Prediction of Wind Turbine Noise using SPSS Modeler

Nattapat Charoentangprasert* Netnapid Tantamsapya[1]** Chatpet Yossapol***

(Received: *October 25, 2022*; Revised: *March 22, 2023*; Accepted: *March 23, 2023*)

ABSTRACT

IBM SPSS Modeler was used to develop a noise prediction model for the wind power plant located in Nakhon Ratchasima province, Thailand. Four individual models (CHAID, CART, Linear, and Neural network) and their ensemble were developed and compared. The model's inputs are distance, time, wind speed, wind direction, temperature, humidity, and pressure. The output is the equivalent sound pressure level. From the field measurement, the average sound level (43.0-47.8 dB(A)) was higher for the measurement point closer to the wind turbine. The measured sound at various times of the day shows higher sound levels in the morning and evening, indicating the effect of human activity. The most suitable technique was the Ensemble model, where the cross-validation for training and testing provides RMSE (10.08%) and MAE (5.89%).

Keywords: Noise prediction model, SPSS, Wind turbine noise

[1] *Corresponding author: netnapid@sut.ac.th*

**Graduate student, Department of Environmental Engineering, Institute of Engineering, Suranaree University of Technology, Thailand*

***Associate Professor, Department of Environmental Engineering, Institute of Engineering, Suranaree University of Technology, Thailand*

****Lecturer, Department of Environmental Engineering, Institute of Engineering, Suranaree University of Technology, Thailand*

Introduction

It is well known that wind turbine generates a noise perceived as an annoyance by the residents nearby. The noise can be noticeable and annoying when it exceeds the environmental background noise [1]. Studies on the noise disturbance generated by the wind turbine confirm the impact on sleeping problems that potentially affect human health, such as dizziness, anxiety, and depression [2]. In many countries, a noise impact assessment and mitigation measures are required before obtaining a wind farm permit. A wind turbine and noise emission model are essential for environmental planning, especially for wind farms near noise-sensitive receivers [1]. Wind turbine noise prediction models are classified into three main types. The first type estimates the overall emitted sound power level as a function of turbine parameters such as the rotor diameter, rated power, and wind speed. The second type considers different noise generation mechanisms, such as low-frequency noise, inflow turbulence noise, and airfoil self-noise. The third type relates to noise generation mechanisms, source directivity, and surrounding atmospheric conditions. The third type can be used to understand the interaction between wind turbine operation, noise generation, and surroundings [3]. Data mining involves using data analysis tools to find patterns and relationships in large data sets to build a model and find hidden associations and features. It incorporates analysis and prediction using statistical models, machine learning, and mathematical algorithms, such as neural networks or decision trees. Many researchers used a data mining approach for wind farm noise to develop a noise model [4]. Several factors, including turbine parameters, wind speeds, background noise, and climate, influence the wind turbine noise that impacts receptors.

The IBM SPSS modeler was utilized in this study to apply several numerical analytic methods, including classification and regression tree (CART), chi-squared automatic interaction detector (CHAID), linear regression (LR), artificial neural networks (ANNs), and then applied models to construct an ensemble model.

Accordingly, the objectives of this work are the following: 1) Use meteorology and noise data to develop a noise prediction model; 2) Compare the four individual models (CHAID, CART, Linear, and Neural network) and their ensemble in developing a noise prediction model; 3) Conduct initial performance assessment statistics; 4) Validate the results by assessing the performance metric. The software used is IBM SPSS Modeler, a multi-functional big data analytic tool. The results are a noise prediction model for the wind turbine farm, where the model's inputs are distance, time, wind speed, wind direction, temperature, humidity, and pressure. The output is the equivalent sound pressure level in decibels (dB) (Leq).

Methodology

A. Study area

The study area was a wind power plant and its vicinity. The wind power plant was located in Nakhon Ratchasima province. The study area was 400 m from the boundary of the wind power plant and covered an area of 3.25 sq. km. The surrounding area was rural, and agriculture was cultivated, including cassava, cane, and corn. The area was flat, with little difference in elevation. The majority of the area was unaffected by terrain features like hills, trees, and buildings that could activate sound propagation. The study area is shown in Figure 1.

