IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 June 2024, accepted 19 July 2024, date of publication 29 July 2024, date of current version 6 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434547

RESEARCH ARTICLE

Voice Frequency-Based Gender Classification Using Convolutional Neural Network for Smart Home

NASARUDDIN NASARUDDIN[1], **(Member, IEEE),**
MUHAMMAD AGUNG P. PRATAMA TRESMA[1], **MASDUKI KHAMDAN MUCHAMAD**[1],
AND ZAHRUL FUADI[2], **(Member, IEEE)**
[1]Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[2]Department of Mechanical and Industrial Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia

Corresponding author: Nasaruddin Nasaruddin (nasaruddin@usk.ac.id)

This work was supported by the Ministry of Education, Culture, Research, and Technology, Indonesia, under Grant 590/UN11.2.1/PT.01.03/DPRM/2023.

ABSTRACT The smart home's functional requirements should include the capability to differentiate between various user categories, such as gender and voice recognition. The data-driven Internet of Things (IoT) can present challenges for the elderly and people with disabilities, but voice recognition technology could offer an effective solution. In addition, developing an accurate gender prediction model for voice recognition is still challenging due to the large time variation and randomness. Therefore, we propose gender classification and detection models based on voice frequency using Convolutional Neural Networks (CNN) with ResNet50 and ResNet101 architectures to enhance smart home functionality. We also introduce an algorithm for converting voice frequencies into images to speed up the recognition and detection processes. The research method involves converting voice frequencies into images to expedite the recognition and detection processes. The CNN models were trained and tested with various learning rates using audio datasets. Performance was evaluated through simulations that measured training accuracy, validation accuracy, recall, precision, and F1 scores. The simulation results show high training accuracy: ResNet50 achieved 99.67% and ResNet101 achieved 99.82%. The validation accuracy of the models also exceeded the accuracy of traditional CNN models in previous studies. The simulation results based on recall, precision, and F1 score for each proposed model are 99.3%, 100%, and 99.65%, respectively. Finally, we successfully used the ResNet50 model to create a low-latency smart home prototype. Thus, this paper significantly contributes to the practical applications of voice-based gender recognition in smart home environments with high accuracy and efficiency in detection.

INDEX TERMS Voice frequency, convolutional neural networks (CNN), ResNet50, Resnet101, smart home.

I. INTRODUCTION

The global population of elderly people and individuals with special needs continues to increase. The demand for smart home technology is becoming important to assist them in living comfortably and independently in their daily activities at home. Smart home systems make it easier and more convenient for homeowners to manage their homes and also help save costs [1], [2]. Deep learning for smart homes has been rapidly increasing since 2017, with Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)/Long Short-Term Memory (LSTM) being actively applied in recent years. The analysis of keyword networks revealed that CNN is closely associated with functions such as activity recognition, fall detection, and automation, while RNN and LSTM are linked to areas such as energy management, activity recognition, authentication, and anomaly detection. The research also identified that 14% of

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

the selected studies considered older people as a specific population target, highlighting the importance of addressing the needs of the elderly in smart home technology [3]. A smart home is a technology-based residence where all home equipment can be managed and centralized in one place, allowing for local or remote control. Currently, the Internet of Things (IoT) plays an important role in supporting smart homes by utilizing data information [4]. Overall, deep learning in smart homes aims to offer a simple, flexible, reliable, and affordable smart home automation system that addresses security, privacy, and decision-making challenges for IoT devices [5]. IoT is attracting significant attention and growth potential because IoT technologies and solutions make our lives easier and offer many things we can experience in our homes. For instance, it enables the control of electronic devices like lights, refrigerators, TVs, and doors in a smart home, allowing for functions such as opening, closing, and locking. Since IoT is based on data information, it requires a certain level of operational expertise for users. This could sometimes be inefficient for certain groups of users, such as the elderly and disabled. To solve the problem, voice recognition can be used. IoT-based voice recognition involves artificial intelligence connected to the internet [6], [7]. The smart home concept is a result of integrating IoT to provide a more efficient household experience through the use of a voice control system [8].

Voice processing and detection is a way of communicating with electronic equipment in a smart home [9]. The IoT bot intelligence system is designed to emulate an interactive conversation with the user. The conversation that occurs between a computer and a human is a response to a program that has been declared in the program database [10]. The Bot Intelligence system's primary issue is adjusting to societal conversation changes by leveraging patterns already present in the database. To produce a correct response, matching responses/answers that have previously been determined must be transformed into deep learning. The smart home is one of the IoT applications that uses deep learning [11]. Thus, speech recognition in computing becomes more efficient at overcoming this problem [12]. The main objective of speech recognition in personalization is to achieve better usage acceptability and higher quality of experience. The advantage of presence detection using audio as a modality is its lower computational complexity, lower cost, and better acceptability among users. The user's acceptability is tightly connected with data privacy. In the case of audio, a well-considered design can lead to local processing, without the need to use cloud-based speech-processing services [13]. Smart home systems can offer personalized solutions to improve the overall user experience by identifying the individuals responsible for the home. This includes room settings, activity schedule notifications, and other personalized adjustments based on user preferences. Moreover, voice recognition can help smart home systems identify users interacting with the device. This is useful for providing access to specific user-related settings and personal information. For this reason, a voice recognition system based on deep learning to identify users is crucial for smart homes.

Speech recognition is one of the significant technologies in the context of the dynamic development of the Industry 4.0 era [14], [15]. This technology has wide applications, one of which is in the analysis of voice frequency characteristics to identify a person's gender. This is based on the noticeable differences between male and female voices, especially regarding intonation frequency. Voice frequency can be used as an effective method to distinguish the gender of individuals [16]. In general, the frequency of men's voices ranges from 65 to 150 Hz, while the frequency of women's voices ranges from 210 to 500 Hz [17]. Therefore, voice characteristics are an effective way to recognize and distinguish one individual from another. Convolutional Neural Network (CNN) is used to classify gender based on voice due to its effectiveness in capturing spatial and temporal features of speech signals. CNN also aims to address the challenging task of recognizing age and gender from speech signals. This has been a complex issue in speech processing because of the challenge of extracting salient high-level speech features and designing suitable classification models. This approach enables CNNs to effectively learn and detect important gender-related features from speech spectrograms, leading to high performance and robustness in gender classification [18]. Currently, with advancements in information technology, gender recognition, and identification can be achieved using computer algorithms. Digital signal processing, Python programming, and the development of CNN have all been utilized in developing voice recognition and gender identification applications. Apart from that, information technology is also used in data processing and voice feature extraction to construct voice models and match them with the voice to be tested.

There have been several previous studies on voice-based gender detection and recognition using machine learning [19], [20]. However, creating an accurate gender prediction model for voice recognition remains complex and challenging. One of the main problems in gender recognition is signal processing due to its large time variation and randomness. This also poses challenges when using deep learning in selecting human voice features to determine and classify gender. Significant progress has been made in developing algorithms for gender detection based on voice characteristics, but there remains a need for greater accuracy and reliability. Voice-based gender detection systems with high accuracy are needed with the emergence of technology that exploits gender information to improve performance. A limitation of current speech extraction models is their ability to classify speakers based on age and gender [18]. Therefore, there is a need for further exploration of different CNN architectures and optimization techniques to enhance these models' robustness and generalizability. The integration of gender recognition models into smart home functions needs further exploration to create more comprehensive and intelligent smart home systems.

