Failure Detection from the Knocking Sounds Using Convolutional Neural Network

Punyapat Areerob*, Rithea Sum*, Chanon Khongprasongsiri†, Sudchai Boonto*
*Department of Control System and Instrumentation Engineering,
King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand
{punyapat.aree, sudchai.boo, rithea.sum01}@kmutt.ac.th
†National Astronomical Research Institute of Thailand
chanon@narit.or.th

***Abstract*—Toilet quality assurance is a crucial process in ensuring lavatories meet rigorous performance, durability, and hygiene standards. The current standard of Maximum Performance (MaP) testing faces challenges, leading researchers to explore innovative approaches such as sound source classification for quality assurance. This approach involves collecting a diverse dataset of lavatory sounds and extracting relevant acoustic features. Deep learning models, particularly convolutional neural networks (CNNs), are trained on these features to accurately classify sound sources. The trained models were evaluated and compared by considering metrics such as classification accuracy, computational complexity, and model parameters. This paper performs these tests and chooses the most effective model to enhance the quality assurance process for toilets. Incorporating sound source classification techniques has several benefits, including the optimization of testing processes, non-intrusive performance assessment, and efficient resource utilization through targeted testing and troubleshooting. By improving the standards of lavatory quality, this approach ensures enhanced performance, durability, and hygiene of lavatories.**

***Index Terms*—Toilet quality assurance, sound classification, deep learning, high accuracy, low trainable parameters.**

I. Introduction

Toilet quality assurance is an indispensable process of great importance to the industry. This ensures that the toilets meet stringent standards in terms of performance, durability, and hygiene. As an essential fixture in daily life, lavatories play a crucial role in maintaining proper sanitation and providing comfort to users. Whether used in residential, commercial, or public settings, the quality of the lavatory is paramount for ensuring a hygienic and convenient experience for individuals. In the manufacturing industry, quality assurance processes for lavatories are implemented to guarantee that products adhere to specific regulations and standards. This involves rigorous testing and evaluation of various aspects such as flushing efficiency, water usage, structural integrity, and durability.

By conducting thorough quality checks, manufacturers can identify and rectify any potential flaws or defects, ensuring that lavatories meet or exceed industry standards. However, the current standard of Maximum Performance (MaP) for lavatories testing require experience worker to knock each side of the lavatories which often require more time to test and possible human error could be happening all the time that can result in financial losses and inefficient resource utilization. To overcome these challenges, researchers are actively exploring innovative approaches, one of which involves implementing sound source classification for quality assurance in this case. The process of sound source classification for quality assurance begins with the collection of a diverse dataset comprising the recordings of lavatory sounds. These sounds were captured using microphones strategically placed to capture sound from every axis, allowing for comprehensive coverage as shown in Fig. 1. Then the quality of the lavatories is identified using the sound classification method.

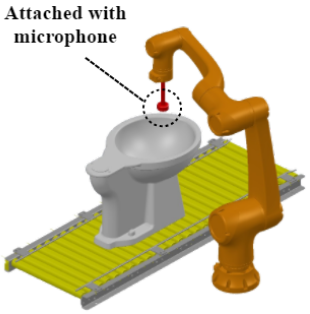


Fig. 1: Methods for collecting sound data for classification of toilet quality

Sound classification has been used in various applications which are categorized as Automatic Speech Recognition (ASR) [1] and Environmental Sound Classification (ESC) [2]. ESC is commonly used to classify the urban sound [3], [4] as it is essential for the applications of smart cities [5]. In the industrial field, ESC is used to identify the sound of interest that happens during the production [6]. In this paper, sound classification is used to identify the quality of the lavatory during the production process in the industrial field.

The classification of sounds is commonly done in three steps starting with sound signal pre-processing, feature extraction from the processed signal, and sound classification [3]. The pre-processing part involves the segmentation of the recorded sound file into a wide range of acoustic characteristics and variations. Then, relevant acoustic features were extracted from the collected dataset to capture the unique characteristics of the different sound sources. These features may include the frequency distribution, amplitude patterns, and temporal characteristics of the recorded sounds. Advanced signal processing techniques are employed to transform audio data into meaningful representations that can

be used by deep learning models. Finally, the model is used to train and classifies the desired signal.