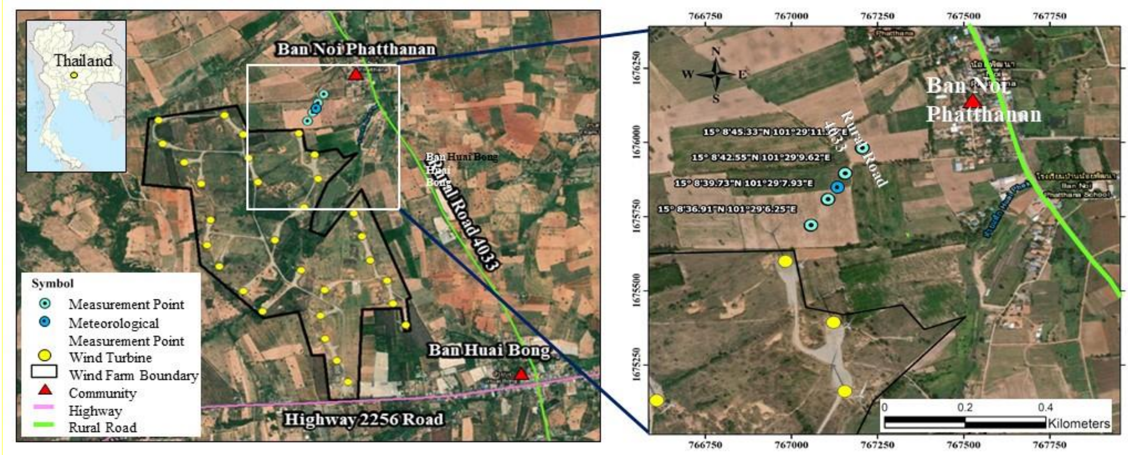


Figure 1 Study area and field measurement locations

B. Method and equipment

The sound level was measured with a calibrated PULSAR Model 44 S/N 1864 Sound Level Meter. The sound level meter was set with a frequency weighting of “A” according to the international standard IEC 61672:2003 [5] to represent human hearing. Meteorology ambient conditions, including wind speed, direction, temperature, humidity, and atmospheric pressure, were measured with the NovaLynx Anemometer. The geographical positions of the measurement points were determined using a Garmin eTrex 10 handheld GPS. ESRI’s ArcGIS 10.1 software was used to create the maps.

C. Data collection and data preparation

Field measurement data was taken in 5-min intervals for three days (From 1.00 pm, 20 February 20th, 2023, to 1.00 pm, February 23rd, 2023), a total of 864 times per point. The measurement was taken at the minimum measurement frequency recommended by USEPA, fifty times per 10 minutes, to ensure sufficient data for modeling [6]. The measurement locations were at the northeast corner of the wind power plant. The sound level measurement was performed at four points with distances of 100 m intervals up to 400 m. Additionally, the meteorology ambient conditions were performed between these measurement points as shown in Figure 1. Measurement data was processed into a consistent and usable form. Data processing included data cleaning, data structuring, data transformation, and data filtering.

D. SPSS Modeler

SPSS Modeler is data mining and analytics software used to build a predictive model. This research applied various algorithms to predict sound levels using field measurement data, including sound levels, wind speed, wind direction, temperature, humidity, and atmospheric pressure. The field measurement data was divided into two datasets, with a ratio of 70% for training and 30% for testing.

The auto-numerical node was used to generate a variety of algorithms in a single modeling run. The node explores every possible model and ranks each candidate model based on the correlation between predicted and observed values for each model. CHAID, CART, Linear, and Neural network models were possible to automatically create, and compare default models of continuous numerical outcomes from the auto-numerical node. Default values were set in the auto-numerical node. Four models were individual constructs that were then applied to construct ensemble models that were proposed for increasing accuracy. The brief descriptions of the prediction models used here are as follows:

CHAID (Chi-squared Automatic Interaction Detection) is a decision tree algorithm that builds a decision tree by recursively splitting the data into subsets based on the most significant differences between the target variable and predictor variables. CHAID is a popular algorithm for categorical target variables. It is used to identify the most important predictors that determine the target variable.