This paper proposes a deep learning model to detect gender-based voice frequency using an audio dataset. The audio dataset is then converted into a spectrogram image, which will be utilized for training and testing a CNN architecture using VisualStudioCode. For this reason, this paper introduces a voice-to-image conversion algorithm. Based on literature studies, this research explores classification using audio datasets and deep learning for gender detection, focusing on the utilization of the ResNets architecture. In this paper, the model processes the results of the input audio and then analyzes them using two ResNet models: ResNet-50 and ResNet-101. Each ResNet model is tested and evaluated to determine its performance. This research aims to conduct gender identification based on voice frequency spectrograms and evaluate the performance of two models using various evaluation metrics, including accuracy, precision, recall, and F-score. This paper also explores the potential applications of speech recognition technology in smart homes. This research proposes an integrated prototype of speech recognition technology with smart homes to automatically adjust internal settings, such as lighting and mosquito traps, based on occupant gender identification. Thus, creating a more personalized and connected experience in the smart home.

The following are the contributions of this paper.

a. We developed voice frequency-based gender detection and classification models using CNN. This paper uses the architecture of ResNet50 and ResNet101 for accurate gender detection in smart home environments. This aims to improve real-time gender detection by adapting these models for smart home applications.
b. We introduce a new algorithm to convert voice frequencies into images. This voice-to-image conversion technique is designed to harness the power of deep learning models, particularly CNNs, to improve the speed and accuracy of the recognition process. This technique advancement enables more efficient processing of audio data, facilitating faster and more precise gender classification.
c. We obtain an efficient ResNet deep learning architecture model for gender detection based on voice frequency, specifically designed for smart home applications. The proposed model shows high accuracy and performance compared to existing models and demonstrates the effectiveness of deep learning in a smart home application. The proposed ResNet50 and ResNet101 architectures result in remarkable performance, demonstrating the superiority of the proposed models.
d. We develop a prototype of smart home systems equipped with smart devices, applying the ResNet50 model. The prototype demonstrates that it has low latency and can be deployed in the real world. Our system also explores the potential applications of speech recognition technology in smart homes and provides users with a personalized and connected experience.

In summary, this paper contributes to developing an advanced gender classification model, presents a novel voice-to-image conversion algorithm, demonstrates high model performance, develops a practical smart home prototype, and improves the overall user experience through personalized smart home applications. These contributions collectively advance the field of smart home technology, especially in the application of deep learning to voice recognition.

This paper is organized as follows. Chapter I presents the introduction, providing an overview of the research objectives, the significance of gender recognition in smart home technology, and the potential applications of voice recognition in artificial intelligence-connected smart homes. The related works are provided in Chapter II, which reviews studies in the field of deep learning, activation functions, transfer learning, and gender and region detection from human voice using deep learning methods. Chapter III outlines the methodology employed in this research, detailing the implementation model and prototype of the proposed ResNet model for smart home applications. It also discusses the training and validation methods used for the Convolutional Neural Network (CNN) models. In Chapter IV, the results of the simulation are presented, including training accuracy, recall, precision, and F1 score of the proposed models. This section also discusses the prototype testing results and the potential applications of the proposed model in smart home systems. Chapter V offers conclusions drawn from the study, summarizes the paper's contributions, and discusses potential future research areas in integrating deep learning and smart home technology.

II. RELATED WORK

Audio signal processing refers to the manipulation and analysis of audio signals using various techniques and algorithms. This involves converting sound waves into electronic signals, which can then be modified, enhanced, or analyzed for a variety of applications. The field has seen significant progress over the past 25 years, driven by consumer expectations, technological capabilities, and the trend toward data-driven methods based on machine learning and deep learning. Audio signal processing encompasses various areas such as speech enhancement, noise reduction, speaker separation, dereverberation, and acoustic scene analysis. It also plays an important role in applications such as audio coding, telecommunication, music retrieval, and sound field analysis. The field continues to evolve with an increasing emphasis on data-driven methods and the integration of other methods such as vision for speaker localization and separation [21]. The purpose of the study in [22] is to audio signal examine how deep learning methods may be used for processing. The study covered several pertinent audio processing categories, such as music analysis, ambient sound analysis, and speech processing. The paper examines different deep learning models and feature representations utilized in these domains while exploring applications such as audio recognition, synthesis, and signal quality enhancement. It also examines the existing challenges and limitations in this field and proposes potential future research directions. Finally, this paper underscores the importance of applying data augmentation techniques and employing suitable performance evaluation metrics in deep learning for audio signal processing.

Another study in [23] the use of deep learning algorithms in the process of identifying and monitoring the behavior of elderly individuals in a healthcare environment was investigated. The study presents a comprehensive review of recent advances in this domain, examining developments, methods, and applications of smart devices in the context of identifying elderly behavior in real-world situations. It focuses on the exploration of deep learning algorithms, especially convolutional neural networks, and provides a taxonomy of data derived from relevant literature. Furthermore, the research highlights the challenges encountered and casts an eye toward the future by detailing potential research directions. This will serve as a foothold for the formulation of new deep learning algorithms and machine learning-based smart devices to support healthcare for elderly individuals.

A framework that aims at predicting future activities in a smart home environment by utilizing data on the activities of daily living (ADL) is presented in [24]. ADL refers to the routine activities that individuals typically perform daily, such as eating, bathing, dressing, and household chores. In the context of smart homes, ADL data is collected through sensors to monitor and analyze residents' behavior. This data can be used to recognize and predict resident activities, which can be used in various applications such as healthcare, energy consumption, recommendation systems, and anomaly detection. ADL data is a valuable asset for understanding residents' behavior and providing optimized services in smart home environments. The proposed framework integrates data embedding algorithms with neural networks in bidirectional long short-term memory (BiLSTM). The evaluation of the framework has been conducted on a dataset depicting a real ADL situation and resulted in a high level of accuracy. In addition, the study also provides an in-depth understanding of neural networks, LSTM, and word embedding techniques as a theoretical foundation. A discussion of previous work related to human activity recognition and prediction is also outlined in this paper. Overall, the research prioritized achieving accurate smart home activity prediction in the future by combining data embedding algorithms and BiLSTM models.

Convolutional neural network (CNN) is the most well-known and common algorithm in deep learning (DL). The main advantage of CNNs is their ability to automatically identify relevant features without human assistance. CNNs have been widely applied in various fields, including computer vision, language processing, and facial recognition. A commonly used type of CNN, which is similar to a multi-layer perceptron (MLP), consists of many convolution layers beginning with a sub-sampling (pooling) layer, while the final layer is the FC layer. An example of CNN architecture for image classification is illustrated in Figure 1 [25].

Using CNN can be effective in classifying gender and age based on voice recognition [26]. The study uses open data and audio pre-processing methods to train the algorithm. It also uses an architecture comprising 4 Conv2D layers and uses the Batch Normalization method. However, there are limitations to neural networks, such as the need for a large amount of training data and long training time. The study suggests the use of more complex architectures and further research on the dependency of accuracy on parameters, such as batch size and optimization algorithms. Figure 2 depicts the "Max Pooling" procedure used in CNNs for machine learning and computer vision. It can be explained as follows.

1. There are sixteen cells in this matrix, which are separated into four colored portions of four cells each. Every cell contains a numerical value. For instance, the numbers (29,15,0,100) are present in the green-colored region. \
2. The label 2×2 pool size indicates the size of the pooling operation.
3. This matrix has four cells representing the max pooling operation of the top section. These values are 100, 184, 12, and 45, which are the maximum numbers from each cell.

