Enhancing Building Energy through Regularized Bayesian Neural Networks for Precise Occupancy Detection and Energy Optimization

Abdullahi Yahaya[1], Abdulhameed Babatunde Owolabi[1,2], Dongjun Suh[1*],

[1]Department of Convergence and Fusion System Engineering, Kyungpook National University, Sangju, 37224, South Korea
[2]Regional Leading Research Center for Smart Energy System, Kyungpook National University, Sangju 37224, South Korea

Email: yahaya@knu.ac.kr[1], owolabiabdulhameed@gmail.com[1,2], dongjunsuh@knu.ac.kr[1*]

Abstract

Accurate occupancy detection is crucial for optimizing building energy management. However, challenges arising from limited data, sensor noise, and intricate dynamics often undermine precision. To address these challenges, we propose a two-step approach. Firstly, we incorporate supplementary features derived from building information to enhance data quality and predictive capacity. Subsequently, a physics-based regularizer is integrated into a Bayesian Neural Network Model, ensuring adherence to constraints and reliable uncertainty estimation, thus elevating prediction precision. Utilizing open-access data from an office building experiment encompassing occupancy profiles, electricity consumption, and indoor environmental data, our approach consistently outperforms conventional models in all test cases, achieving accuracy levels ranging from 96% to 99% when compared to BNN, GBM, SVM, and NB models. Finally, we have developed a design framework to streamline data input, model training, and evaluation through an accessible GUI interface, enabling support for utilizing the model in different building domains. In a world with a growing emphasis on sustainability, this model and framework provide a promising path for accurate occupancy detection, promoting energy conservation, and enhancing energy efficiency.

Keywords: Occupancy detection, Building optimization, Physics-based regularizer, Energy management, Bayesian Neural Networks, Predictive modeling

1. Introduction

The global energy landscape is significantly shaped by buildings, which are major contributors to energy consumption and greenhouse gas (GHG) emissions [1]. In developed regions like the United States (US) and Europe, buildings represent approximately 40% of the total primary energy consumption [2]. Commercial buildings, particularly office spaces, demonstrate significantly high levels of energy use intensity, accounting for approximately 17% of energy consumption within the US commercial building sector [3]. In Asia, where the residential sector ranks second in energy consumption, buildings play a pivotal role. For instance, South Korean buildings alone account for approximately 25% of the nation's total energy consumption [4], while China has recorded a steadily increasing energy usage, with annual increments exceeding 10% [5]. This data underscores the need to address the high energy consumption of buildings as a critical step in mitigating global energy challenges and reducing carbon emissions.

The International Energy Agency (IEA) identifies six key factors determining building energy consumption: occupant behavior, weather conditions, indoor design standards, building envelope, building operation and maintenance, and building energy and service systems [6]. Among the critical energy consumers within buildings, heating, ventilation, and air-conditioning (HVAC) systems stand out, accounting for a substantial portion of overall energy usage [7]. The behavior of occupants significantly influences energy consumption patterns within buildings. For example, HVAC systems rely on estimated occupancy data to control temperatures and airflow effectively [8]. Similarly, lighting systems use occupancy sensing to turn lights on or off in areas like doorways and meeting rooms [9]. Occupancy detection is crucial in building automation systems, regardless of whether they are deployed in commercial or residential settings. Numerous research studies have consistently demonstrated that precise occupant presence detection can result in significant energy savings, typically ranging between 30% to 42% [10]. Precisely assessing and predicting building occupancy status is essential for achieving energy efficiency and sustainable building management. It facilitates the optimal control of energy usage, resulting in reduced energy consumption while ensuring the comfort of occupants. Given the substantial impact of buildings on global energy consumption, it is imperative to prioritize initiatives to mitigate their environmental footprint [11].

1.1. Literature Review

In recent years, there has been a growing focus among researchers on predicting building occupancy, recognizing its substantial influence on energy efficiency. The established correlation between occupant behavior and energy consumption highlights the significance of this area of study [12]. For instance, implementing occupancy-based management systems for lighting and HVAC has demonstrated notable potential for achieving energy savings [13]. Currently, diverse methodologies are employed for estimating building occupancy, encompassing direct and indirect approaches [14]. Direct methods involve the utilization of technologies that directly detect human presence, such as camera sensors dedicated to people counts [15], radio frequency identification (RFID) [16], optical tripwires [17], and Wireless Fidelity (Wi-Fi) [18]. However, direct methods face challenges concerning high hardware costs, privacy concerns, intrusiveness, and installation complexities [19]. Conversely, indirect methods detect occupancy by monitoring changes in environmental parameters such as temperature, CO_2 levels, luminosity, humidity, or sound in indoor spaces. These changes, induced by human activities, can be captured using indirect solutions based on the Internet of Things (IoT) [20], [21], incorporating vibration detectors, microphones, CO_2 concentration monitors, and sensors for temperature, light, and humidity. Indirect methods have demonstrated effectiveness as a viable alternative to direct approaches [22].

Data-driven methodologies show significant promise in enabling accurate building occupancy prediction due to their simplicity and effectiveness [23]. Many researchers have conducted experiments exploring the use of various machine learning techniques such as Neural Networks (NNs) [24], Decision Trees (DTs) [25], Hidden Markov Models (HMMs) [26], and Support Vector Machines (SVMs) [27], to predict occupancy levels in various building scenarios, consistently yielding impressive results. Notably, the research conducted by Ryu et al. [10] employed an indirect approach to data collection. They developed an occupant prediction model using Decision Trees (DT) and Hidden Markov Models (HMM), considering measured electricity consumption from appliances, lighting, and indoor and outdoor CO_2 concentrations within a building. Sayed et al. [28] employed environmental data to infer occupancy information using information theory metrics with ML approaches to determine the best strategy for predicting occupancy patterns. A recent study demonstrated a 20.3% energy saving by implementing occupant-centric demand-driven control compared to conventional baseline control. The findings revealed an inverse relationship between an individual office's energy saving potential and occupancy count [29].

Neural Networks (NNs) have gained popularity for accurately modeling datasets with complex problem structures [30]. However, in critical tasks that demand precise occupant detection within robust engineered systems, where reliance on model output is crucial, the deterministic prediction model provided by traditional NNs may fall short. Bayesian Neural Networks (BNNs) have been used to tackle this challenge. They combine Bayesian inference and deep neural networks to provide high representation power and quantifiable uncertainty estimates

[31]. Substantially, traditional NNs cannot quantify epistemic uncertainty, which pertains to systematic uncertainty arising from model uncertainty. In contrast, BNNs offer a means to quantify this epistemic uncertainty, thereby capturing uncertainty inherent to the models. By incorporating prior distributions on neural network weights, BNNs achieve bias reduction. Furthermore, BNNs facilitate the estimation of a posterior distribution on the predicted model, aiding in quantifying uncertainty in the output. This, in turn, enhances decision-making in downstream tasks [32].

A table summarizing reported accuracies, experimental methodologies, sensors used, and model parameters relevant to the scope of this research is provided in Table 1.

Table 1. Previous reported model accuracies on occupancy detection.

Ref.	Methodology	Parameters/Sensor	Approach	Accuracy
[33]	Statistical classification models were used, including Linear Discriminant Analysis (LDA), Classification Regression Trees (CART), and Random Forest (RF)	Data from light, temperature, humidity, and CO2	Indirect	Accuracies range from 95% to 99%. The LDA, CART, and RF models showed the best results.
[34]	SVM, k-nearest neighbor (KNN), Artificial Neural Network (ANN), Naive Bayesian (NB), tree augmented naive Bayes network (TAN), decision tree (DT).	Twelve ambient sensor variables are used for occupancy modeling, including CO2, door status, and light variables.		Accuracy ranges from 96.0% to 98.2% for single-occupancy and multi-occupancy offices. DT technique yielded the best.
[35]	Radial basis function neural network	Lighting, sound, Reed sensor, C02, temperature, RH, PIR		Accuracy for the number of occupants ranges from 63.23–66.43%
[36]	Support Vector Machine (SVM), K-nearest neighbor (KNN), and thresholding	Electric power consumption (W)	Direct	59-90%
[24]	Gradient boosting, support-vector network, feed-forward neural network, and Deep Neural Network (DNN).	Energy consumption and occupant count		DNN model exhibits slightly better prediction accuracy, obtaining R^2 score of 0.87%

1.2. Research Gap and Contribution

Following a comprehensive review of the literature on energy conservation, HVAC control, and occupancy prediction models, several noteworthy points have been identified, revealing the following gaps:

- The impact of implemented models is often domain-specific, limiting their applicability to other domains for occupancy prediction.
- Experimental data used often involves a relatively small sample size for training the models, which can hinder model accuracy. Research by [37] highlights the importance of data volume in data-driven approaches for accurate model prediction.
- Only a limited number of studies discussed the inherent uncertainty in the data generated from sensors, which can introduce ambiguity in model predictions.
- None of the reviewed literature presented a graphical user interface specifically designed for easy comprehension by energy managers to show the effect of model performance. The focus primarily revolved around feature importance and testing different model accuracy [33].
- Obtaining ground-truth occupancy data is challenging in most studies due to privacy concerns and difficulties in accurately detecting individuals from sensors.
- The literature reveals the need for model improvement, data quality improvement, addressing uncertainty in data, and developing a user-friendly interface for easy interpretation of model performance.

This paper extends the existing literature and bridges the identified research gaps by introducing an innovative approach. Our method uses a BNN model integrated with a Physics-based regularizer (PBR) to develop a highly accurate prediction model for occupant behavior, specifically emphasizing occupancy detection. This approach is designed to capture the distribution of uncertainties while integrating domain-specific environmental data, ultimately enhancing the model's predictive capabilities and ensuring its robust development. This research seeks to make significant contributions to optimizing HVAC control strategies, energy management, and overall building energy efficiency by achieving more accurate occupancy predictions. Here are the primary contributions of this research:

- **Two-Step Approach for Improved Accuracy**: We propose a two-step approach to enhance occupancy prediction accuracy in buildings. In Step 1, we tackle data limitations by merging domain environmental data with building experimental data. In Step 2, we introduce a PBR into a BNN model, significantly improving prediction accuracy.

- **Versatile Model Framework**: We have developed a model framework for building occupant detection that can be effectively applied across various building types. This adaptable framework goes beyond merely estimating current occupancy within a fixed domain, enhancing its practical applicability and relevance.
- **User-Friendly GUI Tool**: A user-friendly Graphical User Interface (GUI) has been designed to showcase the model's real-time performance. This tool empowers energy users, building managers, and stakeholders to understand the impact of different factors on occupant status within a building. It facilitates decision-making and the implementation of effective energy management strategies.

These contributions set this research apart from previous studies and offer advancements in accurately predicting building occupant status, providing valuable insights and practical applications for energy optimization and management. The research framework is illustrated in Figure 1. The other sections of this paper are structured as follows: Section 2 presents the research methodology, including an overview of the data and models. Section 3 discusses the experimental approach. Section 4 discusses the results, including details about the proposed framework. Finally, Section 5 presents the conclusions drawn from this research.