CART (Classification and Regression Trees) is another decision tree algorithm that builds a decision tree by recursively splitting the data into subsets based on the predictor variables that best predict the target variable. CART is used for categorical and continuous target variables. It can also be used for classification and regression tasks.

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. In SPSS Modeler, linear regression models can be used for simple and multiple linear regressions that depend on the number of independent variables. The dependent variable is continuous, and the independent variables can be either continuous or categorical.

Neural networks are a type of machine learning algorithm that is designed to recognize patterns in data. In SPSS Modeler, neural network models can be used for classification and regression tasks. The neural networks are particularly useful when the relationships between the predictor variables and target variables are complex and non-linear. The neural network model in SPSS Modeler allows for the customization of the number of hidden layers and neurons in each layer, as well as the activation function used in the model.

An ensemble model is a machine learning technique that combines multiple individual models to improve the overall performance of the prediction. The idea behind ensemble models is that by combining multiple models, the strengths of each model can be leveraged, and the weaknesses can be mitigated.

To evaluate the prediction accuracy of the individual models and ensemble models, the Predictor Importance Charts were produced to find the relative importance of each predictor in estimating the model. The most appropriate model was selected from 5 types of models by comparing the model’s performance. The results of five models were merged, The performance error of the developed model was evaluated using R-squared (R^2), Root Mean Squared Error (RMSE), and the Mean

Absolute Error (MAE), which expresses the average model-prediction error in the units of the variable of interest [7]. The smallest error model was selected as a prediction model [8]. The expressions of these parameters are given in Eq. (1) – (3).

$$R^2 = 1 - \frac{\sum_{□}(□_{□}-□_{□})^2}{\sum_{□}(□_{□}-\overline{□_{□}})^2} \qquad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{□}\sum_{□=1}^{□}(□_{□} - □_{□})^2} \qquad (2)$$

$$\text{MAE} = \frac{1}{□}\sum_{□=1}^{□}|□_{□} - □_{□}| \qquad (3)$$

Where	$□_{□}$	=	the measured values
	$□_{□}$	=	the predicted values
	$\overline{□}_{□}$	=	the mean values

The gain chart is a visual representation of the performance of a predictive model. Gains are defined as the proportion of hits in each increment relative to the total number of hits in the tree [9]. Finally, the gain charts were plotted to evaluate the performance of the model.