The input matrix is divided into sections—in this case, 2 × 2 sections—and the maximum value is extracted from

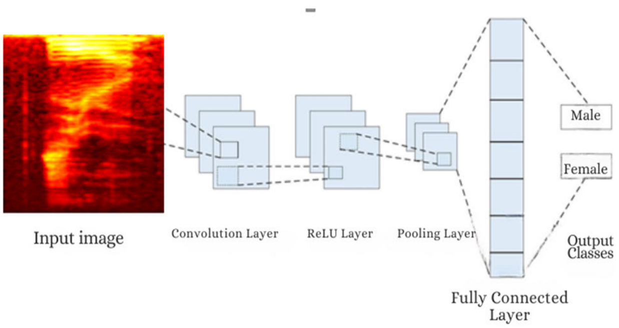
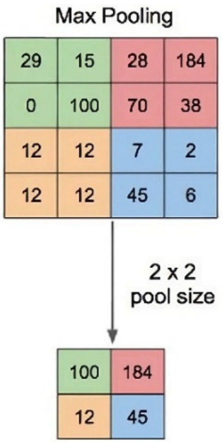


FIGURE 1. An example of CNN architecture for image voice classification [25].



29	15	28	184
0	100	70	38
12	12	7	2
12	12	45	6

100	184
12	45

FIGURE 2. An example of Max-Pooling [27].

each sector during the max pooling operation. For instance, since 100 is the largest number in the green-coloured section (29, 15, 0, 100), 100 appears in the corresponding position on the bottom matrix. This process is repeated for every section of the input matrix to create the output matrix. The max pooling operation operates in this manner.

Max pooling is a feature of CNNs that can be utilized in the speech gender recognition process. The following layer of the CNN can then use this matrix as input. CNNs can be trained to identify patterns in voice characteristics that distinguish between male and female voices. This can be used to predict the gender of a given voice in the context of voice gender detection. It should be emphasized that this is merely a basic illustration. In reality, the procedure would involve matrices of much larger dimensions, along with potentially multiple convolutional and pooling layers. Furthermore, speech gender recognition typically requires other procedures like voice feature extraction and normalization [28]. Max-pooling is a subsampling technique in neural networks that is used to reduce the dimensionality of data by taking the maximum value of each window of a fixed size over a larger dataset [29].

The fully connected layer is the final component in the CNN architecture that connects all features extracted from the previous layer to the final output. The fully connected layer consists of neurons fully interconnected with each neuron in the previous layer so that each neuron in the fully connected layer receives input from every feature extracted from the previous layer. Figure 3 is a description of the 3 fully connected layers, i.e., the input layer, the hidden layer, and the output layer. The input layer is where the input is a combination of all matrices obtained in the pooling layer process, where the matrix has been converted into a 1-dimensional vector (flattening). The hidden layer, the layer between the input layer and the output layer, is responsible for processing input data and extracting increasingly abstract and complex feature representations. Each neuron in the hidden layer receives weights and biases that are adjusted during the network's training process to learn patterns present in the input data and output layer, the last layer of the neural network. It produces a prediction or output from the network based on the feature representation that has been learned by the hidden layer [30].

This architecture was created to overcome the problem of gradient loss in very deep neural networks by introducing residual connections. This residual connection allows the network to learn the residual function, eliminating the need to directly learn the underlying base mapping [32]. Table 1 describes the structure of the ResNet architecture, where ResNet50 and ResNet101 each consist of 5 parts; conv1, conv2, conv3, conv4, and conv5. The difference is that in ResNet50 the conv4 section is convolved 6 times, while in ResNet101, it is convolved 23 times. This makes ResNet101 have a deeper layer than ResNet50.

Using spectrograms as input to the system and the importance of gender information in reducing classification variability has been discussed in [33]. It is shown that the use of neural network model development in identifying gender in speech using the Common Voice dataset was very effective. The study also discussed the advantages of using the ResNet50 architecture, which has a better performance in gender classification in speech compared to other specialized and traditional models. In addition, ResNet50 also can overcome the vanishing gradient problem that often occurs in deep-training neural networks. Although ResNet50 has more layers than ResNet34, it still has higher accuracy and the capability to overcome the overfitting problem. However, the use of ResNet50 also has a trade-off between accuracy and model size, so it needs to be considered depending on the objectives and available resources. The pre-trained ResNet 50 achieved 98.57% accuracy, which is better than the traditional ML approach and previous works reported with the same data set [33]. Previous research in [34] focuses on developing an Internet of Things (IoT)–based elderly fall detection model using a deep learning approach, which aims to improve smart home care. They utilize a combination

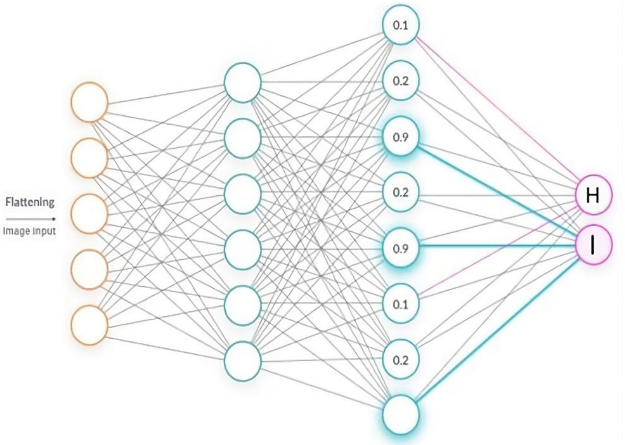


FIGURE 3. Final component fully connected layer in convolutional neural network (CNN) architecture [30].

TABLE 1. Structure of the ResNet architecture [31].

Layer name	Output Size	50-layer	101-layer
Conv1	112×112	7×7,64, stride 2	
Conv2_x	56 × 56	3×3 max pool,64, stride 2	
		$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$
Conv3_x	28 × 28	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$
Conv4_x	14 × 14	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times 6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times 23$
Conv5_x	7 × 7	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times 3$
	1 × 1	Average pool, 1000-d fc, SoftMax	
FLOPs		3.8×10^9	7.6×10^9

of IoT devices, smartphones, and deep learning algorithms to recognize and detect falls in a smart home environment. Video data obtained from IoT devices undergo pre-processing before being directed to a CNN for extraction of relevant features. The model training process involves several stages, including hyperparameter tuning and the use of appropriate classification algorithms. When the model detects a fall incident, notifications are automatically sent to nurses and hospital management. Validation of the model was conducted using a previously collected dataset in fall detection, and the results demonstrated a high level of accuracy. The study also presents the details of the working process of the proposed model, which includes data acquisition, pre-processing, feature extraction, parameter tuning, and classification phases.

III. MATERIAL AND METHOD

A. DATASETS

The dataset used in this paper was downloaded from Kaggle (https://www.kaggle.com/datasets). The dataset consists of 3000 audio files with two class labels: male and female. The dataset has the following characteristics: (1) German accent; (2) age range of 25-29 years old; and (3) duration of each audio between 0.1 and 1 second. The audio was converted into an image based on a frequency spectrogram with a size of 224 × 224 pixels. Furthermore, the dataset is divided into 80% for training, 10% for validation, and 10% for testing. This dataset will influence the training outcomes and model correctness. Therefore, the age, accent, and gender used in this model will be more accurate when the user's attributes are similar to those in the existing dataset.