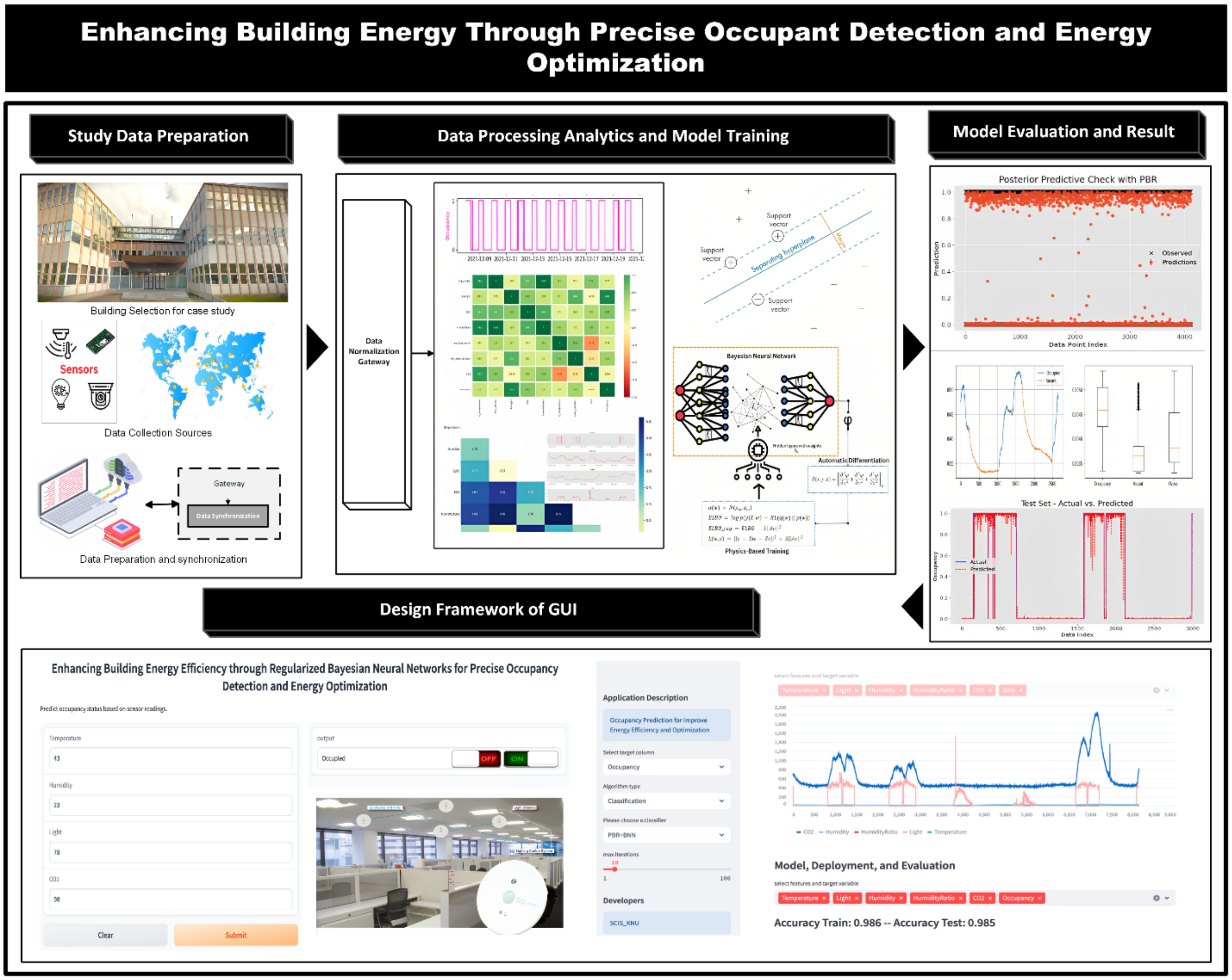

Enhancing Building Energy Through Precise Occupant Detection and Energy Optimization
Study Data Preparation
Data Processing Analytics and Model Training
Model Evaluation and Result
Building Selection for case study
Sensors
Data Collection Sources
Gateway
Data Synchronization
Data Preparation and synchronization
Data Normalization Gateway
Support vector
Separating hyperplane
Bayesian Neural Network
Automatic Differentiation
Physics-Based Training
Posterior Predictive Check with PBR
Prediction
Observed
Predictions
Data Point Index
Test Set - Actual vs. Predicted
Occupancy
Actual
Predicted
Data Index
Design Framework of GUI
Enhancing Building Energy Efficiency through Regularized Bayesian Neural Networks for Precise Occupancy Detection and Energy Optimization
Predict occupancy status based on sensor readings.
Temperature
43
Humidity
23
Light
78
CO2
58
Clear
Submit
output
Occupied
OFF
ON
Application Description
Occupancy Prediction for Improve Energy Efficiency and Optimization
Select target column
Occupancy
Algorithm type
Classification
Please choose a classifier
PBR+BNN
max iterations
10
1
100
Developers
SCIS_KNU
select features and target variable
Temperature
Light
Humidity
HumidityRatio
CO2
date
Model, Deployment, and Evaluation
Occupancy
Accuracy Train: 0.986 -- Accuracy Test: 0.985

Figure 1. Research framework entailing all involved activities of the research

2. Research Methodology

The method for developing the proposed models for occupancy prediction adopts a two-step approach. In the first step, an occupancy detection model is developed using deep learning techniques, specifically the BNN, and includes additional features derived from building information and domain environment factors. The second step further enhances the prediction model by incorporating a physics-based regularizer. This regularizer ensures that the predicted occupant status closely adheres to the physical constraints and principles governing the building environment, thereby improving the model's accuracy. Below, several methodologies relevant to the proposed approach are concisely introduced.

2.1. Data collection and Preparation

The dataset utilized in this research was obtained from a study conducted at the University of Mons in Belgium [33]. This experiment occurred during the winter season in February and involved monitoring an office building. The following variables were under observation: Temperature (°C), Relative Humidity (φ), Light (Lux), and CO_2 (ppm). The office's occupancy status was determined through a digital camera sensor. Data collection was executed through an indirect sensor-based approach utilizing cost-effective monitoring equipment. For a comprehensive understanding of the monitor sensors, the features are described in Table 2. For more in-depth information about the experiment and data generation, please refer to [33]. Figure 2 illustrates the relationship between each data feature and the occupant status within the building.

Table 2. Monitoring sensors equipment **[33].**

Sensors	Parameter	Resolution	Range
Telaire 6613	CO2	1 ppm	0–2000 ppm
DHT22	Humidity	0.1%	0–100%RH
TSL2561	Light	1 Lux	1–40,000
DHT22	Temperature	0.1◦C	−40–80 ◦ C

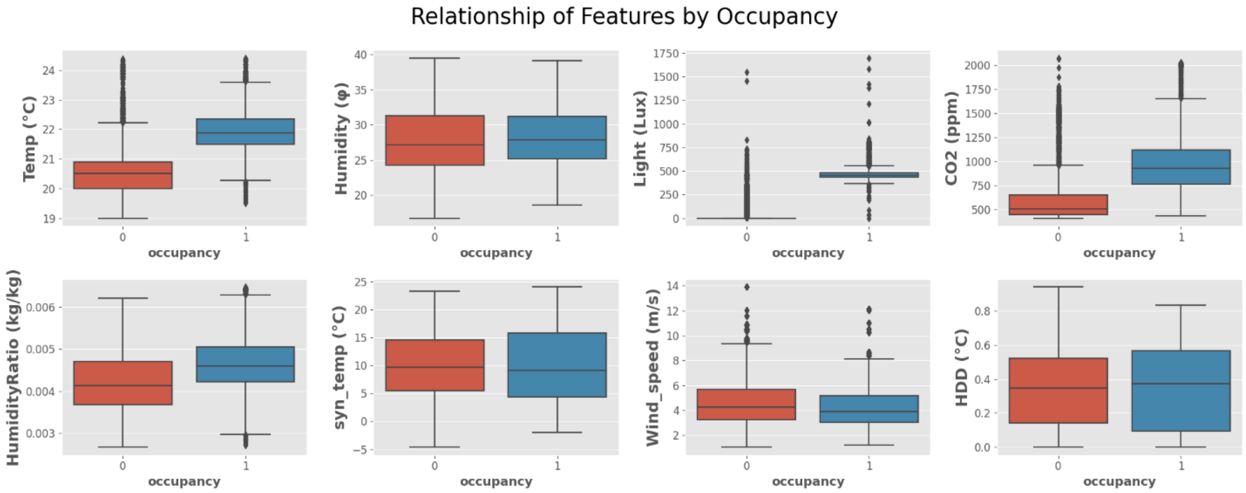


Figure 2. Relationship of Features by Occupancy

2.2. Bayesian Neural Networks (BNNs)

BNNs integrate neural networks and Bayesian probability modeling, aiming to combine their strengths. In BNNs, weights of the neural network are treated as random variables with predefined prior distributions. Unlike standard neural networks, which provide single-point estimates, BNNs generate a distribution over possible outcomes. Bayesian inference is then applied to calculate the posterior distribution over the weights by sampling from this output distribution [38]. Specifically, BNNs use probability distributions for neural network weights, spreading uncertainty across the model. Bayes' rule updates the prior weight distribution based on observed data to get the posterior distribution, capturing learned weight uncertainty. These weight distributions also reveal insights into parameter learning from data [39]. BNNs blend neural networks' representational power and flexibility with the probabilistic interpretations and uncertainty modeling of Bayesian methods. BNNs can offer probabilistic prediction guarantees and generate parameter distributions from observations [40]. This feature helps infer the nature and distribution of neural network parameters within the parameter space. These characteristics make BNNs appealing to both researchers and practitioners.

Notation

Table 3 presents the notations utilized in introducing the basics of BNNs. In contrast to traditional neural networks, BNNs handle model parameters "w" denoted as probability distributions represented by "$w \sim t(\theta)$" where "θ" refers to a deterministic parameter set governing the distribution of "w." This incorporation of the parameter distribution, known as epistemic uncertainty, effectively mitigates the problem of overfitting frequently encountered in NNs [41].

