An assessment of stingless beehive climate impact using multivariate recurrent neural networks

Noor Hafizah Khairul Anuar1,2, Mohd Amri Md Yunus3, Muhammad Ariff Baharudin1, Sallehuddin Ibrahim1, Shafishuhaza Sahlan1, Mahdi Faramarzi4
1School of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, Malaysia
2Center for Electrical Engineering Studies, Universiti Teknologi MARA Johor Branch, Masai, Malaysia
3Frontier Materials Research Alliance, Control and Mechatronics Engineering, School of Electrical Engineering Universiti Teknologi Malaysia, Skudai, Malaysia
4Data Science and Analytic Groups, Digital Department, Sembcorp Industries Ltd, Singapore

ABSTRACT

A healthy bee colony depends on various elements, including a stable habitat, a sufficient source of food, and favorable weather. This paper aims to assess the stingless beehive climate data and examine the precise short-term forecast model for hive weight output. The dataset was extracted from a single hive, for approximately 36-hours, at every seven seconds time stamp. The result represents the correlation analysis between all variables. The evaluation of root-mean-square error (RMSE), as well as the RMSE performance from various types of topologies, are tested on four different forecasting window sizes. The proposed forecast model considers seven of input vectors such as hive weight, an inside temperature, inside humidity, outside temperature, outside humidity, the dewpoint, and bee count. The various network architecture examined for minimal RMSE are long short-term memory (LSTM) and gated recurrent units (GRU). The LSTM1X50 topology was found to be the best fit while analyzing several forecasting window sizes for the beehive weight forecast. The results obtained indicate a significant unusual symptom occurring in the stingless bee colonies, which allow beekeepers to make decisions with the main objective of improving the colony’s health and propagation.

Keywords: Climatic impact, Gated recurrent units, Long short-term memory, Short-term forecast, Stingless bee

1. INTRODUCTION

Stingless bees (Meliponies) are eusocial insects like honeybees (Apis), which produce honey, propolis, and beeswax [1]. Over 600 stingless bee species with 56 different names of genera live in tropical and subtropical areas of the world [2]. Meliponiculture is stingless beekeeping with supportable activities suitable for honey production and keeping excellent pollinators of the environmental and agricultural fields [3]. An example of issues studied in the area of beekeeping includes the foraging activity pattern [1], [4], the pollinations factors and effects [5], the environmental influences [6], beekeeping and rearing techniques [7]–[9], and the quality of honey and propolis production [10]. The increase in worldwide temperature may uncomfortably pressure the colonies while residing in their habitat. To prevent overheating, bee colonies attempt to regulate the internal hive temperature between 33 °C and 36 °C [11]. High temperature may lead
to situations includes colony death [11], swarming event [12], [13], low quality of honey [14], [15], and colonies disappearance [16]. Temperature and humidity are the two crucial determined parameters to analyze the beehive, as well as hive weight and foraging activity data to resemble colonies’ health level [17]–[22], swarming events [15], [23]–[25], and yields productions status [26], [27].

The neural network (NN) approach uses computational methods to learn the information directly from massive data without relying on a specific equation. For example, a study [28] used a NN on environmental data, hive temperature, and weight to forecast the beehive health status in beekeeping. The study [28] indicates an imminent collapsing state for beekeepers using classification algorithms such as k-nearest neighbors (k-NN), random forest (RF), and NN. The approach, like support vector machine (SVM) and machine learning methods, was used in the detection of the queen bee state condition as reported in [28], [29]. In addition, Arruda et al. [30] applied an algorithm (RF, multilayer perceptron (MP), and support vector machine (SVM)) to the foraging pattern data from radio frequency identification (RFID) system to classify the stingless bee species. In [31], a fuzzy logic approach is used on temperature data to identify the honeybee hive state, either health, death, or weak.