B. PROCESS

The classification process for CNN is shown in Figure 4. It started by acquiring the datasets from the source, converting voice into images, and using them for the classification model. The model was designed with characteristics of a 10^{-4} learning rate size, by [35], and an image dimension of 224 × 224, following the ResNet architecture [36], [37]. 1500 images based on male voice frequency and 1500 images based on female voice frequency were used to train the model, with two classes and several epochs of 200 by [38] and [39]. If the designed model is appropriate and no errors occur, it will be further evaluated and tested. Otherwise, it will be redesigned. The evaluation process was conducted to determine if the model was functioning correctly. This process was observed regarding accuracy, learning curve, and confusion matrix. During the testing process, the model was tasked with generating output, which was evaluated by analyzing the accuracy and the loss values.

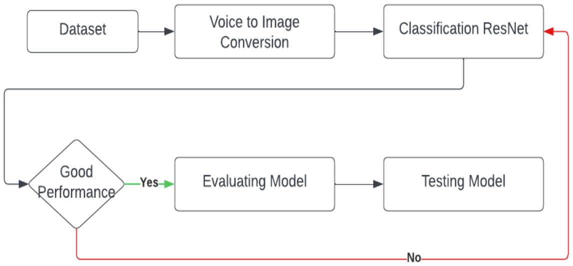


FIGURE 4. Process classification using convolutional neural network.

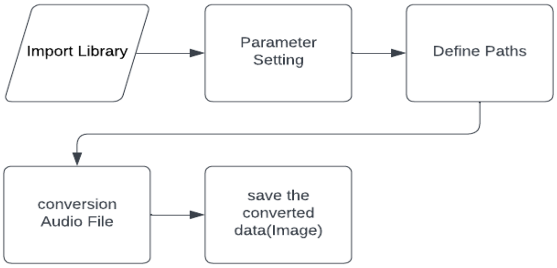


FIGURE 5. Voice to image conversion.

C. VOICE TO IMAGE CONVERSION

The voice-to-image conversion process is illustrated in Figure 5. The process starts with importing the essential libraries. These libraries include the routines and methods required for audio conversion. Next, the audio conversion parameters are configured. These characteristics could include elements such as the bit rate, sampling rate, and audio format. The next step is to specify the paths for saving the audio file and image data. After completing these preliminary stages, the audio file is transformed using the imported libraries and the specified settings. Finally, the converted data, now in the form of an image, is saved in the previously specified location. The audio is converted into an image representation using a spectrogram transform, with values such as n_fft, hop_length, n_mels, and output_size specified. This process involves importing libraries such as os, NumPy, Matplotlib, librosa, PIL, and skimage.transform, specifying the spectrogram parameters, defining the audio data path (data_dir), and the path where the spectrogram is stored (output_dir). Next, there is a loop that iterates through the audio file. In this loop, the program retrieves the audio path, computes the spectrogram, converts it to a logarithmic scale, adjusts its size to 224 × 224 dimensions, and normalizes the values within the range of [0, 1]. The spectrogram is then converted into a color representation using a "heat" color map and stored as an image with values in the range [0, 255]. The entire audio data was processed to enable effective speech pattern recognition for gender detection using a CNN model. The algorithm for voice-to-image conversion is presented in Algorithm 1.

D. CLASSIFICATION USING RESNET ARCHITECTURE

The classification process involves training and validating datasets using the ResNet50 and ResNet101 architectural structures shown in Figure 6. The main objective of this stage is to generate a model. The result of this classification process is an output consisting of two classes identified as male and female. In the ResNet classification stage, the dataset is divided into test data, which constitutes 10% of the overall

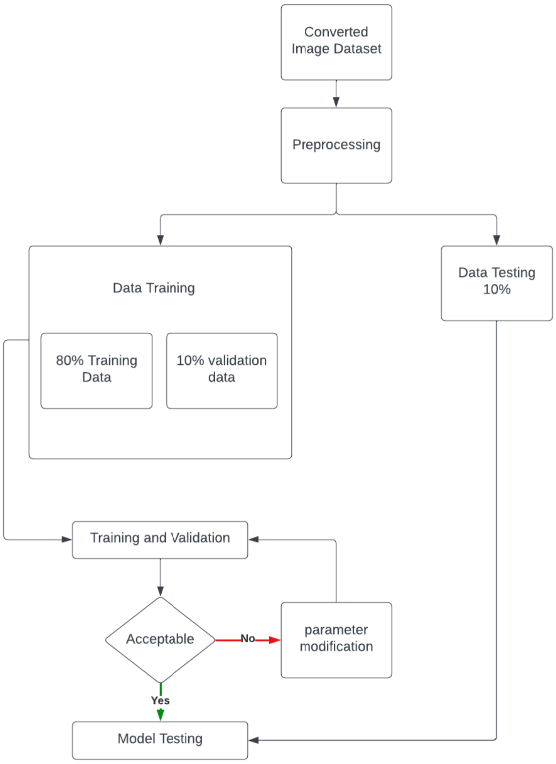


FIGURE 6. Classification using RESNET architecture.

data, and training data, which constitutes 80% of the overall data. These data differ from validation data, which constitutes the remaining 10% of the overall data, as explained in Sec. III-A.

E. TRAINING AND VALIDATION

The process starts with inputting the frequency image dataset that will be used in training the model, as shown in Figure 7. Next, the relevant classes, namely male and female, are labeled. Before starting the training, the network parameters, such as the size of the convolution filter, the number of residual blocks, and the size of the fully connected layers, were carefully set. Parameter settings include using the SGD optimizer, running 200 epochs, using a batch size of 32, and setting the learning rate at 10^{-3}, 10^{-4}, and 10^{-5}. During training, convolution serves as the core operation in a convolutional neural network. At this stage, the input images or features are processed through a series of convolution filters. These filters play an essential role in extracting key features from the image, such as edges, textures, and other patterns. Residue blocks are utilized in the ResNet architecture to train deep networks using shortcut connections, which helps prevent the problem of gradient limitation.

Max pooling is utilized to decrease the dimensionality of the convolved features, thereby enhancing their invariance to minor shifts in objects in the image. The extracted features are then converted into vectors and passed to the fully connected layer to be associated with the classes in the classification problem. Finally, the network is terminated with a SoftMax activation layer that generates class probabilities, which are required in multiclass classification. Once all these stages are completed, the resulting model is ready to be used in the testing phase, where its reliability will be tested.