Table 3: Fundamental notation of a BNN

Notation	Definition
$\boldsymbol{a}$	The training inputs
$\boldsymbol{b}$	The prediction outputs
$\boldsymbol{w}$	The BNN parameters, such as weights
$\mathcal{D}$	All training variables, including input and output variables
θ	The parameter vectors for distributions of **w**
a^*	The new inputs
b^*	The new input corresponding to predictions

In this research study, we seek to ascertain the posterior distributions of parameters $\boldsymbol{w}$ denoted as $p(w|\boldsymbol{a},\boldsymbol{b})$ given a set of n input data vectors “ $\boldsymbol{a}$ " and their corresponding outputs " $\boldsymbol{b}$ " Subsequently, an estimation of the actual distribution $p(b|a)$ is made by leveraging the estimated posterior distribution $p(w|a,b)$.

$$\overbrace{p(\mathbf{w} \mid \boldsymbol{a},\boldsymbol{b})}^{\text{Posterior}} = \frac{\overbrace{p(\boldsymbol{b}|\boldsymbol{a},\mathbf{w})}^{\text{Likelihood}}\overbrace{p(\mathrm{w})}^{\text{Prior}}}{\underbrace{p(\boldsymbol{b}|\boldsymbol{a})}_{\text{Evidence}}} \tag{1}$$

In practical applications, the distribution space of parameter "$\boldsymbol{w}$" is commonly specified as a prior, represented by "$p(\boldsymbol{w})$." The likelihood function, denoted as "$p(\boldsymbol{b}|\boldsymbol{a},\boldsymbol{w})$," is derived from the training dataset. The normalizer, "$p(\boldsymbol{b}|\boldsymbol{a})$," referred to as evidence [42], ensures the appropriate evaluation of the posterior distribution of "$\boldsymbol{w}$," denoted as " $p(\boldsymbol{w}|\boldsymbol{a},\boldsymbol{b})$" utilizing Eq. (1). In contrast to deterministic models, BNNs derive their predictive outcomes from the estimated posterior distribution of parameter "$\boldsymbol{w}$" instead of relying on a single deterministic value. The predictive distribution is acquired by performing an integration process over "$\boldsymbol{w}$" as denoted in Eq. (2).

$$p(b^* \mid a^*) = \int p(b^* \mid a^*, \boldsymbol{w}) p(\boldsymbol{w} \mid \boldsymbol{a},\boldsymbol{b}) d(\boldsymbol{w}) \tag{2}$$

As demonstrated in Eq. (2), the predicted output, denoted as b^*, is obtained through the process of integration over parameter " $\boldsymbol{w}$ ". This integration can be viewed as the expectation of $\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{a},\boldsymbol{b})}[p(b^* \mid a^*, \boldsymbol{w})]$, which represents the average prediction over the distribution of "$\boldsymbol{w}$" given the input-output pairs (a, b). However, this expectation is challenging to compute directly for practical cases in neural networks [43].

2.2.1. Variational Inference

Variational inference has emerged as a prominent method for addressing the intractability challenge encountered in posterior inference of BNNs [44]. This technique entails approximating the intricate posterior distribution, $p(\boldsymbol{w} \mid X, Y)$, with a simpler and more manageable distribution, $q(\boldsymbol{w} \mid \theta)$. The quantification of the dissimilarity between these two distributions is achieved through the application of the Kullback–Leibler (KL) divergence, denoted as $KL\big(q(\cdot) \mid\mid p(\cdot)\big)$. By minimizing the KL divergence with respect to the variational parameters θ, the optimal variational distribution $q(\cdot)$ is obtained [45]. In the context of BNNs, a vital component is the Evidence Lower

Bound (ELBO), which serves as a valuable lower bound on the negative log-likelihood function [46]. This ELBO plays a central role in formulating the objective function during the training process, as illustrated in Eq. (3). By optimizing this objective function, BNNs learn to capture the intricate dependencies among the model's parameters, leading to enhanced predictive performance and model interpretability.

$$-\widehat{ELBO} \coloneqq \underbrace{KL\big(q_\theta(w)|p(w)\big)}_{\text{KL divergence}} \quad \underbrace{-\log\big(p(D \mid w)\big)}_{\text{Negative Log-Likelihood (NLL)}} \tag{3}$$

2.3. Physics-based regularizer (PBR)

The PBR constrains the range of mode posteriors to guide the neural network towards behaviors consistent with engineering domain knowledge [47]. To incorporate this framework, a general constraint function, denoted as "$f(\boldsymbol{a}, \boldsymbol{b})$," is introduced for the output variable "$\boldsymbol{b}$". Importantly, "$f(\cdot)$" can be defined based on either the input "$\boldsymbol{a}$" or the output "$\boldsymbol{b}$". The formulation of this method is expressed in Eq. (4) as follows:

$$PBR: \max_{\theta} \mathcal{L}(\theta; \mathcal{D}) + \boldsymbol{w}\big(p(\boldsymbol{b} \mid \mathcal{D}, \theta)\big) \tag{4}$$

where $\mathcal{L}(\theta; \mathcal{D})$ represents the marginal likelihood of "$\mathcal{D}$," and $w(\cdot)$ denotes a regularization function applied to the posterior over latent or output variables "$\boldsymbol{b}$". It is essential to note that the posterior here pertains to a different learning model, rather than the traditional Bayesian posterior. In this context, the PBR is utilized to restrict the output of non-Bayesian method. To extend the applicability of PBR to Bayesian methods, a proposed approach known as RegBays [48] adopts the variational Bayesian inference framework, as illustrated in Eq. (5). This extension enables the integration of PBR principles into the Bayesian model, allowing for enhanced regularization and guidance of predictions through the lens of domain-specific knowledge.

$$\begin{aligned} &\text{RegBayes: } \inf_{q(\mathbf{w})} \mathrm{KL}\big(q(\mathbf{w}) \parallel p(\mathbf{w} \mid \mathcal{D})\big) + V(\boldsymbol{\xi}) \\ &= \inf_{q(\mathbf{w}),\xi} \mathrm{KL}\big(q(\mathbf{w}) \parallel P(\mathbf{w})\big) - \int_x \log P(D \mid \mathbf{w}) q(\mathbf{w}) d\mathbf{w} + V(\boldsymbol{\xi}) \\ &\qquad\qquad \text{s.t.} q(\mathbf{w}) \in \mathcal{P}_{\text{post}}(\boldsymbol{\xi}), \end{aligned} \tag{5}$$

The KL-divergence, denoted as $KL\big(q(w)|p(w \mid \mathcal{D})\big)$, quantifies the dissimilarity between the desired post-data posterior distribution, $q(w)$, over model weights "w", and the actual posterior distribution, $p(w \mid \mathcal{D})$. This measure captures the estimated and actual distribution discrepancy given the observed data. The regularization term "V" is a function of the slack variable ξ (i.e., $V(\xi) = |\xi|_{\beta}$ *or* $V(\xi) = \xi$). This function offers a flexible means to incorporate additional information, such as domain knowledge, into the model. By leveraging this regularization, we can effectively impose constraints on the model weights to align with specific knowledge or requirements. In the scope of this research, we aim to extend the RegBayes methodology into a Bayesian deep learning framework, enabling its application to complex data and various types of knowledge constraints. This extension allows us to address constrained BNNs, where incorporating additional knowledge and handling complex data patterns play crucial roles in achieving enhanced model performance and interpretability.

2.4. Physics-based Bayesian Neural Network (PBR-BNN)

This section demonstrates the core optimization problem addressed in the proposed PBR-BNNs and the corresponding algorithm used for PBR-BNNs. The optimization problem and algorithms are designed to integrate physical constraints and domain knowledge into the BNN framework, thereby enhancing the model's predictive capabilities and interpretability. The central objective is to harness the principles of PBR to bolster predictive Accuracy and integrate specialized domain knowledge into the BNN framework. This approach ensures that the predictions align with engineering principles and physical constraints. The emphasis is on the output constraint, wherein any valid input "$\boldsymbol{a}$" is subject to a constraint based on the conditional expectation of knowledge constraint functions " $f(\boldsymbol{a}, \boldsymbol{b})$ ". This expectation, denoted as "$E[\,f(a, b) \mid \mathcal{D}, w\,]$" pertains to the predictive output "$\boldsymbol{b}$" and input set "$\boldsymbol{a}$", as influenced by the dataset "$\mathcal{D}$" and trained weights "$\boldsymbol{w}$".

The Bayesian estimation procedure is executed while incorporating constraints applied to both the outputs and inputs of the models. To establish feasible constraints on the posterior distribution estimation "$q(\boldsymbol{w})$," the concept of a slack penalty function "$V(\xi)$" was introduced. The introduction of the slack variable "ξ" aims to generalize the constraints, thereby accommodating controlled violations Eq. (6).

$$\min_{q(w),\xi} K\, L[q(w)|P(w)] - E_{q(w)} \log P\,(\,y \mid \mathcal{D}, w\,) + \lambda U(\xi) \tag{6}$$

$$Subject\ to: q(\boldsymbol{w}) \in Q_{\xi}, Q_{\xi} = \left\{q(w): E_{q(w)} E[\,f(x,y) \mid \mathcal{D}, w\,] \geq -\xi\right\}$$

Through the presented formulations, we establish a versatile approach that accommodates diverse levels of constraint enforcement, allowing for robust predictions while accounting for the degrees of occupants' status prediction within the building environments. The PBR-BNN architecture, as illustrated in Figure 3, comprises five hidden layers, each containing 50 neurons. The chosen activation function for these layers is the sigmoid function, which facilitates the transformation of input data. The network's design considers boundary conditions by incorporating the PBR function. This function evaluates the proximity of the neural network's predicted scores at each boundary to the prescribed boundary conditions.

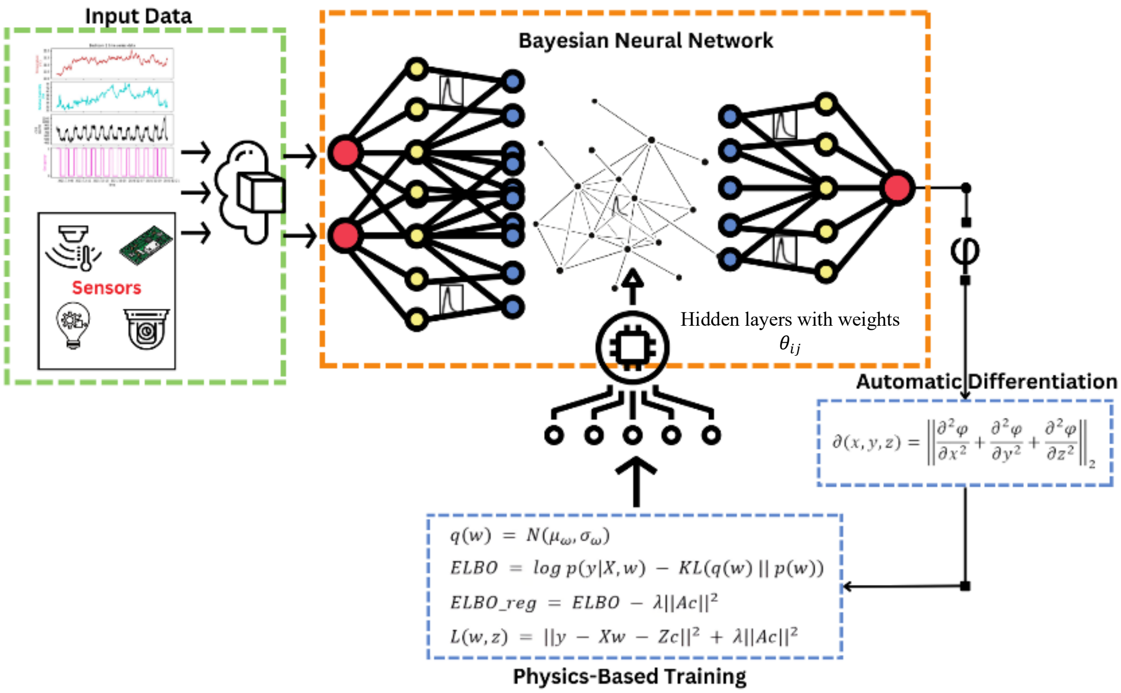

Input Data
Sensors
Bayesian Neural Network
Hidden layers with weights θ_{ij}
φ
Automatic Differentiation
$\partial(x,y,z) = \left\|\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + \frac{\partial^2\varphi}{\partial z^2}\right\|_2$
$q(w) = N(\mu_\omega, \sigma_\omega)$
$ELBO = log\, p(y|X,w) - KL(q(w)\, ||\, p(w))$
$ELBO\_reg = ELBO - \lambda||Ac||^2$
$L(w,z) = ||y - Xw - Zc||^2 + \lambda||Ac||^2$
Physics-Based Training

Figure 3. Architecture of the proposed PBR-BNN during training.