Recurrent neural networks (RNN) add feedback within the neurons that allows the network to handle the variable-length sequences. RNN presents the ability to store internal memory to manage dynamic temporal behavior naturally. In [32], image processing through the fast-region-based convolutional neural networks (fast-RCNN) approach amounts to the pollen carried by the honeybee forager into the hive. As an improvement from the RNN concept, Hochreiter and Schmidhuber [33] introduced long short term memory (LSTM) cells that benefit from long term dependencies. Kachole et al. [34] used image processing to count the bee traffic data at the hive entrance and forecast the activity and health using the LSTM method. Meanwhile, the gated recurrent units (GRU) and LSTM approach are used in [35] to predict the bee activity levels through activity data and environmental parameters. This research paper focuses on a method/approach for processing and analyzing stingless beehive climate data impact to forecast the beehive yields. The deep learning RNN approach was compared using LSTM and GRU to establish the best topology. Several forecasting window sizes have also been tested to improve model performance for future development. The forecast output assessed in this paper is the hive weight concerning the other environmental parameters (temperature, humidity, and sum of incoming and outgoing bees at the hive funnel). The output helps beekeepers estimate the yields produced and colony growth. The hive weight forecast output helps to determine the best time range for harvesting and the process of colony propagation (to separate the colonies if an overcrowded situation happens). On the other hand, the study may determine the correlations between observed parameters that most influence the colony’s health. The suggested model also predicts unusual symptoms, which assists beekeepers in determining appropriate actions for their hives to improve the colony’s propagation process.

2. RESEARCH METHOD

2.1. Data collection

Data collection using wireless stingless bee management and monitoring system for ~36 h from Dec 7, 2018, to Dec 9, 2018. In this paper the data stamps are set up at intervals of seven seconds. The monitoring system consists of a weight sensor (load cell), temperature and humidity sensors (DHT22), and bee counter controlled using a NodeMCU microcontroller. The bee counter system used two pairs of infrared sensors for counting the incoming and outgoing bees through the hive entrance funnel. The data from a microcontroller were pushed to Firebase’s real-time database and can be accessed through web or android applications. The technical detail of the monitoring system is as discussed in previous works reported in [36]. The data were continuously collected in real-time and saved to google firebase's real-time database. The extracted dataset is from a single hive of Heterotrigona itama. Figure 1 shows the overall architecture of the monitoring system used for this study.

2.2. The input variables

The sensors system observed the hive weight in kg, inside and outside hive temperature and humidity in °C and % respectively, and bee traffic at hive entrance (total number of bees going in and out from funnel). The bee counter consists of 2 pairs of infrareds (IRs) emitters and photoresistors. The bee counter system recognizes a departing bee when the first pair of IR is detected, followed by the second IR pair. Then, the value “1” is added to the current total bee count in the programming by the microprocessor. Meanwhile, an arriving bee is detected when the second IR pair is detected, followed by the first IR pair, and then a value of “-1” is added to the current total count.

Dewpoint temperature is a calculated variable accesses in this project. The dewpoint is the temperature at which the air is the most saturated. The inside hive relative humidity (RHin) and the inside hive temperature (Tin) need to be determined to calculate the dew point. Condensation occurs when the inside
hive temperature is below the dewpoint ($T_{dp}$). Depending on the surface types, condensation is estimated to occur at $|T_{dp} - T_{in}| < 1.5 \degree C$ [37] for the greenhouse area and at $|T_{dp} - T_{in}| < 3 \degree C$ [38] for the standard building area. The dewpoint temperature was calculated using the inside hive temperature and humidity data [39]; as represented in (1).

$$T_{dew} = 237.3 \left( \frac{[RH(\%)] + 17.27 \cdot T}{12.56 + 237.3 \cdot T} \right)$$  \hspace{1cm} (1)$$

where $RH$ is inside hive relative humidity in $\%$, $T_{in}$ is inside temperature in $\degree C$, and $T_{dew}$ is dewpoint temperature in $\degree C$.