Algorithm 1 Proposed Voice-to-Image Conversion

```
1. Read the audio signal as input.
      Read audio signal as input s($\tau$)
      Apply time window to the audio signal w($\tau$-t)
2. Short Time Fourier Transform (STFT)
      For each point in time (t) in the audio signal:
      Calculate the STFT (t, $\omega$) value.
         STFT (t, $\omega$) $= \int_{-\infty}^{\infty} S(\tau)\, \mathrm{w}(\tau - \mathrm{t})\mathrm{e}^{\{-\mathrm{j}\omega\tau\}} \mathrm{d}\tau$
         $\omega$ is the frequency
3. Quadratic Modulus
      For each calculated STFT (t, $\omega$) value:
      Calculate the square modulus.
      M (t, $\omega$) = |STFT (t, $\omega$) |2
4. Conversion to Mel Scale
      For each value of linear frequency (F) associated with
      $\omega$:
      Convert linear frequency (F) to mel frequency (mel(f))
         mel(f) $= 2595 * \log10(1 + \mathrm{F}/700)$
5. Intensity of Mel-Spectrogram Image
      For each pair (t, $\omega$) or (t, F) in the spectrogram image:
      Find the frequency value mel that corresponds to
      $\omega$ or F (use mel(f))
      Calculate the intensity of pixels in the mel-spectrogram
      image (I (t, m))
         I (t, m) = M (t, mel$-$1(m))
         m is the frequency index of mel corresponding to
         mel(f)
6. Mel-Spectrogram Image
      Each pair (t, m) in the mel-spectrogram image now has
      a corresponding intensity.
      The mel-spectrogram image can now be used for
      visualization or further analysis
```

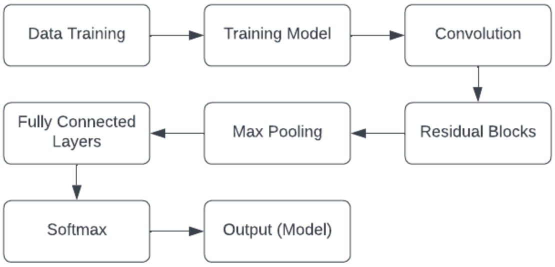


FIGURE 7. Training and validation.

F. EVALUATION METHOD

The standard evaluation method for classification models includes the confusion matrix, accuracy, recall, precision, and F1 score [40]. The confusion matrix provides a complete output and performance of the model.

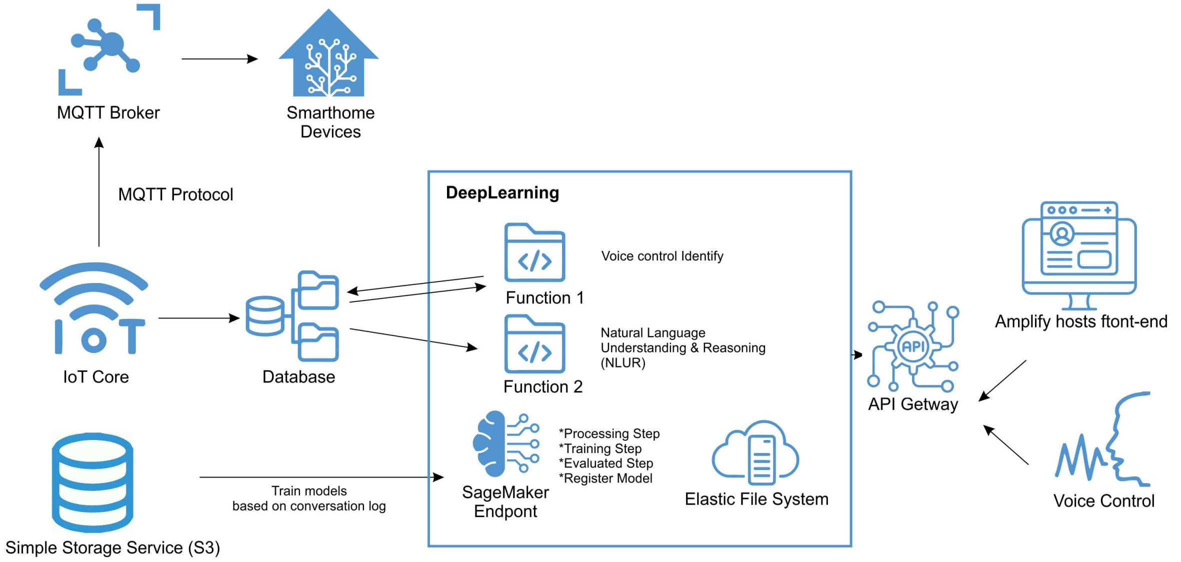


FIGURE 8. The implementation model for a smart home system.

The performance evaluation of the classification model comprises 4 terms: true positive (TP), true negative (TN), true positive (TP), and false positive (FP). Each standard of the performance evaluation can be described as follows.

Accuracy refers to the comparison of correct predictions (both positive and negative) against the entire dataset. It measures the proximity of the predictions to the actual true value. It can be calculated as:

$$\textbf{Accuracy}=\frac{(\boldsymbol{TP}+\boldsymbol{TN})}{(\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{TN}+\boldsymbol{FN})} \tag{1}$$

Recall is the proportion of correct predictions in the positive category to the total data that truly belongs to the positive category. Furthermore, recall can be interpreted as a measure of the model's ability to identify data that belongs to the positive category compared to the actual data that belongs to that category. It can be calculated as:

$$\textbf{Recall}=\frac{(\boldsymbol{TP})}{(\boldsymbol{TP}+\boldsymbol{FN})} \tag{2}$$

Precision describes the ratio of accurate positive predictions to the total predictions identified as positive. Precision represents the level of accuracy between the data and the prediction results obtained from the model. It can be expressed as:

$$\textbf{Precision} = \frac{(\boldsymbol{TP})}{(\boldsymbol{TP}+\boldsymbol{FP})} \tag{3}$$

F1 Score is the weighted average comparison between recall and precision. It can be presented as:

$$\textbf{F1Score}= 2*\frac{(\boldsymbol{Recall} * \boldsymbol{Precision})}{(\boldsymbol{Recall}+\boldsymbol{Precision})} \tag{4}$$

The sigmoid function, also known as the logistic function or squashing function, is a mathematical function commonly used in various applications, especially in machine learning and artificial neural networks. Sigmoid functions have an S-curve shape that is useful for converting numerical inputs into values that fall within the range of 0 to 1. The most common sigmoid function is the logistic function, also known as the logit function. The sigmoid function is given by [39]:

$$f(x) = \frac{1}{1+e^{-x}} \tag{5}$$

where $\mathbf{f}(\mathbf{x})$ is the output of the sigmoid function for input $\mathbf{x}$, $\mathbf{e}$ is the Euler number (approximately 2.71828), and $\mathbf{x}$ is the numeric input to be converted: a value between 0 and 1. The sigmoid function takes the input value x and produces a value between 0 and 1. When x approaches minus infinity (-∞), $\mathbf{e}^{(-x)}$ approaches zero, thus f(x) approaches 0. Conversely, when x approaches plus (+∞), $\mathbf{e}^{(-x)}$ approaches infinity, so f(x) approaches 1, in the middle of the range of x values, the sigmoid function produces a value of about 0.5 [39].

IV. IMPLEMENTATION MODEL FOR SMART HOME

In this paper, we also introduce a proposed ResNet-based deep learning implementation model for smart homes. The implementation model for a smart home has several parts, as illustrated in Figure 8, and can be described as follows.

A. INTERNET OF THING (IoT)

The IoT protocol used in implementing the model for smart homes is Message Queuing Telemetry Transport (MQTT), which is specifically designed for communication with low bandwidth or limited to IoT devices. The MQTT protocol operates based on the publish/subscribe principle. In traditional network communications, home appliances and servers communicate with each other directly. Connected household appliances request resources or data from the server, and then the server processes and sends back responses as operational commands for household appliances. However, MQTT uses a publish/subscribe pattern to distinguish message senders (publishers) from message recipients (home devices) via message brokers. These brokers manage communications between publishers and subscribers. The broker's task is to filter all incoming messages from publishers and distribute them correctly to subscribers.