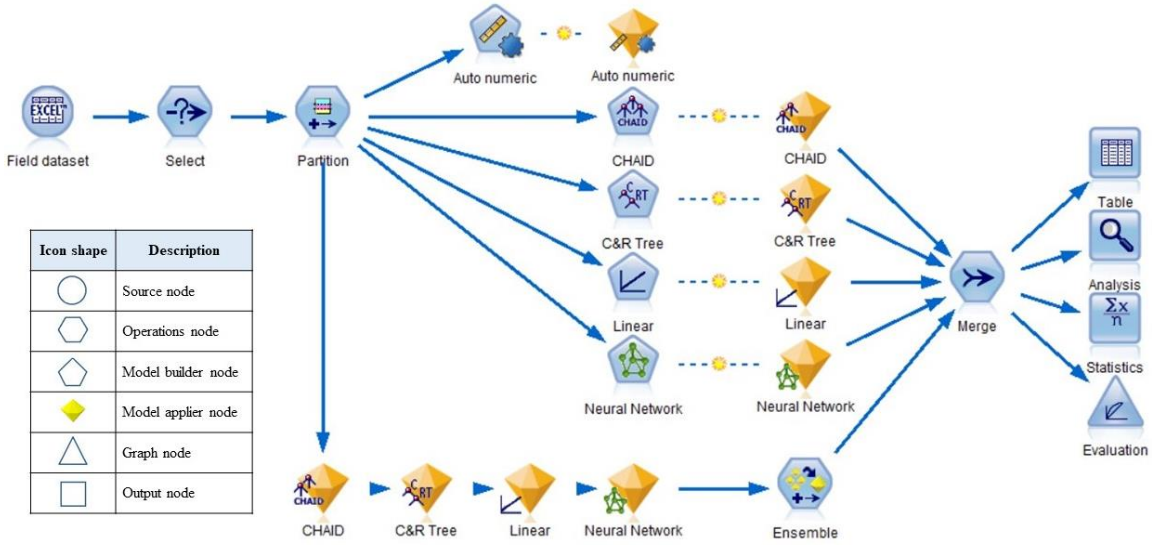


Figure 2 SPSS modeler flow

The modeling steps can be graphically illustrated as SPSS modeler flow, as shown in Figure 2. Nodes in the IBM SPSS Modeler are represented by a specific shape to indicate their function [10]. The source node (circle) imports data into the modeler from a different format. The operations node (hexagon) modifies the data in some way and returns the modified data to the modeler stream. The model builder node (pentagon) generates models from the data in the modeler. The model applier node (gold diamond) defines a container for the generated model that is returned to the modeler canvas. The graph node (triangle) generates a graph or report from the data in the modeler. The output node (rectangle) provides the means to obtain information about data and models. These node shapes work together to facilitate data processing and analysis in the IBM SPSS Modeler.

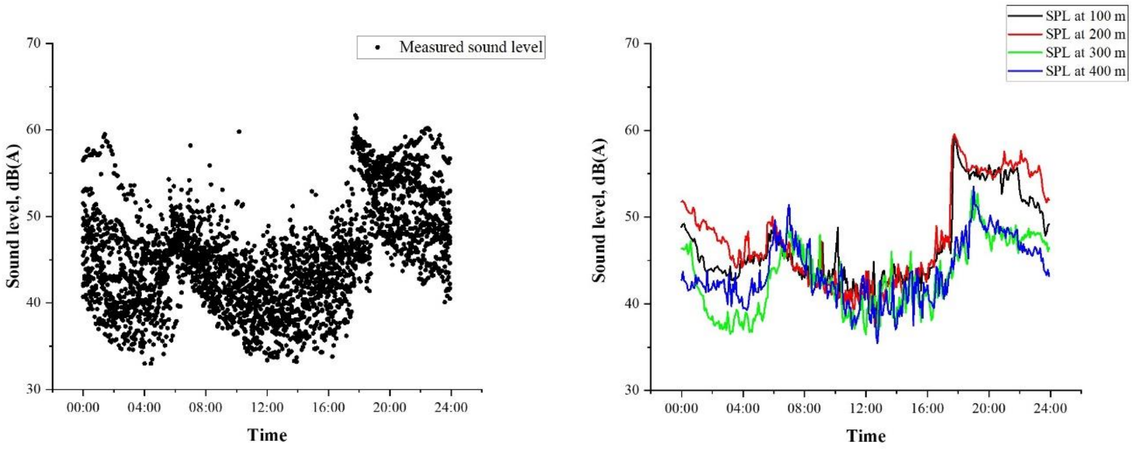
Results

A. Field measurement data

The ranges and averages of the field measurements from four measurement points are shown in Table 1. Comparing meteorological parameters between the field measurement and historical data obtained from the Thai Meteorological Department (TMD) of Nakhon Ratchasima province from 1990-2019 reveals that the measurement data is within the range of the historical data. The average sound level was higher at the measurement point closer to the wind turbine. A plot between sound level and time of the day for all measurement points over 72 hours is shown in Figure 3.

Table 1 Field measurement data

Parameters	Units	Field measurement data				Historical data
		100 m	200 m	300 m	400 m	(1990-2019)
Sound level (Mean±SD)	dB(A)	46.8 ±5.4	47.8 ±6.3	43.0 ±4.9	43.6 ±4.6	-
Wind direction	Degree	62.1±54.8				-
Wind speed	m/s	1.2±1.1				0.9–1.3
Temperature	°C	28.5±2.7				24.4–30.1
Humidity	%	67.7±2.9				62.0–81.0
Pressure	hPa	998.5±0.4				997.7–1,013.8



a) All measured data

b) At different point

Figure 3 Plots between sound level and time

From Figure 3, the difference between the sound at various times of the day can be seen. The higher sound level around morning and evening indicated the effect of human activity from the road and village nearby. The U.S. The Environmental Protection Agency (EPA) defines daytime sound levels as those that occur between the hours of 7.00 am and 10.00 pm, and nighttime sound levels as those that occur between 10.00 pm and 7.00 am [11]. The high noise levels in the daytime compared to the nighttime are typical for a quiet residential area.