In the context of the physics-based training in Figure 3, each weight within the neural network is assumed to conform to a Gaussian distribution characterized by its mean, "μ," and standard deviation "σ" expressed as $\boldsymbol{w} \sim N(\mu, \sigma)$. To learn the model parameters, a variational inference approach is employed, approximating the posterior distribution over the weights using a Gaussian distribution expressed as $q(w) = N(\mu_w, \sigma_w)$, The parameters of this variational distribution, "μ_w" and "σ_w" are iteratively learned throughout the training. This process minimizes the ELBO objective function, denoted in Eq. (7).

$$ELBO = log\, p(y|X, w) - KL(q(w)\, ||\, p(w)) \tag{7}$$

where $p(y|X, w)$ signifies the likelihood of the data given the weights, and $KL(q(w)\, ||\, p(w))$ represents the Kullback–Leibler divergence between the variational distribution "$q(w)$" and the prior distribution "$p(w)$" over the weights. To ensure the integration of the PBR, we introduce an additional term into the Evidence Lower Bound (ELBO) function. This term functions as a penalty, encouraging the model to comply with the energy conservation constraint. We express this augmented objective function as Eq. (8):

$$ELBO_{reg} = ELBO + \lambda\left|\left|Ac\right|\right|^2 \tag{8}$$

where "A" signifies a matrix encapsulating the relationships between supplementary features and the model coefficients "w" as previously detailed. The parameter λ controls the strength of the regularizer while $\left|\left|Ac\right|\right|^2$ denotes the squared Euclidean norm of the deviation from the energy conservation constraint.

Finally, combining both equations, we introduce a generalized equation encompassing the PBR for occupancy detection Eq. (9).

$$L(w, z) = \left|\left|y - Xw - Zc\right|\right|^2 + \lambda\left|\left|Ac\right|\right|^2 \tag{9}$$

where 'y' represents the vector of occupancy labels, 'X' is the matrix of input features, 'w' signifies the vector of model coefficients, 'Z' denotes the matrix of additional features,'c' is the vector of coefficients establishing the relationship between additional features Z and model coefficients w, 'A' stands for the matrix encoding the physical associations between these features, and 'λ' acts as the hyperparameter controlling the strength of the regularizer.

2.4.1. Algorithm of the proposed PBR-BNN

During model implementation, we emphasized aligning the model with domain-specific knowledge. This alignment was realized by imposing constraints at each distribution point. This approach guaranteed the precise determination of occupants' statuses in full compliance with the specified constraints. A detailed explanation of the implementation algorithm is provided in Table 4.

Table 4: Algorithmic approach to Physics-based Bayesian neural network

```
Bayesian Neural Network with Physics-Based Regularizer for Occupancy Detection
1.  Prepare and process datasets for model training and testing.
2.  Define the Bayesian Neural Network model architecture:
      - Specify the activation function, number of layers, and number of neurons
3.  Initialize the parameters and hyperparameters of the model:
      - Set the prior weight, variance, likelihood function, epoch, and learning rate.
      - Given $y = f(x, \theta)$, where $y$ is the predicted occupancy, $x$ is the input features, and $\theta$
      denotes the model's parameters.
4.  Define the physics-based regularizer to enforce domain-specific constraints:
      - Construct the objective function to enforce model penalty and prevent overfitting using
      Equation (6) and Equation (9)
5.  Implement Bayesian inference technique to capture model uncertainty:
      - Update the ELBO using Bayesian inference
6.  Initialize $k = 0$.
While k $<$ $k_{max}$ do
      - Estimate the loss for each prediction loop $k$.
      - Update the weights from $\theta_n$ to $\theta_{n+1}$ using stochastic gradient descent algorithm
        - Increment $k$ by 1.
End while
7.  Optimize the combined loss function in training the model:
      - Define the loss function $L(\theta) = L_{data}(\theta) + \lambda R(\theta)$, where $L_{data}(\theta)$ represents the data
      loss, $\lambda$ is the regularization strength, and R(θ) is the regularizer term.
8.  Evaluate the model's performance:
      - Assess the model using appropriate evaluation metrics.
9.  Deploy the trained model in a production environment for real-time occupancy predictions.
```

3. Experimental Methodology

The implementation process of this model is visually depicted in Figure 4, illustrating the systematic flow of steps involved in harnessing the predictive power and uncertainty modeling capabilities of BNNs. Throughout the implementation of this research, the processing trajectory is segmented into four distinct phases. Each phase holds significance within the overall processing stage and contributes to the exploration of the proposed framework. Below, we provide an explanation of each of these phases:

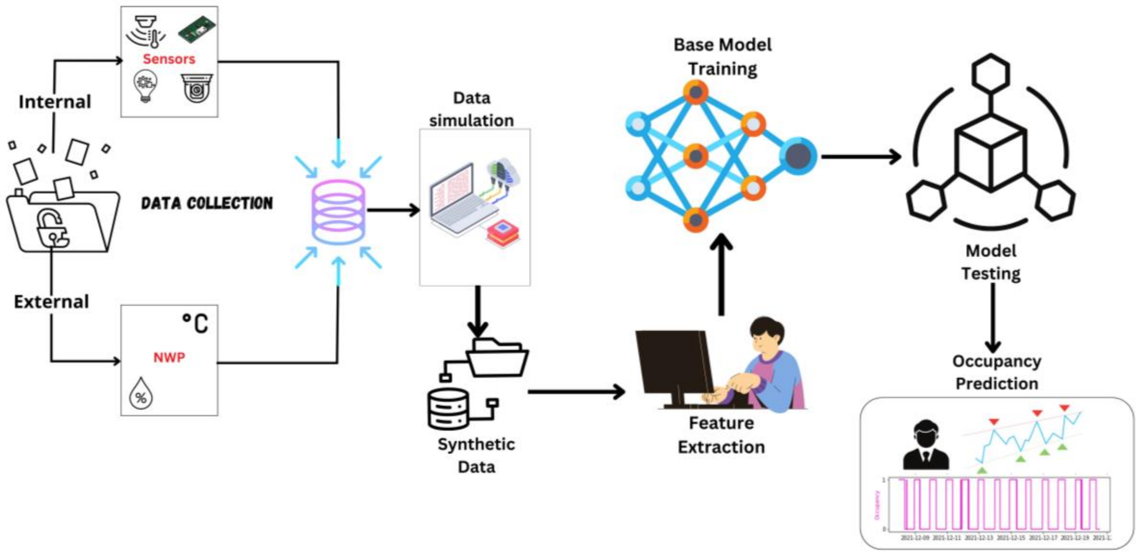


Figure 4. Process flow of implementing the model

Phase 1: Date Preparation, Sorting, and Merging Stage.

In this initial phase, the focus was on enriching the available dataset with domain-specific environmental and building-related energy information extracted from the office building's description and the timing of the experiment. The objective was to augment the data pool and enhance its quality, thereby bolstering the model's predictive capability with synthetic features. The data collection experiment was carried out during the winter season when the indoor climate of the office building was carefully controlled to ensure the comfort of its occupants. Consequently, two pivotal metrics, namely heating degree days (HDD) and cooling degree days (CDD), assumed significance due to their substantial impact on occupants' conditions within a building [47]. HDD quantifies the extent of heating required to maintain optimal indoor temperature based on external weather conditions. For the computation of HDD values, environmental parameters such as temperature and wind speed were acquired, with the average temperature serving as the basis for HDD calculation. By adhering to the ASHRAE-recommended base temperature of 18°C [49], the cumulative degree days for the office building were deduced through the summation of absolute temperature deviations from the prescribed base temperature for each day of experimentation, as denoted in Eq. (10).

$$\mathrm{HDD} = \sum_{i=1}^{n} |T_a - T_b| \qquad (10)$$

The symbol $\sum_{i=1}^{n}$, denotes a summation notation implying the aggregation of values across each day, ranging from 1 to n. The expression $|T_a - T_b|$ represents the absolute discrepancy between the mean temperature (T_a) and the reference base temperature (T_b). Utilizing the absolute value ensures that the outcome remains non-negative, irrespective of whether the mean temperature is higher or lower than the base temperature.

Phase 2: Data Preprocessing and Feature Extraction

This phase aimed to uncover insights into the relationships among data attributes, as highlighted in Figure 2. This visual representation vividly illustrates the correlation between input data and the distribution patterns of occupancy status. It's worth noting that certain features within the dataset exhibited outliers, particularly in the Light and CO2 variables. To enhance data coherence, we took a structured approach. First, each feature underwent normalization. Next, we identified and subsequently removed outliers. This involved retaining only those data points where feature values fell within a range of three standard deviations from the mean. This measure was implemented to avoid the potential impact of abnormal values that could otherwise skew the model's inference based on the dataset.

Phase 3: Feature Selection and Model Training

After eliminating outliers and extracting significant data attributes, we conducted an extensive correlation analysis to identify features strongly linked to occupancy data. To optimize model efficiency in identifying vital features, we employed the Spearman rank correlation coefficient (Eq. 11) [50]. Features with absolute correlation coefficients exceeding 0.3 were selected for inclusion in model training. Subsequently, we divided the combined dataset into two subsets: a training set, which accounted for 70%, and a testing set, constituting 30%. The model underwent training using this dataset, and we graphically depicted the train-validation loss outcomes in Figure 5.

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(M_i - \bar{M})^2} \tag{11}$$

Here, the calculation results of the correlation coefficient are encapsulated within Eq. 11. wherein 'r' symbolizes the coefficient outcomes. 'y_i' represents the actual occupancy values pertinent to the prediction tasks, while 'M_i' corresponds to the values of each feature. Further clarification stems from $\bar{y}$ and $\overline{\mathrm{M}}$, signifying the average values of 'y_i' and 'M_i', respectively. It is noteworthy that the proximity of the absolute value of r to 1 or -1 indicates a heightened degree of correlation.