Several benefits can be realized by incorporating sound source classification techniques into quality assurance (QA) processes during production. First, this approach allows for the optimization of testing processes, as sound-based evaluation can be conducted alongside or in conjunction with existing methods. Second, the use of sound as a diagnostic tool can help minimize product damage during testing, as it offers a non-intrusive way to assess performance. Third, the implementation of sound-source classification can improve the efficient utilization of resources by enabling targeted testing and troubleshooting.

QA of lavatory products can be achieved by employing a robust deep learning model for quality classification during production. It is crucial for this model to demonstrate high accuracy in classifying lavatory quality while utilizing a minimal number of training parameters. By employing a lower number of training parameters during the training process, it is possible to enhance the processing speed of the data when implemented on hardware devices, such as FPGA. In this paper, we propose a CNN model and compare it with the state-of-art models in terms of classification accuracy and the required training parameters on the same dataset.

II. PROBLEM FORMULATION

The training of deep learning models, specifically convolutional neural networks (CNNs), involves utilizing extracted acoustic features to enable differentiation between various sound sources. These models are designed to learn and recognize the patterns and characteristics present in audio data, thereby facilitating accurate classification. During the training process, the models were exposed to labelled examples from the dataset, which served as reference points for learning and generalization. The models optimized their performance by repeatedly adjusting their parameters to classify different sound sources effectively. This iterative adjustment ensures that the models become more proficient in accurately categorizing and distinguishing between acoustic patterns and characteristics.

Once the deep learning models were trained, their performances were evaluated and compared to determine the most accurate and efficient approach for sound source classification in lavatory quality assurance. This process involves conducting tests on a separate set of unseen data, commonly called a validation or test dataset. This dataset comprised sound recordings that were not utilized during the training phase. The trained models were then applied to this dataset to classify sound sources and predict their respective classes. Several metrics are considered to assess the performance of the models. The primary metric used was classification accuracy, which measures the percentage of correctly classified sound sources from the total number of samples in the validation dataset. While comparing different models, we carefully analyzed these metrics to identify the model that achieved the highest accuracy and demonstrated the most efficient results. The efficient evaluation considered factors such as computational complexity and the number of parameters associated with each model. These factors are important because they can influence the practicality and resource requirements of deploying the model in real-world applications. By comparing the performance of different deep learning models, we aim to determine the model that provides superior accuracy and efficiency in sound source classification tasks. Through this evaluation process, we can select the most effective approach to predict the acquired dataset to enhance the quality assurance process for toilets. The chosen model ensures the accurate and reliable classification of sound sources with the least usage of training parameters, leading to improved performance, durability, and hygiene standard in lavatory production.

III. EXPERIMENTAL SETUP

A. Dataset preparation

Initially, we use a high-quality USB microphone to capture the sound from multiple subsets of toilets, distinguishing between good and bad. The recorded sound was captured from the 3-axis of each toilet. The recordings were then categorized into six classes: D1, D2, D3, F1, F2, and F3. In this categorization, the initial letters D and F represent good and bad sound detection, respectively. At the same time, indexes 1, 2, and 3 denote the three axes (two sides and one front) from which the toilet sounds were recorded. Then, we performed segmentation on the recorded samples by removing the silent sounds and trimming the samples into multiple segments, which serve as our pre-processed dataset. Finally, we archived a total of 3629 samples, of which the distribution of each class is shown in Table I.

TABLE I: The distribution of each class

D1	D2	D3	F1	F2	F3
528	651	544	560	592	754

Since the recorded sound is in .m4a format, each recording is first converted to .wav format before performing feature extraction. This conversion is necessary because .wav files represent time-series data, which is easier to analyze and extract features from.

B. Feature extraction

Mel-spectrogram, an image-based representation of the sound signal where the frequencies are converted to the mel scale as shown in Fig. 2, will be utilized as the training and testing dataset for the classification method in this paper. The conversion from the time series data to Mel-spectrogram was done using librosa [7], a python package for processing sound signals. Then, each Mel-spectrogram is labeled accordingly based on its class (good or bad). After being labeled, the processed dataset has been divided into a training set (80%) and testing set (20%).