The measured sound level, 33.0–61.7 dB(A), was lower than Thailand's standard, which sets an average level of 70 dB(A) for 24 hours and a maximum level of 115 dB(A)[12]. However, some measurements exceed the WHO's recommended value, 45 dB (A), for the wind turbine noise and the WHO's recommended value for community noise in outdoor living areas, 55 dB LAeq [2]. This means that noise in the study area could potentially be harmful to human health. Hence, the mitigation measures should be implemented to protect residents in the study area.

B. Data Preparation

The field measurement data used for model input was within a wind turbine's cut-in speed condition. The cut-in speed is when the wind turbine blades start to rotate and generate power. The wind turbines at the study site are the G114-2.0 MW model, which has a cut-in wind speed of 2.5 m/s. The remaining dataset (n = 576) was divided into training and testing. A ratio of 70/30 for training and testing datasets was a popular ratio, and it was considered the best ratio for training and validating the models [13]. The number of training data was 399 (69.3%), and testing was 177 (30.7%). The distribution plot of the training and testing datasets with sound levels is shown in Figure 4.

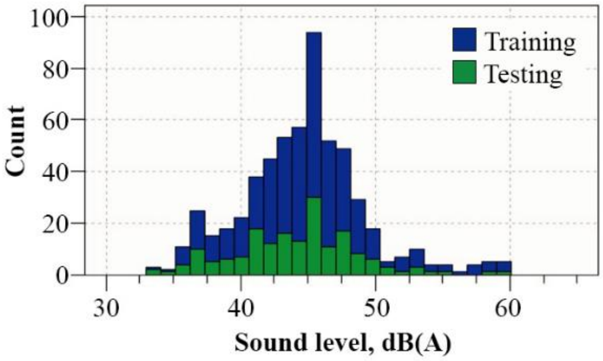


Figure 4 The distribution of the training and testing dataset

C. Modeling

The modified datasets were used to generate models from the auto-numerical node with default values. When an automated modeling node is executed, the node estimates candidate models. The model candidate provided four modeling methods: CHAID, CART, Linear, and Neural network. The ensemble model combines the other models to produce one optimal predictive model. The default ensemble method is voting, the voting operates by counting how many times each potential predicted value is selected, and then choosing the value with the greatest cumulative count.

D. Predictor Importance

The predictor importance chart helps indicate the relative importance of each predictor in estimating the model. In Figure 5, the predictor importance chart of the CHAID, CART, Linear, and Neural network models reveals that distance is the primary predictor, followed by temperature, time, and wind speed.