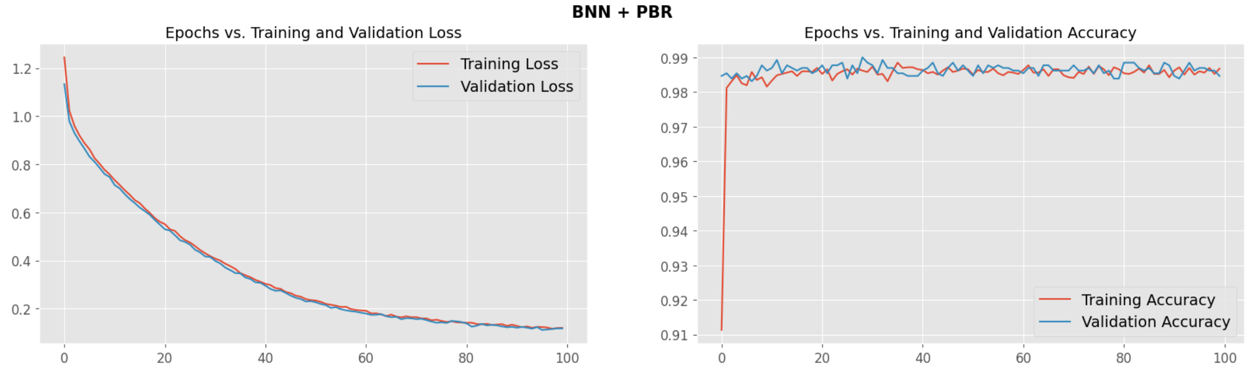


Figure 5. Graphical validation of training loss vs. training accuracy using the proposed model

Phase 4: Model Testing and Validation

In this phase, the trained model underwent testing using a designated subset of the dataset specifically allocated for evaluation. We formulated three distinct testing scenarios to evaluate the model's performance utilizing the dataset detailed in Table 5. The results obtained from these assessments are comprehensively analyzed and discussed in Section 4.

Table 5. Description of the testing cases for model performance evaluation

Cases	Evaluation Description
Case Test 1	Evaluation with the dataset excluding synthetic features
Case Test 2	Evaluation with comprehensive dataset including synthetic features
Case Test 3	Evaluation considering door status (open and close)

Test Case 1 was assessed using the unaltered experimental dataset, while Test Case 2 utilized the merged dataset with synthetic features, as described in Phase 1 of Section 3. In contrast, Test Case 3 incorporated a dataset that accounted for doors' opening and closing status throughout the experiment. It's worth emphasizing that the models were evaluated across all three test cases, each involving various time intervals. The same input parameters were utilized for each model across the test cases to ensure fair and unbiased comparisons. The proposed PBR-BNN underwent training in conjunction with established models, specifically the Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and Naive Bayes (NB). These selections were made deliberately, considering their widespread recognition and demonstrated effectiveness within the scope of this study.

GBM recognized for its proficiency in handling complex datasets and achieving high predictive precision, achieves this by sequentially combining multiple "weak" models, typically decision trees. Subsequent models correct errors made by their predecessors, iteratively improving overall prediction accuracy [51]. GBM's capacity to learn from mistakes and cultivate a resilient ensemble of models renders it an asset to this study. Conversely, SVM is a widely adopted model in classification tasks, primarily due to its ability to identify optimal hyperplanes that effectively separate distinct classes within datasets. By maximizing the margin between classes, SVM enhances the model’s capacity to discern data patterns [52]. Its versatility in handling linear and non-linear data separation and its demonstrated effectiveness across various real-world applications underscores SVM's significance and reliability as a benchmark in this research. Finally, although a simple model, NB excels in classification and regression tasks. NB simplifies computations by leveraging Bayes' theorem and assuming feature independence [53]. Its effectiveness and precision underscore its importance as a comparative model alongside the proposed model in this research. By utilizing the strengths of GBM, SVM, and NB, this study endeavors to achieve robust and precise predictions compared to the proposed PBR-BNN for the given task.

3.1. Model Evaluation

The model's performance across each test case was evaluated using the following metrics:

3.1.1. Root Mean Squared Error (RMSE)

RMSE quantifies the accuracy of mean prediction by calculating the square root of the average squared difference between predicted and actual values. A model performs better when its RMSE score approaches zero [54].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(XE_i - YE_i)^2} \tag{12}$$

3.1.2. Mean Absolute Error (MAE)

MAE evaluates prediction accuracy by measuring the absolute discrepancies between predicted and actual values [55]. A lower MAE score signifies improved model performance.

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|XE_i - YE_i| \tag{13}$$

3.1.3. Prediction Accuracy

This metric gauges the correctness of predicted occupancy statuses relative to the total number of predictions [55].

$$Acc \ = \frac{P_{exact}}{P_i} \tag{14}$$

Where:

P_{exact} represents the count of exact predicted values aligned with the actual ground truth occupancy status and P_i denotes the total number of predictions.

3.1.4. Coefficient of determination (R^2)

R^2 provides insight into how much of the variation observed in the response variable can be attributed to the predictor variables within the model [55]. This metric generates a value between 0 and 1. A value of 0 indicates that the predictor variables cannot explain any variations in the response variable. In contrast, a value of 1 suggests that the predictor variables can completely account for all the variability without any error.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(X-Y)^2}{\sum_{i=1}^{N}\left(Y-Y_{avg}\right)^2} \tag{15}$$

4. Result and Discussion

Before discussing the outcomes of each model, it's essential to provide details about the technical setup. The simulations and model training were performed on a desktop computer featuring an Intel Extreme Processor i9-7900X Skylake, a NVIDIA GTX1080 Ti 11GB-4WAY graphics card, and 64GB of RAM. All models were implemented using Python programming language, and the corresponding code to replicate the analysis are available on this Github.

During experimentation, each model underwent multiple simulations before recording their respective results. Significantly, the proposed PBR-BNN model demonstrated superior performance compared to the other models across all test cases. Detailed insights into the performance of each model for individual test cases are detailed in Table 6. The hierarchical ranking of the model's performance outcomes across each test case is outlined as follows:

Results of Test Case 1: Following dataset preprocessing and model training, performance evaluation of each model unveiled that the proposed model (PBR-BNN) showcased exceptional performance, achieving an Accuracy of 99.1%. It demonstrated an RMSE of 0.888, an MAE of 0.135, and an R^2 score of 0.955, thereby establishing itself as the top-performing model. In the hierarchy of performance, the BNN model followed closely, securing the second spot with an Accuracy of 99%, accompanied by an RMSE of 0.935, a MAE of 0.144, and an R^2 score of 0.950. Subsequently, the GBM model emerged as the third-best contender, showcasing an Accuracy of 98.6%. Trailing behind, the NB model exhibited a respectable Accuracy score of 96.5%. Lastly, the SVM model recorded the lowest performance within this category, achieving an Accuracy of 95.9%.

Results of Test Case 2: The PBR-BNN model demonstrated its superiority by surpassing all other models in this specific test case. It achieved an exceptional Accuracy of 99.2%, accompanied by a minimal MAE of 0.098, an RMSE of 0.873, and an impressive R2 of 0.967. The GBM model secured the second position in the performance hierarchy, boasting an Accuracy of 98.9%, slightly surpassing the SVM model, which garnered an Accuracy score of 98.8%. The BNN model, although still competitive, landed in the fourth position for this test case, attaining an Accuracy of 98.3%. Conversely, the NB model ranked at the lower end of the spectrum, registering an Accuracy of 97.9%.

Results of Test Case 3: Results of Test Case 3: The model performance evaluation in this particular case is based on two distinct scenarios: open and closed-door statuses. Notably, when the model underwent testing with the open-door status, the proposed PBR-BNN model continued to demonstrate its excellence, achieving an impressive Accuracy score of 97.4%. Similarly, in the closed-door test scenario, it attained an Accuracy of 96.4%. Securing the second position, the GBM model maintained its competitive stance with an Accuracy of 95.5% in the open-door test and 95.7% in the closed-door test. Moving down the hierarchy, the SVM model claimed the third spot during the open-door test, showcasing an Accuracy of 94.8%. However, for the closed-door test, the SVM model shifted to the fourth position, achieving an Accuracy of 92.5%.

Meanwhile, the NB model achieved an accuracy of 94.7% during the open-door test, securing the fourth rank in the performance hierarchy. Furthermore, the NB model emerged as the third-best performing model in the closed-door test, boasting an accuracy of 94.2%. In contrast, the BNN model exhibited relatively lower performance, with an Accuracy of 92.4% for the open-door test and 90.1% for the closed-door test. This placed the BNN model as the least-performing model in this test case.

Table 6. Model performance comparison.

Test Cases	Metrics	GBM	NB	SVM	BNN	PBR-BNN
Case Test 1	RMSE	1.188	1.865	2.015	0.935	**0.888**
	MAE	0.141	0.348	0.406	0.144	**0.135**
	R^2	0.919	0.800	0.766	0.950	**0.955**
	Accuracy	98.6%	96.5%	95.9%	99.0%	**99.1%**
Case Test 2	RMSE	1.066	1.451	1.081	1.292	**0.873**
	MAE	0.114	0.210	0.117	0.167	**0.098**
	R^2	0.950	0.908	0.949	0.927	**0.967**
	Accuracy	98.9%	97.9%	98.8%	98.3%	**99.2%**
Case Test 3	RMSE	2.113 2.068	2.300 2.403	2.267 2.738	2.753 2.789	**1.644** **1.691**
	MAE	0.447 0.428	0.529 0.577	0.514 0.750	0.758 0.795	**0.406** **0.414**
	R^2	0.807 0.742	0.772 0.652	0.778 0.548	0.673 0.531	**0.883** **0.828**
	Accuracy	95.5% 95.7%	94.7% 94.2%	94.8% 92.5%	92.4% 90.1%	**97.4%** **96.4%**

In summary, the proposed PBR-BNN model consistently demonstrated remarkable accuracy levels exceeding 96% across all test cases. This exceptional performance can be attributed to the fundamental concept introduced in this research. The model's deep understanding of the physical environment, combined with the incorporation of a regularizer, undeniably contributed to its outstanding success. This is particularly evident when compared to the BNN model without the regularizer, which displayed inconsistent results across various test cases. Furthermore, the GBM model showcased consistency in its accuracy, likely due to its capacity to blend multiple models during training. However, even with this strength, the GBM model could not outperform the proposed model in any of the test cases. This underscores the inherent efficiency of the proposed model's capabilities and its ability to achieve exceptional results.