3.3. Data preprocessing

In this project, the scientific computing and data analysis process was performed using Numpy and Pandas libraries in Python software. The main subprocess involved data cleaning, data normalization, supervised learning data frame, segmentation, and network design using LSTM/GRU units for training and testing. Firstly, the data cleaning process involves cleaning, indexing (drop the date and time column), resizing, and searching for the missing data. Next, data normalization was performed. This process is essential in stabilizing the supervised learning process by ensuring the input and output ranges are of the same scale. The process rearranges the entire variable data range between zero and one by using the MinMaxScaler features of the scikit-learn library. The normalization scale was interpreted by (2).

$$x_i^*(k) = \frac{\max x_i(k) - x_i(k)}{\max x_i(k) - \min x_i(k)}$$  \hspace{1cm} (2)$$

where $i$ is the number of variables, $k$ is the number of data (index), and $x_i^*(k)$ is the sequence after the normalization process.

Supervised learning is a key component of machine learning and data mining [40]. After the normalization process, the generation of a supervised learning data frame took place. One-dimensional vectors were alternately applied as an output. This project evaluated variables including hive weight and bee count as outputs, run one by one, rather than simultaneously. The chosen output variable is to compare the forecast performance and relevance between both variables. In contrast, the rest of the variables or vectors were set as the influenced input variables (temperature, humidity, dewpoint). Next, the data frame was segmented into train and test sections. In this experiment, the train and test sections were segmented approximately by 75% and 25%, respectively. The data samples to be processed in the train and test sections were 2,500 and 464, respectively. The next steps are to manipulate the number of neurons in the hidden and output layers and a loss function, and an optimizer were determined. In this work, the model is evaluated using the root mean square error (RMSE), which can be obtained from (3).
An assessment of stingless beehive climate impact using multivariate ... (Noor Hafizah Khairul Anuar)

\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \]  

(4)

where \( y_i \) is the actual output, \( \hat{y}_i \) is the output forecast, and \( n \) is the size of the test dataset. This study forecasts a single variable, influenced by the seven input variables. Finally, the training and test losses were tracked by setting the validation data argument in the `fit()` function. The training, test loss, and correct and predicted outputs were plotted at the end of the process.

3.4. The model

The LSTM and GRU [38] networks are superior RNNs that have the competencies for learning long-term dependencies [37]. The Keras model implements multivariate time series forecasting using LSTM/GRU units. Keras is a high-level neural network API written in the python platform and capable of running with several backend packages such as Tensorflow, cognitive toolkit (CNTK), and Theano. In the process of finding the best RNN topology for hive weight and bee activity forecast, six different topologies considered a different number of the hidden layer and neurons. The designed network utilizes the Keras sequential model. The constructed topologies included one hidden layer with thirty-two recurrent units in each layer (1×32), one hidden layer with fifty recurrent units (1×50), two hidden layers with thirty-two recurrent units in each layer (2×32), and two hidden layers with fifty recurrent units per layer (2×50). Hence, the six topologies models are: LSTM 1×32, LSTM 1×50, LSTM 2×32, LSTM 2×50, GRU 2×32, and GRU 2×50.

The input shape is a single timestep with seven features, and the batch size is set as 32. The loss function and optimizer used are the mean absolute error and the efficient Adam version of stochastic gradient descent, respectively [40]. The output layer consists of a dense, fully connected layer to compile the forecast output. The example of potential architectures for a single hidden layer is shown in Figure 2. In this illustration, seven input variables (at \( t=0 \)) and one output variable (at \( t=t+1 \)) are forecast. The evaluated forecast output will be hive weight or bee activity that happens independently, and the unit of the hidden layer for this investigation is LSTM and GRU.

3.5. The multiple input vector window size

The previous evaluation sought to identify the best RNN topology for the data. The best topology from the previous research was employed in the second inquiry, which was aimed at determining the optimal input vector window size. There are four types of window size was chosen to be investigated which are \( \{w1=t_0-1\} \), \( \{w2=t_0-2\} \), \( \{w3=t_0-9\} \), and \( \{w4=t_0-14\} \). Referring to Table 1, \( t_0 \) denotes the current time, \( t_0-1 \) denotes one minute before the event to be predicted, \( t_0-2 \) is for two minutes before the event forecast, \( t_0-9 \) is for nine minutes before the event to forecast, and \( t_0-14 \) is for fourteen minutes before the event to forecast.
Table 1. Evaluated input vectors structure of hive weight forecast