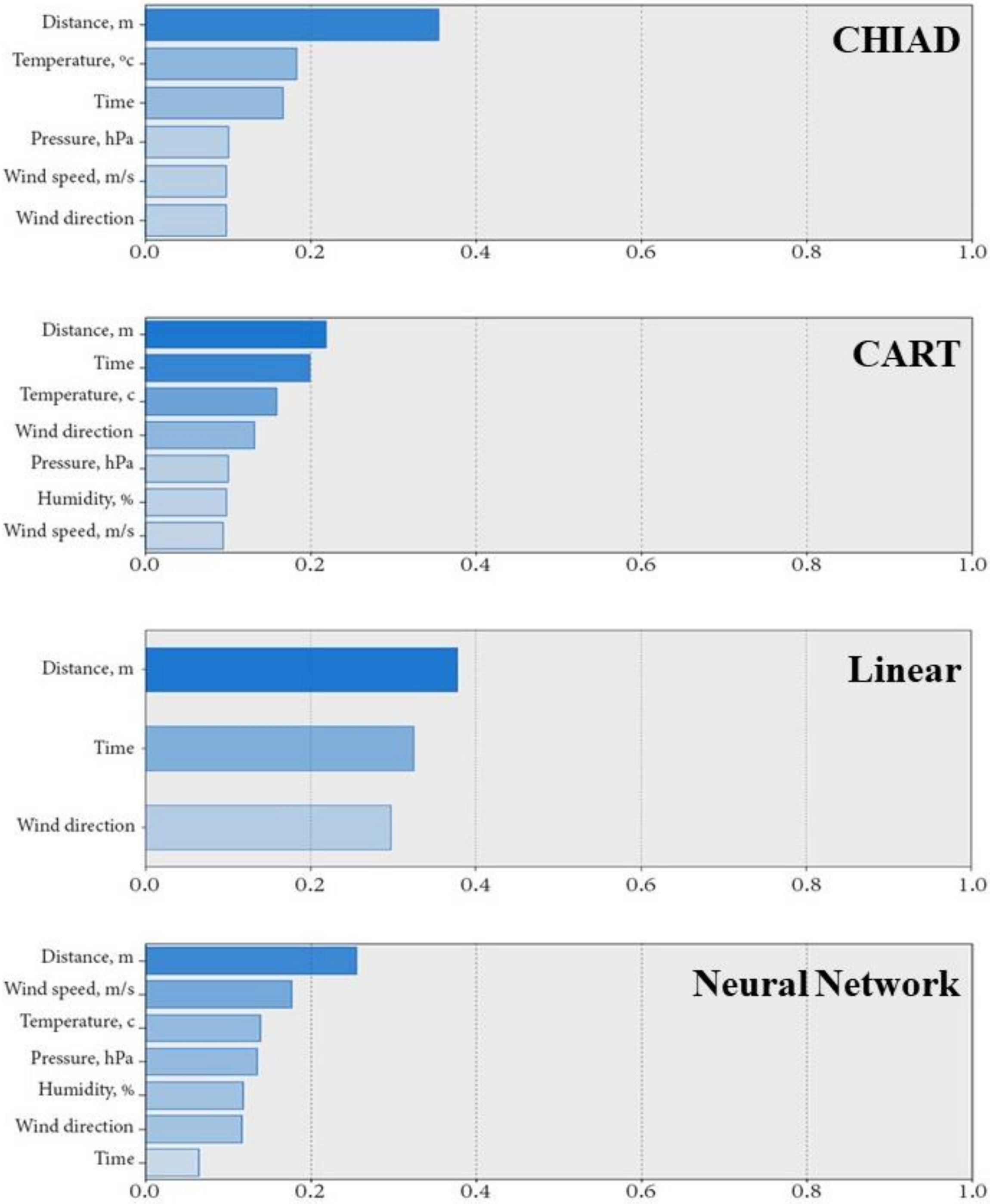


Figure 5 Predictor importance chart

E. Model performance evaluation

Table 2 shows the comparison of the statistical analysis for model evaluation. Considering the R-Squared (R^2), the top 3 best performances were the Ensemble model (0.613), CHAID (0.608), and CART (0.608). Comparing the RMSE and MAE values of the models in Table 2 indicates the Ensemble as the premier model with the lowest values of 2.919 and 2.328, respectively. Therefore, the Ensemble model was selected as a prediction model. The ensemble model was further validated using cross-validation, splitting a dataset into training and testing subsets.

The Ensemble model was further validated using cross-validation by splitting a dataset into training and testing subsets. In this paper, RMSE and MAE are utilized to assess the performance of the forecasting model. As shown in Table 3, The percentage difference between training and testing, RMSE (10.08%) and MAE (5.89%), is low. It indicates that the model is not overfitting [14]. Thus, the proposed model could forecast the sound level with a reasonable level of accuracy. The metrics RMSE and MAE also validate the effectiveness of the model.