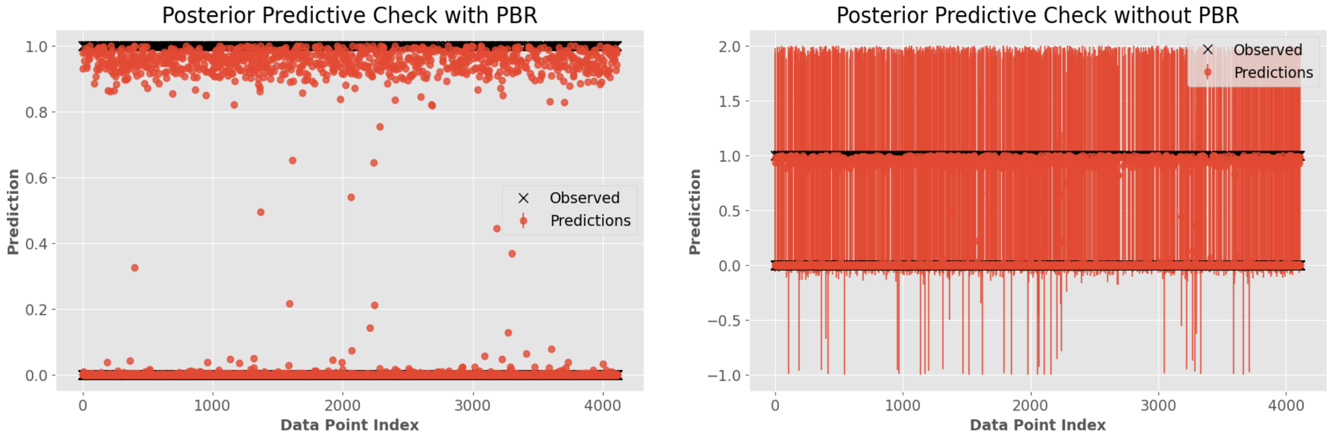
4.1. Graphical Validation

To offer more comprehensive validation and explore the impact of physics-based constraints imposed by the regularizer on the model's predictive performance, we compared two variants of the BNN model. One variant included the PBR, while the other variant had no form of regularization. We created plots to visually illustrate the significant difference in performance between these two models. The generated plots include:

Plot of Posterior Predictive Check

The posterior predictive check is a crucial analysis that evaluates the performance and credibility of a BNN model. This assessment involves comparing the posterior predictive distribution generated by the BNN model with the observed data. The proximity between the posterior predictive distribution and the observed data points indicates the model's effectiveness in capturing underlying patterns within the provided dataset. Moreover, a tightly concentrated and well-calibrated distribution signifies the model's confidence in its predictions, while a broader distribution signifies more significant uncertainty.

To validate the variance of both models, we present Figure 6, which illustrates the distribution of each posterior value (depicted in red) for every sample point compared to the actual data sample (displayed in black). As evident in Figure 6(a), where PBR is present, the y-axis signifies the distribution range for each posterior value during the training phase. Hence, all posterior distribution is guided within this range to ensure accurate prediction. However, in Figure 6(b), where the regularizer is absent during training, the BNN model struggles to accurately guide the posterior distribution in making predictions. The y-axis frequently extends beyond the actual data range, signifying a lack of confidence in the model's output compared to the scenario when the PBR is employed. This inconsistency is reflected in the overall performance ranking of the BNN model.

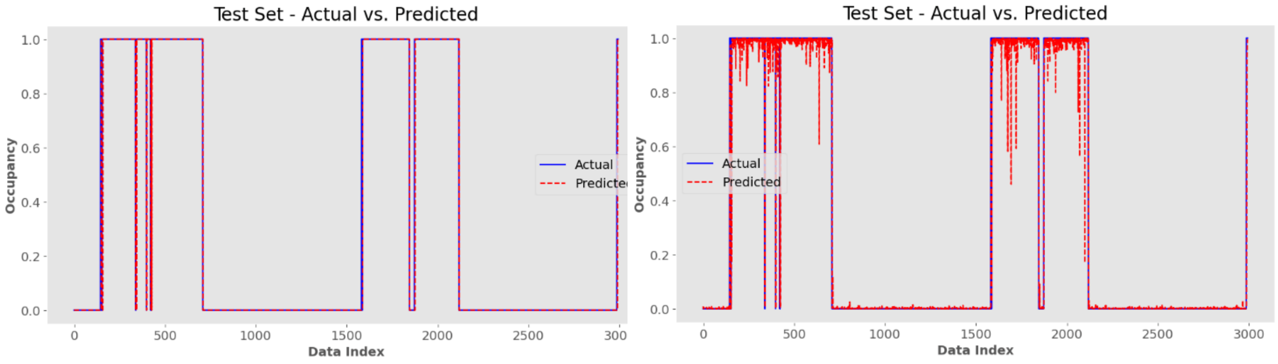


(a) (b)

Figure 6. Check of predictive posterior

Plot of Predicted vs. Actual Values

The purpose of this plot is to evaluate the model's performance through a visual representation of the alignment between actual and predicted values. The transparency of markers ensures clear visualization, especially when data points overlap. Figure 7 compares the actual values (ground truth) and the predicted values obtained from the models. In Figure 7(a), the PBR-BNN model demonstrates a tighter cluster of data points around the diagonal line, indicating enhanced predictive Accuracy. In contrast, the model without regularization exhibits prediction variability, where the predicted curve does not align closely with the actual curves, as shown in Figure 7(b).



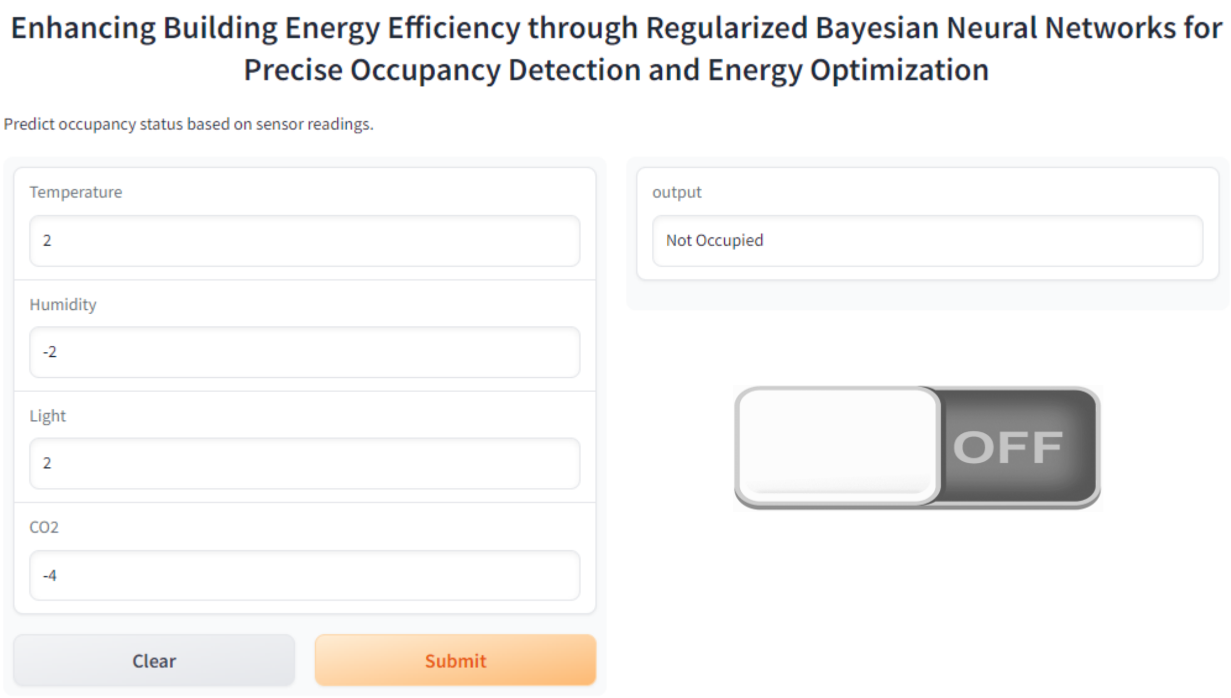
(a) (b)

Figure 7. Actual vs. predicted plot from the trained model

4.2. Design Framework Proposition

In line with the research objectives, prioritizing the smooth integration and straightforward interpretation of the proposed model's performance, we developed the framework GUI in two distinct forms. This framework encompasses various functionalities, including data feature visualization, model training, model evaluation, and occupancy status reporting. This framework aims to make the operational process easily understandable and facilitate the seamless integration of the research findings into a model predictive controller for effectively managing and optimizing energy consumption across diverse domains. The design is structured to accept real-time input variables, as illustrated in Figure 8. Each input corresponds to essential features pertinent to occupant detection within the data-driven approach. Upon input submission, the 'Submit' button encapsulates the proposed model as a functional entity, which evaluates each input variable. Subsequently, it generates an output of the occupant's status, whether 'occupied' or 'not occupied'.

The second form of the designed framework empowers users to upload datasets, conduct diverse exploratory data analyses, choose from the models trained in this research, and assess their performance based on selected features (Figure 9). This design ensures replicability across diverse domains, effectively addressing the challenges of validating and testing models in new environments. A demonstration of this generalized framework has been made available on a web server and is accessible through the provided links (refer to the appendix).



Enhancing Building Energy Efficiency through Regularized Bayesian Neural Networks for Precise Occupancy Detection and Energy Optimization

Predict occupancy status based on sensor readings.

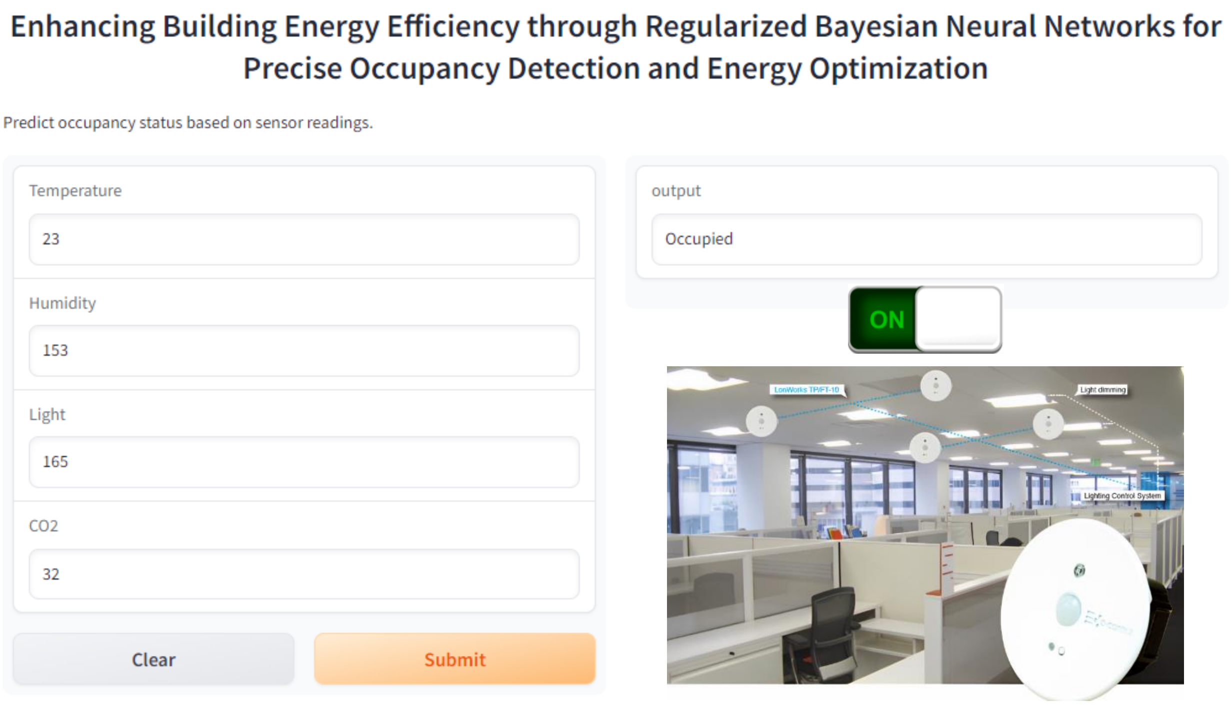
Temperature: 2

Humidity: -2

Light: 2

CO2: -4

output: Not Occupied



Enhancing Building Energy Efficiency through Regularized Bayesian Neural Networks for Precise Occupancy Detection and Energy Optimization

Predict occupancy status based on sensor readings.