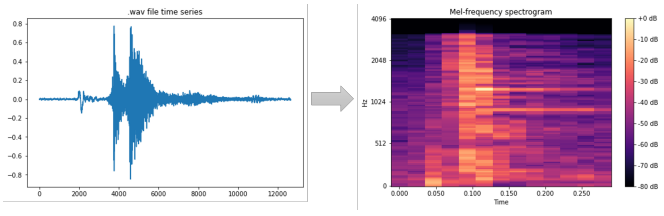


Fig. 2: Conversion from time series data to mel spectogram

C. Model

The resulting features from the feature extraction step are represented as image-based. Therefore, this section presents the proposed CNN model and the state-of-art model for image classification, such as VGG and LSTM, that will be used for evaluating detail. Initially, the mel-spectrogram image needs to be reshaped to the desired input shape of each model.

1) Convolutional Neural Networks: The proposed CNN model used for evaluation with other models is depicted in Fig 3. The model consists of 2 convolutional layers, where each layer utilizes a kernel size of 3*3 to generate a feature map. Each layer are passed through a ReLU activation function and subsequently undergo max pooling with a size of 2*4. The resulting features are then flattened and connected to a fully connected layer with 64 neurons, which is activated by a ReLU activation function. To regularize the features, a dropout layer with a rate of 0.25 is added. Finally, another fully connected layer connects to the output layer and followed by a softmax activation function to produce the output of 2 classification features.

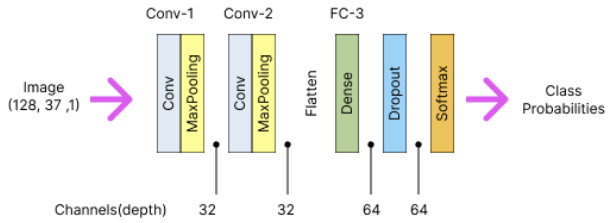


Fig. 3: Proposed CNN model

2) Long Short Term Memory (LSTM): Long Short-Term Memory (LSTM) networks, a special kind of Recurrent Neural Networks (RNN) with an architecture shown in Fig. 4, have been widely adopted for processing sequence data, including sound. In the context of sound processing, LSTM networks can be utilized for tasks such as speech recognition, sound classification, music generation, and audio signal processing.

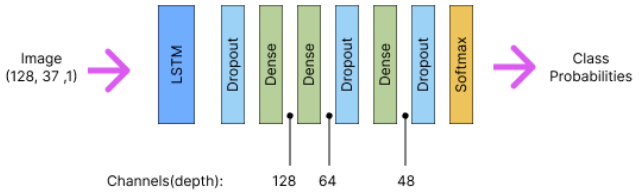


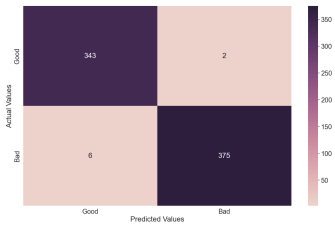
Fig. 4: LSTM model

3) VGG-16: VGG-16 is a Convolutional Neural Network model that consists of 20 layers and is well-suited for image classification tasks [8]. The model has achieved an impressive test accuracy of 92.7% and is ranked among the top-5 performers in the ImageNet dataset. In our problem, the sound signal is transformed into a mel-spectrogram, which is represented as an image. Therefore, VGG-16 is a suitable model to evaluate alongside our proposed CNN model.

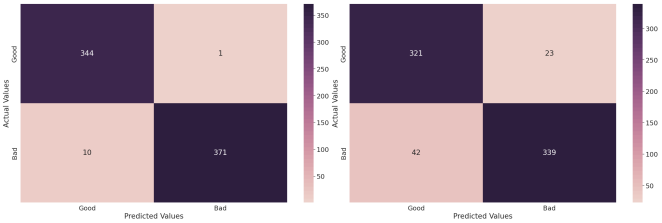
IV. Result

The hyperparameters of these models are trained on the same dataset using NVIDIA GeForce RTX 3090 24-GB GPUs, AMD Ryzen Threadripper 2950X 16-Core Processor, 3.500 GHz CPU, and 128 GB RAM. This paper compares the results in two parts: the accuracy of classifying the sound and the required parameters to train each model.