<table>
<thead>
<tr>
<th>Reference</th>
<th>Input vectors</th>
<th>Forecast output</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td>t_r1</td>
<td>t_o</td>
</tr>
<tr>
<td>w2</td>
<td>t_r1,t_r2</td>
<td>t_o</td>
</tr>
<tr>
<td>w3</td>
<td>t_r1,t_r2,t_r9</td>
<td>t_o</td>
</tr>
<tr>
<td>w4</td>
<td>t_r1,t_r2,t_r9,t_r14</td>
<td>t_o</td>
</tr>
</tbody>
</table>

In this study, the optimal architecture discovered in the previous stage was used to analyze alternative widths for the input window, resulting in varying quantities of preceding data being used to anticipate the next level. It necessitates the RNN's capacity to retain useful information over time. The evaluations were conducted one minute, two minutes, nine minutes, and fourteen minutes before predicting the event.

3. RESULTS AND DISCUSSION

3.1. Statistical exploration

The initial analysis using basic statistical techniques was performed to familiarize with the datasets. Then, the mean, standard deviation, maximum and minimum value, and interquartile range were compute. Table 2 represents the total, mean, standard deviation, minimum and maximum count of the dataset. The total data were 2,965, consisting of the hive weight (Kg), inside and outside hive temperature (°C), inside and outside hive humidity (%), the dewpoint (°C), and the bee count. In addition, the data show a huge variable range where the minimum is 16, and the maximum is 21,282 for the bee count variable.

Table 2. The statistical description of the datasets

<table>
<thead>
<tr>
<th>count</th>
<th>weight (Kg)</th>
<th>temp in (°C)</th>
<th>temp out (°C)</th>
<th>RH out (%)</th>
<th>RH in (%)</th>
<th>DP (°C)</th>
<th>Bee count (counts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>std</td>
<td>0.526321</td>
<td>2.273922</td>
<td>1.574741</td>
<td>5.699065</td>
<td>9.67E-13</td>
<td>2.273628</td>
<td>4465.29884</td>
</tr>
<tr>
<td>min</td>
<td>2.6802</td>
<td>24.3</td>
<td>25.4</td>
<td>68.2</td>
<td>9.99E+01</td>
<td>24.283294</td>
<td>16.33</td>
</tr>
<tr>
<td>25%</td>
<td>3.4188</td>
<td>25</td>
<td>26.2</td>
<td>95.4</td>
<td>9.99E+01</td>
<td>24.983204</td>
<td>34.82</td>
</tr>
<tr>
<td>50%</td>
<td>3.9908</td>
<td>25.7</td>
<td>27.1</td>
<td>98.5</td>
<td>9.99E+01</td>
<td>25.683115</td>
<td>820.58</td>
</tr>
<tr>
<td>75%</td>
<td>4.1948</td>
<td>27.6</td>
<td>28.9</td>
<td>99.6</td>
<td>9.99E+01</td>
<td>27.58287</td>
<td>1773.24</td>
</tr>
<tr>
<td>max</td>
<td>4.5862</td>
<td>32.8</td>
<td>30.7</td>
<td>99.9</td>
<td>9.99E+01</td>
<td>32.782191</td>
<td>21282.6</td>
</tr>
</tbody>
</table>

3.2. Correlation analysis

A correlation analysis is performed on each observed variable to determine the relationship between them in the first evaluation. Based on the correlation analysis result, the beehive weight as an output variable shows a strong correlation with the inside temperature, outside temperature, and dewpoint variables (approximately 70%). The study also indicated that, if bee count is considered as the output variable, other examined variables have a weak relationship (less than 20%). Even though the bee count variable was also being run as an output on the second investigation, to compare between forecast output from hive weight and bee count. Within all the access variables, the inside humidity has the weakest correlation with other vectors. Figure 3 represents the features correlation heatmap for all observed variables. The blue region on the heatmap denotes a positive correlation, implying that the cross-correlation variables are proportional to each other. The red region, on the other hand, denotes a negative correlation, which signifies that all cross-correlation variables are negatively proportional to one another. Brighter areas are considered natural correlations, whereas darken areas are considered a strong correlation. The RNN model evaluations in this study have considered all observed variables as input vectors.