Temperature: 23

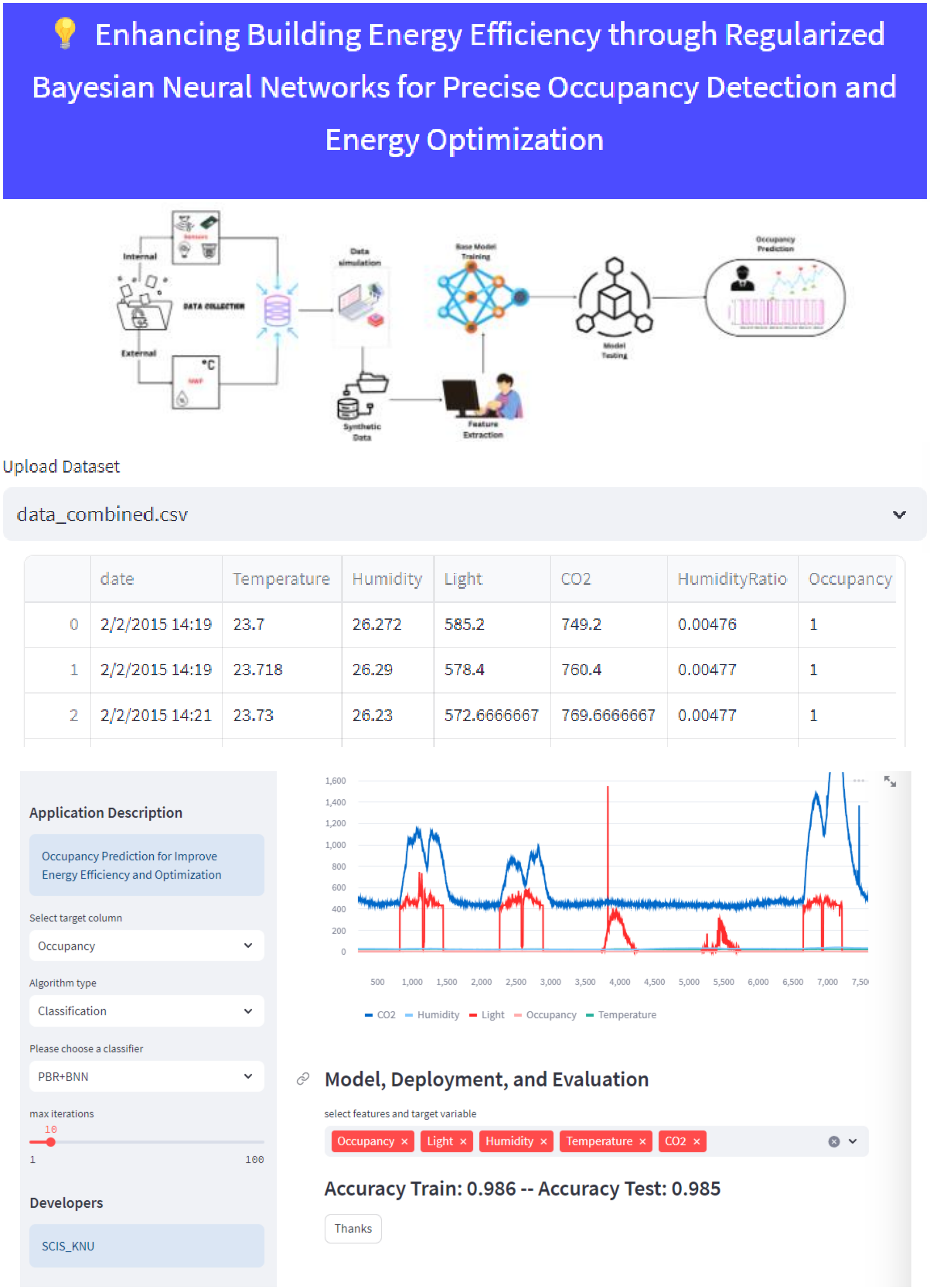
Humidity: 153

Light: 165

CO2: 32

output: Occupied

Figure 8. Real-time prediction of occupancy status



💡 Enhancing Building Energy Efficiency through Regularized Bayesian Neural Networks for Precise Occupancy Detection and Energy Optimization

Upload Dataset

data_combined.csv

	date	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
0	2/2/2015 14:19	23.7	26.272	585.2	749.2	0.00476	1
1	2/2/2015 14:19	23.718	26.29	578.4	760.4	0.00477	1
2	2/2/2015 14:21	23.73	26.23	572.6666667	769.6666667	0.00477	1

Application Description

Occupancy Prediction for Improve Energy Efficiency and Optimization

Select target column: Occupancy

Algorithm type: Classification

Please choose a classifier: PBR+BNN

max iterations: 10 (1 – 100)

Developers

SCIS_KNU

Model, Deployment, and Evaluation

select features and target variable: Occupancy × Light × Humidity × Temperature × CO2 ×

Accuracy Train: 0.986 -- Accuracy Test: 0.985

Thanks

Figure 9. Robust framework design encompassing all models supporting other domains

4.3. Comparative Analysis

To better grasp the significant implications of the outcomes derived from this research, it is essential to conduct a comparative analysis with similar studies, especially those utilizing the same dataset. In a prior study [33], the primary focus was introducing the dataset rather than establishing a predictive model. However, statistical models were applied to the dataset, with the most proficient model being the random forest, achieving an accuracy of 95%. %. Notably, the accuracy of all models employed in this research falls within the range of 96% to 99%, surpassing that of [33]. This improvement is particularly evident in Test Case 1, which aligns with using the same dataset. Likewise, in another study [22], researchers introduced a hybrid model that utilized federated learning and Long Short-Term Memory (LSTM) architecture with the same dataset, achieving an accuracy of 94%. In contrast, our proposed model in this research, particularly in Test Case 1, achieves a substantially higher accuracy score of 99%, clearly outperforming the model presented in [22].

Another study [56] also introduced a synthetic dataset to train a convolutional network with a deep bidirectional long short-term memory (DBLSTM) model. They implemented the concept of transfer learning by applying the same model to the original dataset at three different sample sizes, with the highest recorded accuracy being 93%. In comparison, when we applied our proposed model to Test Case 2, which employs a similar dataset concept, we achieved an accuracy of 99%. This result underscores the clear superiority of our proposed model over the one presented in [56]. This research presents a versatile design framework applicable across diverse domains, with its core goal being the precise detection of building occupancy status, thereby enhancing energy management and optimization. Given the novelty of this approach, making direct comparisons with existing literature proves challenging. Nevertheless, the innovative and comprehensive nature of the proposed methodology sets it apart from previous studies.

5. Conclusion

This research introduces an experimental approach that presents a framework for detecting occupant statuses within buildings, reinforced by a PBR-BNN model, with significant implications for energy optimization. The model consistently demonstrates exceptional performance across various testing scenarios, highlighting its superiority by providing accurate predictions, adhering to constraints, and offering reliable uncertainty estimations. This improved accuracy and dependability are attributed to infusing physics-based insights into the training process, making it suitable for real-world applications that demand precision while adhering to physical laws. Furthermore, the design framework proposed in this research enhances the understanding of the model and facilitates its replicability. An interactive GUI was developed to facilitate data input, model training, and performance evaluation, making it adaptable for practical applications across diverse domains. In pursuing energy optimization, this study provides a clear path by accurately determining occupant statuses, emphasizing transparency, robustness, and real-world applicability. The combined model and framework offer a promising solution for enhancing energy efficiency in building management. To further validate and extend the relevance of this data-driven approach, we recommend future research exploring additional datasets, different building types, and seasonal variations.

Data availability

The data and code are available on https://zenodo.org/record/8073333 (this link is provided in the paper).

References

[1] W. Wuxia, W. Zhang, P. Tien, J. Calautit, and Y. Wu, ‘Building Occupancy Prediction Through Machine Learning for Enhancing Energy Efficiency, Air Quality and Thermal Comfort: Review and Case Study’, Volume 15: Low Carbon Cities and Urban Energy Systems: Part IV, preprint, Nov. 2021. doi: 10.46855/energy-proceedings-8314.

[2] A. Babatunde Owolabi, D. Suh, and G. Pignatta, ‘Investigating the energy use in an Australian building: A case study of a west-facing apartment in Sydney’, *Ain Shams Engineering Journal*, vol. 14, no. 8, p. 102040, Aug. 2023, doi: 10.1016/j.asej.2022.102040.

[3] D. Mora, G. Fajilla, M. C. Austin, and M. De Simone, ‘Occupancy patterns obtained by heuristic approaches: Cluster analysis and logical flowcharts. A case study in a university office’, *Energy and Buildings*, vol. 186, pp. 147–168, Mar. 2019, doi: 10.1016/j.enbuild.2019.01.023.

[4] D. K. Hwang, J. Cho, and J. Moon, ‘Feasibility Study on Energy Audit and Data Driven Analysis Procedure for Building Energy Efficiency: Benchmarking in Korean Hospital Buildings’, *Energies*, vol. 12, no. 15, Art. no. 15, Jan. 2019, doi: 10.3390/en12153006.

[5] P. Xu, E. H.-W. Chan, and Q. K. Qian, 'Success factors of energy performance contracting (EPC) for sustainable building energy efficiency retrofit (BEER) of hotel buildings in China', *Energy Policy*, vol. 39, no. 11, pp. 7389–7398, Nov. 2011, doi: 10.1016/j.enpol.2011.09.001.

[6] EIA. Annual Energy Review, 'Energy Information Administration - EIA - Official Energy Statistics from the U.S. Government'. http:// www.eia.doe.gov/aer/pdf/aer.pdf (accessed May 31, 2023).

[7] Y. Zhou, J. Chen, Z. (Jerry) Yu, J. Zhou, and G. Zhang, 'Short-term building occupancy prediction based on deep forest with multi-order transition probability', *Energy and Buildings*, vol. 255, p. 111684, Jan. 2022, doi: 10.1016/j.enbuild.2021.111684.

[8] J. Lu *et al.*, 'The smart thermostat: using occupancy sensors to save energy in homes', in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, in SenSys '10. New York, NY, USA: Association for Computing Machinery, Nov. 2010, pp. 211–224. doi: 10.1145/1869983.1870005.

[9] X. Guo, D. Tiller, G. Henze, and C. Waters, 'The performance of occupancy-based lighting control systems: A review', *Lighting Research & Technology*, vol. 42, no. 4, pp. 415–431, Dec. 2010, doi: 10.1177/1477153510376225.

[10] S. H. Ryu and H. J. Moon, 'Development of an occupancy prediction model using indoor environmental data based on machine learning techniques', *Building and Environment*, vol. 107, pp. 1–9, Oct. 2016, doi: 10.1016/j.buildenv.2016.06.039.

[11] A. N. Sayed, Y. Himeur, and F. Bensaali, 'Deep and transfer learning for building occupancy detection: A review and comparative analysis', *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105254, Oct. 2022, doi: 10.1016/j.engappai.2022.105254.