A. Accuracy



(a) Prediction result of CNN model



(b) Prediction result of VGG-16 model (c) Prediction result of LSTM model

Fig. 5: The classification accuracy of the (a) CNN, (b) VGG-16, and (c) LSTM display in confusion metrics

The classification accuracy results of the three models are illustrated in Fig. 5. Since 20% of the total dataset is allocated as the testing set, the total number of samples in the test set equals 726. The results show that the proposed CNN model performs comparably to the VGG-16 model, with only eight misclassifications for the proposed CNN model and 11 misclassifications for the VGG-16 model. However, both models outperform the LSTM model, which has a higher misclassification rate of 65, nearly eight times greater than the proposed CNN model.

B. Training Parameters

Although the LSTM model exhibits the lowest performance in training and testing accuracy, it requires only approximately 112,978 parameters to train the given dataset, which is around 135 times lower than the required parameters of VGG-16. However, the proposed CNN model still requires about 0.58 times fewer parameters than the LSTM model while demonstrating significant improvements in training and testing accuracy. The result demonstrates that the proposed CNN model has the best accuracy and utilizes the least parameters to train the model compared to VGG-18 and LSTM.

V. Conclusion

In this paper, we compare three different deep learning models for detecting the quality of lavatories during the production process. Our goal is to choose a model with

TABLE II: The comparison of testing and training accuracy with the total required parameters for training the three classification models

Model	train	test	trainable parameters
VGG-16	0.9991	0.9984	15,255,238
LSTM	0.8711	0.8551	112,978
Proposed CNN	0.9996	0.9903	66,434

high accuracy while minimum trainable parameters are required, making the models suitable for implementation in hardware devices. Additionally, the original data is in audio format. These audio files are processed into image-based mel-spectrograms and used as the dataset for the evaluation. The proposed CNN model is compared with conventional models, such as VGG and LSTM, which have demonstrated outstanding performance in image classification problems. The results show that the proposed CNN model achieves nearly identical training and testing accuracies of over 0.99. However, it is worth noting that the VGG model requires approximately 229 times more parameters than the proposed CNN model, proving that it is inefficient for deployment to hardware devices. In addition, the LSTM model utilizes only about 1.7 times more parameters than the CNN model, but it exhibits lower accuracy in training and testing, approximately 12.85% and 13.65%, respectively. This comparison demonstrates that the proposed CNN model is the most suitable for further implementation on hardware devices to detect the quality of lavatories during the production process.

VI. Acknowledgement

The researchers would like to express their gratitude to King Mongkut's University of Technology Thonburi for providing financial support for research activities through the allocated budget for research funding from the government.

References

[1] M. A. Ambewadikar and M. R. Baheti, "Review on speech recognition system for disabled people using automatic speech recognition (asr)," in *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, 2020, pp. 31–34.

[2] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel teo-based gammatone features for environmental sound classification," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1809–1813.

[3] J. K. Das, A. Ghosh, A. K. Pal, S. Dutta, and A. Chakrabarty, "Urban sound classification using convolutional neural network and long short term memory based on multiple features," in *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, 2020, pp. 1–9.

[4] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 1041–1044. [Online]. Available: https://doi.org/10.1145/2647868.2655045

[5] J. P. Bello, C. Mydlarz, and J. Salamon, *Sound Analysis in Smart Cities*. Cham: Springer International Publishing, 2018, pp. 373–397. [Online]. Available: https://doi.org/10.1007/978-3-319-63450-0_13

[6] A. Moubayed, M. Sharif, M. Luccini, S. Primak, and A. Shami, "Water leak detection survey: Challenges research opportunities using data fusion federated learning," *IEEE Access*, vol. 9, pp. 40 595–40 611, 2021.

[7] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *SciPy*, 2015.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." Computational and Biological Learning Society, 2015, pp. 1–14.