B. DEEP LEARNING

Deep learning models applied to smart homes must be able to recognize complex patterns in sound to generate accurate insights and predictions. Using deep learning methods to automate tasks that typically necessitate human intelligence, such as describing sound frequencies or transcribing sound files into text. Computers use deep learning algorithms to gather insights and meaning from voice message data, which is subsequently transcribed into text and documents. The ability to process human-generated natural text has several use cases, including in smart home controller functions. In this study, we used the ResNet50 model. In the process of implementing DL, there are two main components, as follows.

- Voice Control Identity: Lambda runs code on high-availability computing infrastructure and performs all computing resource administration, including server and operating system maintenance, capacity provisioning and autoscaling, and logging. In voice-controlled identity functions, the user's chat experience is determined by the following components: Intent: (desired action), Utterance (user Input), Prompt (request for data), Slots (required data), and Fulfillment (completed action).
- Natural Language Understanding & Reasoning (NLUR) [8], [41]: It is a top-down approach to deep learning-based voice control models for smart homes. This approach allows users to control various electronic devices through natural language queries. It enhances the user experience by providing a more intuitive way to interact with technology. The NULR approach is designed to enable seamless communication between users and smart home systems. Full-discourse NLUR via domain-specific network-based advanced reasoning techniques have been utilized. In dialogue pairs, when commands indicate that they cannot be answered or executed, the system will prompt the user again in the form of a question. However, the system may continue with random statements such as "I will look for it", "Did I misunderstand you?", "What does that mean?", "That's an interesting question.", "I'll come", "Be back there in a moment", etc.

C. SAGEMAKER ENDPOINT

SageMaker is used as a platform for building and training deep learning models, in this case, it is ResNet50, and subsequently deploying them directly to a production-ready hosting environment. In its implementation, the model and data are stored in a simple storage service (S3) which is optimized to run efficiently on very large data in a distributed environment. The way the deep learning model works using ResNet is for object detection and semantic segmentation. This architecture was created to address the problem of vanishing gradients in very deep neural networks by introducing residual connections. These residual connections enable the network to learn residual functions, eliminating the necessity to directly learn the underlying basic mapping. Testing data resulting from deep learning processing is stored in the Elastic File System (EFS) because EFS provides serverless file storage and is fully elastic. This allows to sharing of file data without the need to provide or manage storage capacity and performance. EFS is built to scale on demand to petabytes without interrupting the application, growing and shrinking automatically as files are added and removed.

Interface design can be directly visible or accessed through specific devices. For example, when you turn on a light switch, the light will come on without you needing to understand the process that occurs behind it. In this study, researchers used a web-based program that allows users to interact with their devices either directly or through a network. Meanwhile, in communication between servers, the Gateway API serves as a connecting bridge. The Gateway API is a key component in Microservices Architecture and other solutions that enable efficient communication between various applications and services for both webmasters and users.

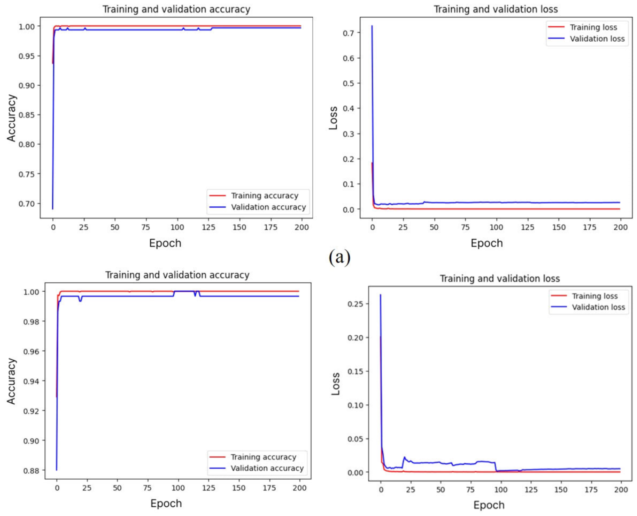
V. RESULT AND DISCUSSION

A. TRAINING AND TESTING MODEL RESNET50 AND RESNET101

In this stage, model detection will be performed using the ResNet50 and ResNet101 architectures by implementing the same appropriate hyperparameters. Next, the model will be evaluated using 200 epochs, where each epoch will be tested with a learning rate of 10^{-4}. From the evaluation results, the model will generate several metrics, such as confusion matrix, training accuracy, training loss, validation accuracy, validation loss, as well as recall value, precision value, and F1-score. Finally, the model will generate a learning curve to assess the stability of the model.

The training results, validation accuracy, and loss of ResNet50 and ResNet101 are presented in Figures 9(a) and 9(b), respectively. The resulting ResNet50 model is unstable from 150 epochs downwards both in terms of accuracy and loss. However, above 150 epochs, it becomes stable again. The ResNet101 model exhibited a decline in validation accuracy during the initial 25 epochs, but later regained stability. In the validation data, there was a decrease at 25 epochs but then stabilized again until epoch 200.

The models created previously and taken the best from ResNet50 and ResNet101 are tested using a dataset comprising 10% of the data and several new datasets containing audio with different accents that were separated during model training. This test evaluates the model's performance in detecting



(a)

(b)

FIGURE 9. Training and validation accuracy and validation loss at a learning rate of 10^{-4} for (a) Resnet50 and (b) Resnet101.

gender and assesses its performance with various datasets other than the training dataset. So, the performance evaluation of the model is determined using the sigmoid function as shown in equation (5).

The model will be considered good if it can achieve a probability close to 1, as shown in Figure 10. The figure displays intriguing results from gender classification experiments utilizing two prominent neural network architectures, ResNet50 and ResNet101. We can observe informative columns that clearly illustrate the classification process of each tested image. Each row of the figure contains valuable information, including the image being analyzed, the assigned label (''Male'' or ''Female''), and the resulting classification probability for each category. The average probability for ResNet 50 is 0.99, while for ResNet 101, it is 1.

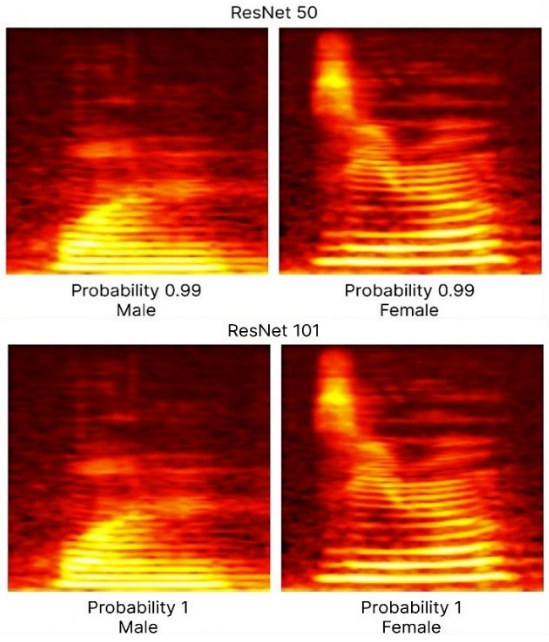


FIGURE 10. Testing Model ResNet 50 and ResNet101.

B. PERFORMANCE EVALUATION

The proposed models are then analyzed and evaluated using a confusion matrix that describes the model's performance in classifying the dataset. The confusion matrix also serves to measure the extent to which the model can capture existing information from the dataset. In addition, the confusion matrix can be used as an evaluation tool to determine the model with the best performance during testing.

