Simul. 4 (2011) 359–369, https://doi.org/10.1080/19401493.2011.577810.

[17] Y.S. Kim, J. Serebic, Impact of occupancy rates on the building electricity consumption in commercial buildings, Energy Build. 138 (2017) 591–600, https://doi.org/10.1016/j.enbuild.2017.09.014.

[18] Z. Wang, T. Hong, M.A. Piette, Predicting plug loads with occupant count data through a deep learning approach, Energy 181 (2019) 29–42, https://doi.org/10.1016/j.energy.2019.05.138.

[19] T. Ekselvague, N. Brown, V. Pakka, D. Fan, Real-time building occupancy sensing using neural-network-based sensor network, IEEE Int. Conf. Digi. Ecosystem. Technol. (2013), https://doi.org/10.1109/DEST.2013.6611339, 114–9.

[20] C. Guzman, L. Rueda, G. Romero, S. Biscans, K. Agbossou, A. Cardenas, Enabling winter behavior analysis on electrically heated residential buildings by smart sub-metering, Proc. IEEE Int. Conf. Ind. Technol. 2018 (2018), https://doi.org/10.1109/ICT.2018.8352351. Febru:1213-8.

[21] F. Wang, F. Huhe, P. Kurniadi, Instrumentation systems for commercial building energy efficiency, 2011 Futur Instum Int Work FIWII 2011 - Proc 2011:21–4, https://doi.org/10.1109/FIWWII.2011.6476826.

[22] Z. Deng, Q. Chen, Reinforcement learning of occupant behavior model for cross-building transfer learning in various HVAC control systems, Build. Energy 238 (2021), 110860, https://doi.org/10.1016/j.enbuild.2021.110860.

[23] R. Ward, C.S.Y. Wong, A. Chong, R. Choudhary, S. Ramasamy, A study on the transferability of computational models of building electricity load patterns across climatic zones, Energy Build. 237 (2021), 110862, https://doi.org/10.1016/j.enbuild.2021.110826.

[24] S. Wang, J. Burnett, H. Chong, Experimental validation of CO2-based occupancy detection for demand-controlled ventilation, Indoor Built Environ. 8 (2003) 377–391, https://doi.org/10.1002/2050-0579 (2013) 430-3.

[25] F. Wang, Q. Feng, Z. Chen, Q. Zhao, Z. Cheng, J. Zou, et al., Predictive control of indoor environment using occupant number detected by video data and CO2 concentration, Energy Build. 145 (2017) 155–162, https://doi.org/10.1016/j.enbuild.2017.04.014.

[26] T.N. Nguyen, B. Michaelis, A. Al-Hamadi, M. Tornov, M.M. Meinecke, Stereo-camera-based urban environment perception using occupancy grid and object tracking, IEEE Trans. Intell. Transp. Syst. 13 (2012) 154–165, https://doi.org/10.1109/TITS.2011.2165705.

[27] K.H. Chao, P.Y. Chen, An intelligent traffic flow control system based on radio frequency identification and wireless sensor networks, Int. J. Distributed Sens. Netw. 2014 (2014), https://doi.org/10.1155/2014/694545.

[28] R. Seng, S. Boner, K. Maaß, T. Häusserrit, O. Hirsch, N. Donner-Banzhoff, et al., Reliable classification: learning classifiers that distinguish aleatoric and epistemic uncertainty, Inf. Sci. 255 (2014) 16–29, https://doi.org/10.1016/j.ins.2013.07.030.

[29] L. Zhu, N. Laptev, Deep and confident prediction for time series at uber, in: IEEE Int. Conf. Data Min. Work. ICDMW 2017 - Novem, IEEE, 2017, https://doi.org/10.1109/ICDMW.2017.19, 103–10.

[30] F. Maah, K. Nguyen, M. Azam, A. Papachristou, J.S.Venne, M. Cheriart, LSTM-based outdoor air temperature prediction framework for HVAC systems in smart
S. Joshi, J.A. Owens, S. Shah, T. Munasinghe, Analysis of preprocessing techniques, L. Xu, M. Hu, C. Fan, Probabilistic electrical load forecasting for buildings using Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: 33rd Int. Conf. Mach. Learn. ICML 2016, 3, PMLR, 2016, pp. 1651–1660.

L. Xu, M. Hu, C. Fan, Probabilistic electrical load forecasting for buildings using Bayesian deep neural networks, J. Build. Eng. 46 (2022), 103853, https://doi.org/10.1016/j.jobe.2021.103853.

A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. (2017), 2017-Decem:5575-85.

M.K. Park, J.M. Lee, W.H. Kang, J.M. Choi, K.H. Lee, Predictive model for PV power generation using RNN (LSTM), J. Mech. Sci. Technol. 35 (2021) 795–803, https://doi.org/10.1007/s12206-021-040-0.

Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: 33rd Int. Conf. Mach. Learn. ICML 2016, 3, PMLR, 2016, pp. 1651–1660.

L. Xu, M. Hu, C. Fan, Probabilistic electrical load forecasting for buildings using Bayesian deep neural networks, J. Build. Eng. 46 (2022), 103853, https://doi.org/10.1016/j.jobe.2021.103853.

A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. (2017), 2017-Decem:5575-85.

M.K. Park, J.M. Lee, W.H. Kang, J.M. Choi, K.H. Lee, Predictive model for PV power generation using RNN (LSTM), J. Mech. Sci. Technol. 35 (2021) 795–803, https://doi.org/10.1007/s12206-021-040-0.

Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: 33rd Int. Conf. Mach. Learn. ICML 2016, 3, PMLR, 2016, pp. 1651–1660.

L. Xu, M. Hu, C. Fan, Probabilistic electrical load forecasting for buildings using Bayesian deep neural networks, J. Build. Eng. 46 (2022), 103853, https://doi.org/10.1016/j.jobe.2021.103853.