Table 2 Comparison of performance metrics of 5 models

Model	R^2	RMSE	MAE
CHAID	0.608	2.871	2.437
CART	0.608	2.871	2.564
Linear	0.276	3.903	3.053
Neural network	0.372	3.848	3.011
Ensemble	0.613	2.919	2.328

Table 3 Ensemble model validation performance metric

Partition	RMSE	MAE
Training	2.818	2.191
Testing	3.134	2.328
% Difference	10.08	5.89

The performances of the models were visually compared using the gain chart plots. The plot presents accumulated gains % to percentile for training and testing datasets. The gain chart in Figure 6 indicates that the models are exemplary because the charts rise steeply toward 100% approximately and then level off.

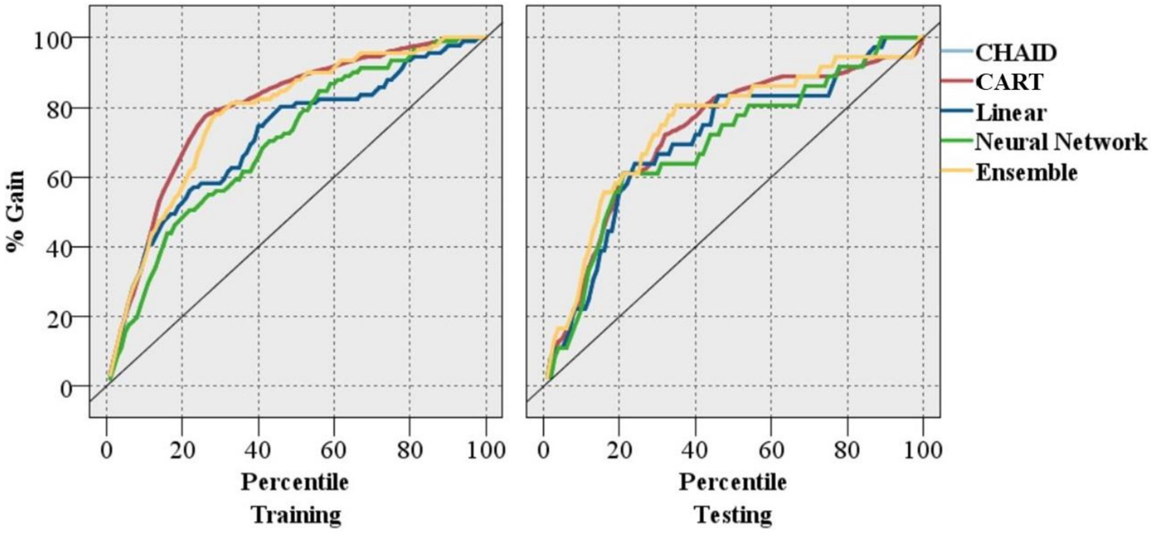


Figure 6 Gain charts

Discussion and Conclusions

The IBM SPSS Modeler was used to develop a noise prediction model, where the model's inputs are distance, time, wind speed, wind direction, temperature, humidity, and pressure. The outputs are the equivalent sound pressure level in A-weighting-network decibels (dBA) (L_{eq}). The noise data from field measurements show the effect of the time of day and human activity from the road and village nearby. To minimize the noise pollution from traffic, mitigation measures are needed, including proper traffic management and strict enforcement of noise rules and regulations [15]. In addition, wind turbine noise mitigation techniques, e.g., installing noise barriers between wind turbines and nearby residences, operational controls that adjust blade rotation speed to wind conditions, and regular maintenance and monitoring of wind turbines can also help reduce noise levels [16].

Five data mining models were applied and compared. Field measurement data at four points measured at 5-minute intervals for 72 hours was used for the models, and it was proved that the data is adequate for the noise prediction model. A comparison study of five models in terms of the R^2 shows that the three best performance models were the ensemble model, CHAID, and CART. Ensemble models showed the most suitable technique. The method involves weighing several individual models and combining them to improve predictive performance [17]. Many researchers observed a better classification performance of the Ensemble model over others [18]. SPSS Modeler is proven to be effective data mining and predictive analytics software that can be used for noise modeling. Similar conclusions were found in the applications of waste management [19] construction materials [20] communications [21] and health care [22].

In addition, the cross-validation of the Ensemble model indicates that the model is not overfitting. Thus, using the proposed model yields satisfactory results and could forecast the sound level with a reasonable level of accuracy. Additional research on machine learning algorithms, e.g., AdaBoost, Random Forest, Extremely Randomized Trees, etc., is recommended for further research.