Fig. 11 shows the confusion matrix for the ResNet50 and ResNet101 architectures. From the results, it can be seen that the model successfully predicted the male class accurately without any prediction errors. However, in the female class, the model successfully predicted 149 images correctly, but there was 1 image from the female class that the model failed to predict accurately. Both models have the same convolution, which could be attributed to being initialized with identical weights, and the training process starts from the same conditions. Both models achieve the same accuracy rate of 99.67%, with a recall rate of 0.933 and a precision rate of 1, as shown in Tables 2 and 3.

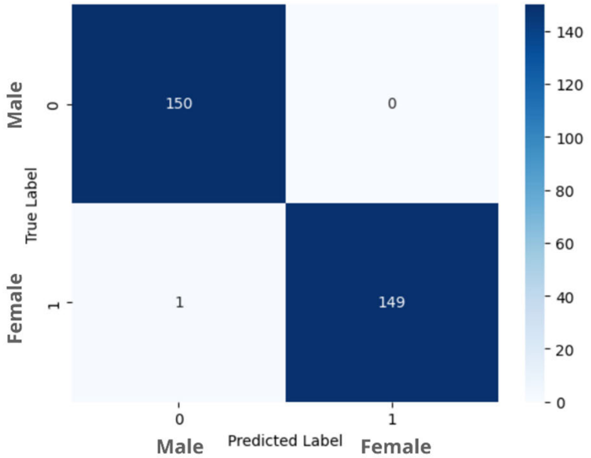


FIGURE 11. Evaluation model.

The main difference between the two models lies in the learning rate results of 10^{-4} and 10^{-5}, respectively. Besides the high accuracy rate, the analysis of the learning curve and the test results indicate that both models, ResNet50 and ResNet101, exhibit good performance and are well-fitting. The evaluation results suggest that both models perform equally well in detecting gender.

Table 4 presents the results of testing each model based on equations (1) to (4), demonstrating that both ResNet50 and ResNet101 can accurately identify gender with a high level of success. These two models achieve the same level of accuracy, namely 99.67%, with a recall level of 0.933 and a precision level of 1. The main difference between the two

TABLE 2. Resnet 50.

Learning Rate	Training Loss	Training Accuracy (%)	Validation Accuracy (%)	Validation Loss
Lr = 10^{-3}	0.0177	99.67	99.33	0.67
Lr = 10^{-4}	0.0028	99.67	99.21	0.79
Lr = 10^{-5}	0.023	99.67	99.09	0.91

TABLE 3. Resnet 101.

Learning Rate	Training Loss	Training Accuracy (%)	Validation Accuracy (%)	Validation Loss
Lr = 10^{-3}	0.0177	99.82	99.33	0.67
Lr = 10^{-4}	0.0018	99.82	99.43	0.57
Lr = 10^{-5}	0.34	99.86	99.33	0.67

TABLE 4. Evaluation results of proposed resnet models.

Resnet architecture	Accuracy (%)	Recall	Precision	F1 Score (%)
Resnet50	99.67	0.933	1.00	99.65
Resnet101	99.67	0.933	1.00	99.65

TABLE 5. Comparison of proposed and conventional models.

Model	Validation Loss	Validation Accuracy (%)
Proposed ResNet50	**0.670**	**99.33**
Proposed ResNet101	**0.670**	**99.43**
Conventional ResNet 50 [33]	0.200	98.57
Conventional ResNet 34 [33]	0.250	97.94
Conventional CNN [26]	7.358	90.08

models lies in the time required for the training process. ResNet50 requires 2 hours of training time, while ResNet101 takes longer. Apart from the high level of accuracy, analysis of the learning curve and test results indicate that the performance of the two models, ResNet50 and ResNet101, can be considered well-fitted and high-performing. The results of this evaluation show that both methods have equivalent performance in detecting gender.

The validation loss and accuracy achieved in this research are compared with those from previous related studies in Table 5. In terms of accuracy, the CNN model in this study demonstrated significantly higher accuracy compared to previous research. The validation loss is quite similar overall. These results indicate that the model in this study effectively detects gender. The same validation loss between two different iterations, in this case, at epochs 50 and 101, may occur for several reasons. One possibility is that the two iterations have identical model architectures, the same weight initialization, and unchanged validation datasets. In this case, if the model has reached a state of convergence by the 50th iteration, then its performance may not change significantly by the 101st iteration. A low validation loss is desirable, ideally close to zero. However, achieving absolute zero may be challenging due to the presence of inherent noise in the data. Good validation accuracy depends on the type of problem being solved. For instance, a model with validation accuracy exceeding 90% would be deemed satisfactory in a binary classification task.

C. PROTOTYPING OF PROPOSED MODEL FOR SMART HOME

Based on the implementation model explained in Chapter IV, the prototype of the voice frequency-based gender detection model has been developed to recognize and distinguish voice characteristics. This prototype comprises smart home devices, a web-based interface capable of recognizing gender voices using ResNet50, and communication through IoT protocols. The developed prototype was tested on several smart home devices including moskiller, smart fan, smart humidifier, smart lamp, etc. A functional test of these, i.e. a smart lamp, is shown in Figure 12, it can work well and can recognize and detect voice commands of both men and women. In addition to gender, this voice detection test is also performed for people of different ages on each smart device. Various devices have been tested successfully and are used in our proposed implementation model for a smart home system. This can enhance the smart home experience by identifying the occupants of the home who are speaking or interacting with the system. This enables customization of preferences and devices based on the active user. The working principle of the prototype is that at the start of use, users need to register or log in to the smart home portal. After successfully logging in, the user must connect the electronic equipment to be controlled to the WiFi network. In the trial, we tested the use of two devices, a mosquito trap, and a lamp, with two users. At the start of testing, a virtual assistant will greet users to receive commands via voice messages.

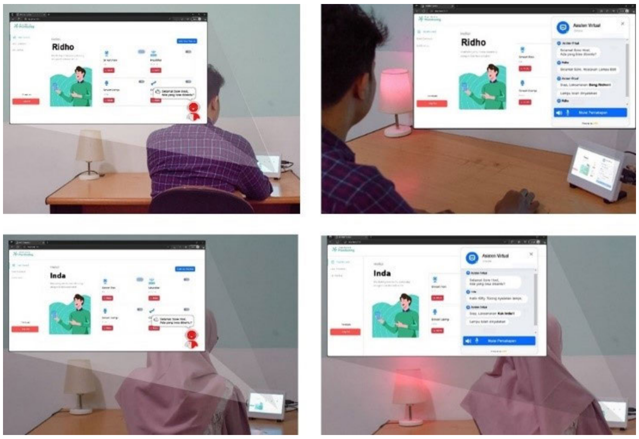


FIGURE 12. Prototype testing gender-based detection for smart lamp.

When the system receives a male voice message, it will identify the gender and greet the user with a male-appropriate address (‘Brother’ or ‘Sir’), based on age analysis conducted during registration. Then, the system connects with the user's name and work order. The term "Brother" is a way of acknowledging male voice frequencies by a deep learning system.