3.3. RNN topologies

The second evaluation aims to investigate the best topology of RNN models for forecasting the selected variables (hive weight and bee count) based on all observed variables as input vectors. Initially, the first experiment forecast the hive weight variable, the designed network was trained and evaluated with 50 times epoch, and a batch size of 32. In the second experiment (bee count variable), the epoch evaluation is set to 200, and the same batch size of 32 for network design. The bee count data consists of a massive range for the maximum-minimum value (from 16 to 20,000) and require more epoch evaluation to stabilize the output fluctuations.

The most significant model found in the investigation of the hive weight forecast is LSTM 1X50 with the lowest RMSE (0.012). Bee count variable represents the high value of RMSE output (≈ 74 – 160).
In the first evaluation, bee count shows a weak correlation with other access vectors as well as the input data range consists of a massive maximum and minimum peak. Figure 4 presents the error RMSE for all experimented network topologies for hive weight forecast. Meanwhile, Figure 5 shows the RMSE output graph of train vs. test for a hive weight forecast using the LSTM 1x50 topology. The forecast output fluctuation is stable approximately at epoch 50 times evaluations.

3.4. Investigations multiple window size on beehive weight forecast

The final evaluation seeks to establish the most appropriate size for the input window. The output of beehive weight forecast based on all input attributes using RNN topology of LSTM1X50 was used since this topology is the best found in the previous evaluation. Table 3 shows the first 5 rows of the supervised learning data frame for the window size of 1 minute before the event forecast (\(w1 = t - 1\)).
From Table 3, the data were grouped into an 8-column table, with the first and seventh columns representing the RNN input vectors and the eighth column representing the anticipated output. The \( \text{var1} \) represents the value of hive weight, while the \( \text{var2} \) until \( \text{var7} \) represents the value of inside temperature, outside temperature, outside humidity, inside humidity, dewpoint, and bee count consequently. The window size of \( w1 \) (1 minute before the event forecast) resulted in the best RMSE of 0.012. Figure 6 represents the RMSE value obtained after evaluation for 50 times epoch for each different forecast window size. The larger the window, the more time-series information can be considered. It may, however, reduce the system's sensitivity, resulting in overly smooth forecasts. In our opinion, it is critical to find the best window size that corresponds to the fractal data dimension.

### Table 3. The example of the first 5 rows of an input-output data frame for the beehive weight variable forecast

<table>
<thead>
<tr>
<th>( \text{var1}(t-1) )</th>
<th>( \text{var2}(t-1) )</th>
<th>( \text{var3}(t-1) )</th>
<th>( \text{var4}(t-1) )</th>
<th>( \text{var5}(t-1) )</th>
<th>( \text{var6}(t-1) )</th>
<th>( \text{var7}(t-1) )</th>
<th>( \text{var1}(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1   0.340923</td>
<td>0.564706</td>
<td>0.943396</td>
<td>0.747634</td>
<td>0</td>
<td>0.564706</td>
<td>0.848004</td>
<td>0.339979</td>
</tr>
<tr>
<td>2   0.339979</td>
<td>0.564706</td>
<td>0.943396</td>
<td>0.741325</td>
<td>0</td>
<td>0.564706</td>
<td>0.838568</td>
<td>0.341658</td>
</tr>
<tr>
<td>3   0.341658</td>
<td>0.564706</td>
<td>0.943396</td>
<td>0.741325</td>
<td>0</td>
<td>0.564706</td>
<td>0.829131</td>
<td>0.342497</td>
</tr>
<tr>
<td>4   0.342497</td>
<td>0.564706</td>
<td>0.943396</td>
<td>0.73817</td>
<td>0</td>
<td>0.564706</td>
<td>0.819689</td>
<td>0.342078</td>
</tr>
<tr>
<td>5   0.342078</td>
<td>0.564706</td>
<td>0.943396</td>
<td>0.741325</td>
<td>0</td>
<td>0.564706</td>
<td>0.810243</td>
<td>0.341343</td>
</tr>
</tbody>
</table>

![Figure 6. The RMSE of different window sizes for hive weight forecast](image)

**4. CONCLUSION**

This study proposed a forecast model for stingless beehive weight output. The observed variables (features) are the internal hive temperature, internal hive humidity, outside hive temperature, outside hive humidity, dewpoint temperature, and bee count. The investigated models were run by using the Keras library in Python software. All features were tested for their significance using the correlation matrix method.