References

1. Watts GR, Pheasant RJ. Identifying tranquil environments and quantifying impacts. Applied Acoustics. 2015;89:122-127.
2. Organization WH. Compendium of WHO and other UN guidance on health and environment. World Health Organization; 2022.
3. Shaltout ML, Yan Z, Palejiya D, Chen D. Tradeoff analysis of energy harvesting and noise emission for distributed wind turbines. Sustainable Energy Technologies and Assessments. 2015;10:12-21.
4. Nedic V, Cvetanovic S, Despotovic D, Despotovic M, Babic S. Data mining with various optimization methods. Expert Systems with Applications. 2014;41(8):3993-3999.
5. International Electrotechnical Commission [Internet]. IEC61672-1 Electroacoustics Sound level meters Part 1: Specifications 2013 [cited 2023 March 19]. Available from: https://webstore.iec.ch/publication/5708.
6. McAleer S, McKenzie A. Guidance note on noise assessment of wind turbine operations at EPA licensed sites (NG3). Environmental Protection Agency, Office of Environmental Enforcement. 2011.
7. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science. 2021;7:e623.
8. Ralević N, Glišović NS, Djaković VD, Andjelić GB, editors. The performance of the investment return prediction models: Theory and evidence. 2014 IEEE 12th International Symposium on Intelligent Systems and Informatics (SISY); 2014: IEEE.
9. The International Business Machines Corporation [Internet]. Gains Charts 2021 [cited 2023 March 19]. Available from: https://www.ibm.com/docs/en/spss-modeler/saas?topic=gains-charts.
10. The International Business Machines Corporation [Internet]. Overview of Nodes 2021 [cited 2023 March 19]. Available from: https://www.ibm.com/docs/en/spss-modeler/saas?topic=SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/clef_nodes_intro.htm.
11. U.S. Environmental Protection Agency Office of Noise Abatement and Control [Internet]. Information on levels of environmental noise requisite to protect public health and welfare with an adequate margin of safety 1974 [cited 2023 March 19]. Available from: https://www.nonoise.org/library/levels74/levels74.htm.
12. Pongpirul K. Noise-induced hearing loss (NIHL) and sound control standards for stone crushers (P. 225). Chulalongkorn Medical Journal. 2020;64(2):225-230.
13. Nguyen QH, Ly H-B, Ho LS, Al-Ansari N, Le HV, Tran VQ, et al. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering. 2021.
14. Kim KI, Simon R. Overfitting, generalization, and MSE in class probability estimation with high□dimensional data. Biometrical Journal. 2014;56(2):256-269.

15. Alam W, Aribam B, Singh WR. GIS based Assessment of Noise Environment of Imphal City, Manipur (India): A Comprehensive Study. Universal Journal of Environmental Research & Technology. 2018;7(1).

16. Bošnjaković M, Katinić M, Santa R, Marić D. Wind Turbine Technology Trends. Applied Sciences. 2022;12(17):8653.

17. Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018;8(4):e1249.

18. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Computer methods and programs in biomedicine. 2018;153:1-9.

19. Kavyanifar B, Tavakoli B, Torkaman J, Mohammad Taheri A, Ahmadi Orkomi A. Coastal solid waste prediction by applying machine learning approaches (Case study: Noor, Mazandaran Province, Iran). Caspian Journal of Environmental Sciences. 2020;18(3):227-236.

20. Chou J-S, Pham A-D. Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. Construction and Building Materials. 2013;49:554-563.

21. Li Y, Yao J, editors. The quantity prediction of 4G customers of China mobile communications corporation based on SPSS modeler. 2016 International Conference on Logistics, Informatics and Service Sciences (LISS); 2016: IEEE.

22. ABDAR M. A survey and compare the performance of IBM SPSS modeler and rapid miner software for predicting liver disease by using various data mining algorithms. Cumhuriyet Üniversitesi Fen Edebiyat Fakültesi Fen Bilimleri Dergisi. 2015;36(3):3230-3241.