[12] V. Oikonomou, F. Becchis, L. Steg, and D. Russolillo, 'Energy saving and energy efficiency concepts for policy making', *Energy Policy*, vol. 37, no. 11, pp. 4787–4796, Nov. 2009, doi: 10.1016/j.enpol.2009.06.035.

[13] J. Kim and J. Drgoňa, *LSTM-based Space Occupancy Prediction towards Efficient Building Energy Management*. 2020.

[14] F. Viani, A. Polo, F. Robol, G. Oliveri, P. Rocca, and A. Massa, 'Crowd detection and occupancy estimation through indirect environmental measurements', in *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, Apr. 2014, pp. 2127–2130. doi: 10.1109/EuCAP.2014.6902229.

[15] D. Liu, Y. Du, Q. Zhao, and X. Guan, 'Vision-based indoor occupants detection system for intelligent buildings', in *2012 IEEE International Conference on Imaging Systems and Techniques Proceedings*, Jul. 2012, pp. 273–278. doi: 10.1109/IST.2012.6295489.

[16] A. Oka and L. Lampe, 'Distributed target tracking using signal strength measurements by a wireless sensor network', *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 7, pp. 1006–1015, Sep. 2010, doi: 10.1109/JSAC.2010.100905.

[17] J. Hutchins, A. Ihler, and P. Smyth, 'Modeling count data from multiple sensors: a building occupancy model', in *2007 2nd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, IEEE, 2007, pp. 241–244.

[18] W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, 'A New Intrusion Detection System Based on KNN Classification Algorithm in Wireless Sensor Network', *Journal of Electrical and Computer Engineering*, vol. 2014, pp. 1–8, 2014, doi: 10.1155/2014/240217.

[19] Y. Zhou *et al.*, 'A novel model based on multi-grained cascade forests with wavelet denoising for indoor occupancy estimation', *Building and Environment*, vol. 167, p. 106461, Jan. 2020, doi: 10.1016/j.buildenv.2019.106461.

[20] F. Cicirelli, A. Guerrieri, C. Mastroianni, G. Spezzano, and A. Vinci, Eds., *The Internet of Things for Smart Urban Ecosystems*. in Internet of Things. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-319-96550-5.

[21] A. Guerrieri, V. Loscri, A. Rovella, and G. Fortino, Eds., *Management of Cyber Physical Objects in the Future Internet of Things: Methods, Architectures and Applications*. in Internet of Things. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-26869-9.

[22] I. Khan, A. Guerrieri, G. Spezzano, and A. Vinci, 'Occupancy Prediction in Buildings: An approach leveraging LSTM and Federated Learning', in *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, Falerna, Italy: IEEE, Sep. 2022, pp. 1–7. doi: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927838.

[23] S. Golestan, S. Kazemian, and O. Ardakanian, 'Data-Driven Models for Building Occupancy Estimation', in *Proceedings of the Ninth International Conference on Future Energy Systems*, in e-Energy '18. New York, NY, USA: Association for Computing Machinery, Jun. 2018, pp. 277–281. doi: 10.1145/3208903.3208940.

[24] P. Anand, C. Deb, K. Yan, J. Yang, D. Cheong, and C. Sekhar, 'Occupancy-based energy consumption modelling using machine learning algorithms for institutional buildings', *Energy and Buildings*, vol. 252, p. 111478, Dec. 2021, doi: 10.1016/j.enbuild.2021.111478.

[25] E. Hailemariam, R. Goldstein, R. Attar, and A. Khan, 'Real-Time Occupancy Detection using Decision Trees with Multiple Sensor Types'.

[26] L. M. Candanedo, V. Feldheim, and D. Deramaix, 'A methodology based on Hidden Markov Models for occupancy detection and a case study in a low energy residential building', *Energy and Buildings*, vol. 148, pp. 327–341, Aug. 2017, doi: 10.1016/j.enbuild.2017.05.031.

[27] M. S. Zuraimi, A. Pantazaras, K. A. Chaturvedi, J. J. Yang, K. W. Tham, and S. E. Lee, 'Predicting occupancy counts using physical and statistical Co2-based modeling methodologies', *Building and Environment*, vol. 123, pp. 517–528, Oct. 2017, doi: 10.1016/j.buildenv.2017.07.027.

[28] A. N. Sayed, R. Hamila, Y. Himeur, and F. Bensaali, ‘Employing Information Theoretic Metrics with Data-Driven Occupancy Detection Approaches: A Comparative Analysis’, in *2022 5th International Conference on Signal Processing and Information Security (ICSPIS)*, Dec. 2022, pp. 50–54. doi: 10.1109/ICSPIS57063.2022.10002508.

[29] Y. Peng, A. Rysanek, Z. Nagy, and A. Schlüter, ‘Occupancy learning-based demand-driven cooling control for office spaces’, *Building and Environment*, vol. 122, pp. 145–160, Sep. 2017, doi: 10.1016/j.buildenv.2017.06.010.

[30] ‘Dreyfus: Neural networks: methodology and applications - Google Scholar’. https://scholar.google.com/scholar_lookup?title=Neural%20Networks%3A%20Methodology%20and%20Applications&author=G.%20Dreyfus&publication_year=2005 (accessed Jun. 14, 2023).

[31] D. F. Specht, ‘Probabilistic neural networks’, *Neural Networks*, vol. 3, no. 1, pp. 109–118, Jan. 1990, doi: 10.1016/0893-6080(90)90049-Q.

[32] R. Krzysztofowicz, ‘Bayesian theory of probabilistic forecasting via deterministic hydrologic model’, *Water Resources Research*, vol. 35, no. 9, pp. 2739–2750, 1999, doi: 10.1029/1999WR900099.

[33] L. M. Candanedo and V. Feldheim, ‘Accurate occupancy detection of an office room from light, temperature, humidity and CO 2 measurements using statistical learning models’, *Energy and Buildings*, vol. 112, pp. 28–39, Jan. 2016, doi: 10.1016/j.enbuild.2015.11.071.

[34] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, ‘A systematic approach to occupancy modeling in ambient sensor-rich buildings’, *SIMULATION*, vol. 90, no. 8, pp. 960–977, Aug. 2014, doi: 10.1177/0037549713489918.

[35] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, ‘A Multi-Sensor Based Occupancy Estimation Model for Supporting Demand Driven HVAC Operations’.

[36] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, ‘Occupancy Detection from Electricity Consumption Data’, in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, Roma Italy: ACM, Nov. 2013, pp. 1–8. doi: 10.1145/2528282.2528295.

[37] Y. Jiang, B. Cukic, T. Menzies, and J. Lin, ‘Incremental Development of Fault Prediction Models’, *Int. J. Softw. Eng. Knowl. Eng.*, 2013, doi: 10.1142/S0218194013500447.

[38] M. Vega and M. Todd, ‘A variational Bayesian neural network for structural health monitoring and cost-informed decision-making in miter gates’, *Structural Health Monitoring*, vol. 21, 2020, doi: 10.1177/1475921720904543.

[39] S. Legler and T. Janjić, ‘Combining data assimilation and machine learning to estimate parameters of a convective-scale model’, *Quarterly Journal of the Royal Meteorological Society*, vol. 148, 2021, doi: 10.1002/qj.4235.

[40] V. Mullachery, A. Khera, and A. Husain, ‘Bayesian Neural Networks’. arXiv, Jan. 30, 2018. doi: 10.48550/arXiv.1801.07710.

[41] E. Hüllermeier and W. Waegeman, ‘Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods’, *Mach Learn*, vol. 110, no. 3, pp. 457–506, Mar. 2021, doi: 10.1007/s10994-021-05946-3.

[42] Y. Gal and Z. Ghahramani, ‘Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning’, in *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, Jun. 2016, pp. 1050–1059. Accessed: Jun. 15, 2023. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[43] J. Huang, Y. Pang, Y. Liu, and H. Yan, ‘Posterior Regularized Bayesian Neural Network incorporating soft and hard knowledge constraints’, *Knowledge-Based Systems*, vol. 259, p. 110043, Jan. 2023, doi: 10.1016/j.knosys.2022.110043.

[44] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, ‘Variational Inference: A Review for Statisticians’, *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, Apr. 2017, doi: 10.1080/01621459.2017.1285773.

[45] S. Kullback and R. A. Leibler, ‘On Information and Sufficiency’, *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951, doi: 10.1214/aoms/1177729694.

[46] D. P. Kingma and M. Welling, ‘Auto-Encoding Variational Bayes’. arXiv, Dec. 10, 2022. doi: 10.48550/arXiv.1312.6114.

[47] D. Saelens, W. Parys, and R. Baetens, ‘Energy and comfort performance of thermally activated building systems including occupant behavior’, *Fuel and Energy Abstracts*, 2011, doi: 10.1016/J.BUILDENV.2010.10.012.

[48] J. Zhu, N. Chen, and E. P. Xing, ‘Bayesian inference with posterior regularization and applications to infinite latent SVMs’, *Journal of Machine Learning Research*, vol. 15, pp. 1799–1847, 2014.

[49] K. Tsikaloudaki, K. Laskos, and D. Bikas, ‘On the Establishment of Climatic Zones in Europe with Regard to the Energy Performance of Buildings’, *Energies*, vol. 5, no. 1, pp. 32–44, Dec. 2011, doi: 10.3390/en5010032.

[50] R. Alaiz-Rodríguez and A. C. Parnell, ‘An information theoretic approach to quantify the stability of feature selection and ranking algorithms’, *Knowledge-Based Systems*, vol. 195, p. 105745, May 2020, doi: 10.1016/j.knosys.2020.105745.

[51] A. V. Konstantinov and L. V. Utkin, ‘Interpretable machine learning with an ensemble of gradient boosting machines’, *Knowledge-Based Systems*, vol. 222, p. 106993, Jun. 2021, doi: 10.1016/j.knosys.2021.106993.

[52] G. Wu, C. Li, L. Yin, J. Wang, and X. Zheng, ‘Compared between support vector machine (SVM) and deep belief network (DBN) for multi-classification of Raman spectroscopy for cervical diseases’, *Photodiagnosis and Photodynamic Therapy*, vol. 42, p. 103340, Jun. 2023, doi: 10.1016/j.pdpdt.2023.103340.

[53] H. Zhang and L. Jiang, ‘Fine tuning attribute weighted naive Bayes’, *Neurocomputing*, vol. 488, pp. 402–411, Jun. 2022, doi: 10.1016/j.neucom.2022.03.020.

[54] J. Frost, ‘Root Mean Square Error (RMSE)’, *Statistics By Jim*, May 06, 2023. https://statisticsbyjim.com/regression/root-mean-square-error-rmse/ (accessed Sep. 11, 2023).

[55] A. Bajaj, ‘Performance Metrics in Machine Learning [Complete Guide]’, *neptune.ai*, Jul. 21, 2022. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide (accessed Sep. 11, 2023).

[56] M. Weber, C. Doblander, and P. Mandl, ‘Towards the Detection of Building Occupancy with Synthetic Environmental Data’. arXiv, Oct. 08, 2020. Accessed: May 22, 2023. [Online]. Available: http://arxiv.org/abs/2010.04209