In future works, it is recommended to vary the dataset season for example data in sunny, rainy, windy, fruity, or flowering. Hence, different locations and landscapes of the dataset such as urban, rural, highland, and lowland could also be considered. It is also recommended to investigate other potential attributes such as light intensity, solar irradiance, gas content, and air pressure.

**ACKNOWLEDGEMENTS**

For financial support, the authors are thankful for a Research University Grant from Vote 20H22 and 21H06 via the Research Management Centre Universiti Teknologi Malaysia.

**REFERENCES**


An assessment of stingless bee hive climate impact using multivariate … (Noor Hafizah Khairul Anuar)


BIOGRAPHIES OF AUTHORS

Noor Hafizah Khairul Anuar received the B.Eng in Electrical Telecommunication from Universiti Teknologi Malaysia (UTM) in 2008 and M.Sc in Electrical Telecommunication Engineering and Information Technology from Universiti Teknologi MARA (UiTM) in 2012. She is currently pursuing her Ph.D. in Electrical Engineering at Universiti Teknologi Malaysia (UTM) in sensor development, instrumentation, and machine learning. She is a lecturer at the College of Engineering, Universiti Teknologi MARA, Johor. Email: noorhafizah2575@ultm.edu.my.

Mohd Amri Md Yunus received the B.Eng. degree and the M.Eng degree in electrical engineering from Universiti Teknologi Malaysia, Skudai, Malaysia, in 2002 and 2005, respectively. In 2011, he received a Ph.D. in sensing technology from Massey University, Palmerston North, New Zealand. He is currently a Senior Lecturer in the Control and Mechatronics Engineering Division. His research interests include process tomography, planar electromagnetic sensors, and sensing technology. Email: amri@fke.utm.my.

Muhammad Ariff Baharudin received his B.Eng. (Mechatronics) and M.Eng. (Electronic & Communications) from Universiti Teknologi Malaysia (UTM) in 2008 and 2009, respectively. He joined the Faculty of Electrical Engineering as a Tutor in 2008 after completing his B.Eng. degree. He then completed his M.Eng. and Ph.D. in Functional Control System (Mobile Computing & Communications) from Shibaura Institute of Technology, Tokyo, Japan. Currently, he is a Senior Lecturer in the Communication Engineering Division, School of Electrical Engineering, UTM. Dr. Ariff is actively involved in research as a principal investigator and consultancy projects focusing on Communication systems, the Internet of Things (IoT) system, and WiFi network planning and deployment. His research interest includes Software Defined Networks, Fog Computing, and IoT. Email: mariff@utm.my.

Sallehuddin Ibrahim is currently an Associate Professor at the Control and Instrumentation Engineering Department, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Malaysia. He received his Ph.D. in process tomography from Sheffield Hallam University, UK, in 2000. He has also presented many papers at various conferences. His current research interest is in the field of instrumentation, sensors, and tomography. Email: sallehudding@utm.my.
Shafishuha Maz Zurina received a Master of Engineering (MEng) in Control Systems from the University of Sheffield between 1998 and 2002. She earned a Doctor of Philosophy (Ph.D.) in Control Systems from the University of Western Australia (UWA) between 2005 and 2010. Her current areas of specialization are Process Control System Modeling and Optimization, Supervisory Control and Optimization, and Energy Management. Email: shafis@utm.my.

Mahdi Faramarzi received his Ph.D. in Electrical and Electronics Engineering from Universiti Teknologi Malaysia, in 2016. He is currently employed at Sembcorp Industries Ltd in Singapore as a Senior Data Scientist. He is a Data Scientist with over 10 years of expertise in Artificial Neural Networks, Classical Machine Learning, Deep Learning, Image Processing, and Metaheuristic Optimization. Email: mahdi.Faramarzi@sembcorp.com.