Likewise, if the voice message received is from a woman, the system will recognize her and greet her with a female address ('Sis' or 'Madam'), depending on the age analysis during registration. Then, the system connects with the user's name and work order. The term "Sis" is used by a deep learning system to recognize female voice frequencies. In this conversation, the user refers to the virtual assistant as Kitty because there is an option to assign a name to the assistant. Users can assign any name to the Virtual Assistant system, which is then stored in the Elastic File System (EFS). The operation of the lamp through the prototype, as shown in Figure 12, is facilitated by communication between the virtual assistant and the user. Initially, the virtual assistant greets the user, following which the user issues the command to turn on the lamp. Deep learning will process the command to recognize male or female voices that have been registered in the system, and then the lights will be turned on. The lamp can also be personalized for both men and women with the desired color and ambiance.

In our proposed smart home model, various smart devices have been used to measure the system's performance in terms of latency. Latency is a substantial performance aspect to consider when designing smart home systems and implementing smart home devices. Measurements are performed 30 times in real-time for each smart device. The measurement results show that the average latency of each device is different, as shown in Figure 13.

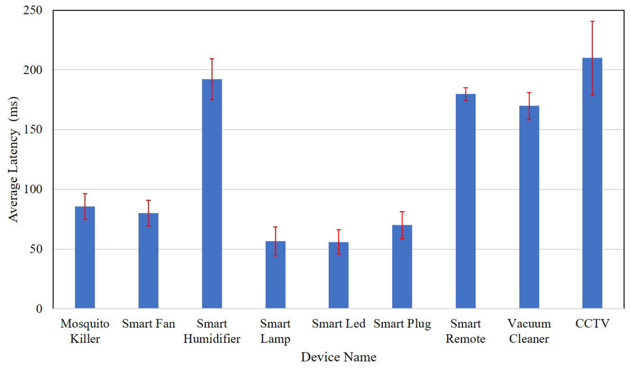


FIGURE 13. Average latency measurement of the implemented prototype model on various smart devices for smart home systems.

The average latency of Smart LED is the lowest, about 56 ms. It simply receives commands to turn on, off, or change brightness, allowing for quick response. So, its functionality is simpler than other devices like CCTV cameras or others. Smart LED is specially designed for quick response to commands and easy data processing. On the other hand, CCTV has the highest average latency of 210 ms. CCTV cameras record video in a relatively large format and depend on the resolution used. Therefore, sending the data to the server takes longer, which increases the delay. Network congestion can also result in increased CCTV latency. These average latencies are also lower than the experimental results of a previous study [42] without deep learning. This shows that deep learning can provide low latency in smart homes.

TABLE 6. The standard deviation of the latency of smart devices.

Device Name	The standard deviation of latency (ms)
Mosquito Killer	10.70
Smart Fan	10.83
Smart Humidifier	17.15
Smart Lamp	12.09
Smart Led	10.23
Smart Plug	11.26
Smart Remote	5.52
Vacuum Cleaner	11.22
CCTV	30.88

The measurement results show that data communication latency differs for each device. To understand the latency consistency level, the measurement data's standard deviation is calculated, which shows how much the average latency varies in the 30 experiments performed as shown in Table 6. The standard deviation value of the smart remote is the smallest, indicating that the latency data tends to be close to the average latency. The standard deviation value of CCTV latency indicates that the latency data is spread over a range of values greater than the average latency.

Based on the measurement and analysis results, the latency can be affected by several factors, such as the unstable environmental conditions that cause latency, and the size of the data transmitted through the prototype, which causes different latency, data transmission distance, and the impact of hardware such as a smart humidity controller is modified by adding a microcontroller and a relay. The smart humidity developed in this paper focuses on the convenience of an air purifier, where relays activate and deactivate the device based on user commands. The smart lamp uses last-will messaging technology (LWT) to store and transmit client states to MQTT clients. The sensor is then connected to an MQTT broker to maintain a connection and pass messages.

The prototype can enhance the smart home experience by identifying the occupants of the home who are speaking or interacting with the system. This enables customization of preferences and devices based on the active user. Furthermore, utilizing information about the user's gender can enhance the smart home experience by providing a more personalized touch. In addition, the smart home can optimize energy usage based on who is in the home at the time, such as lighting or room temperature settings that are tailored to the preferences of the occupants. Thus, implementing this model can make the smart home more intelligent, responsive, safe, and efficient.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a novel approach to gender classification in smart homes, leveraging voice frequency-based features and convolutional neural networks (CNN) to achieve exceptional accuracy and performance. This paper introduces a voice-to-image conversion algorithm that will be utilized in two deep-learning architecture models: ResNet50 and ResNet101. Based on the training and testing results, the two proposed ResNet50 and ResNet101 architectures are highly effective in gender detection in smart homes. The implementation of the proposed ResNet50 and ResNet101 architectures has yielded remarkable results, achieving a high accuracy rate of 99.66%. The model also demonstrated outstanding recall, precision, and F1 score, surpassing conventional methods and showcasing the effectiveness of deep learning in smart home applications. This paper has also developed a prototype that utilizes ResNet50 on smart devices within a smart home. Prototype testing results show that these smart devices have detected male and female voices. So, the proposed model has great potential for application in smart home systems. Prototype testing of the ResNet50 model on smart devices confirms its practical viability. It successfully identifies male and female voices, showcasing the potential for real-world deployment and integration into smart home systems.

Further research could focus on refining the integration of deep learning models with smart home technology to enhance automation capabilities, particularly in creating personalized experiences tailored to individual residents based on gender recognition. Exploring advanced algorithms to enhance security measures and energy efficiency in smart homes offers a promising path to future development. By using deep learning techniques, more efficient and sustainable solutions can be developed. Collaboration across disciplines and industries can foster smart home technology innovation, driving advancements that prioritize user experience, sustainability, and efficiency. This collaboration ultimately helps realize the vision of intelligent, interconnected homes.

ACKNOWLEDGMENT

The authors thank the research partner of the Department of Industry and Trade, Aceh Province.

REFERENCES

[1] D. Pal, T. Triyason, S. Funilkul, and W. Chutimaskul, "Smart homes and quality of life for the elderly: Perspective of competing models," *IEEE Access*, vol. 6, pp. 8109–8122, 2018, doi: 10.1109/ACCESS.2018.2798614.

[2] A. Zielonka, M. Wozniak, S. Garg, G. Kaddoum, M. J. Piran, and G. Muhammad, "Smart homes: How much will they support us? A research on recent trends and advances," *IEEE Access*, vol. 9, pp. 26388–26419, 2021, doi: 10.1109/ACCESS.2021.3054575.

[3] J. Yu, A. de Antonio, and E. Villalba-Mora, "Deep learning (CNN, RNN) applications for smart homes: A systematic review," *Computers*, vol. 11, no. 2, p. 26, Feb. 2022, doi: 10.3390/computers11020026.

[4] L. Y. Rock, F. P. Tajudeen, and Y. W. Chung, "Usage and impact of the Internet-of-things-based smart home technology: A quality-of-life perspective," *Universal Access Inf. Soc.*, vol. 23, no. 1, pp. 345–364, Nov. 2022, doi: 10.1007/s10209-022-00937-0.

[5] M. Umer, S. Sadiq, R. M. Alhebshi, M. F. Sabir, S. Alsubai, A. A. Hejaili, M. M. Khayyat, A. A. Eshmawi, and A. Mohamed, "IoT based smart home automation using blockchain and deep learning models," *PeerJ Comput. Sci.*, vol. 9, p. e1332, May 2023, doi: 10.7717/peerj-cs.1332.

[6] M. Nadji-Tehrani and A. Eslami, "A brain-inspired framework for evolutionary artificial general intelligence," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5257–5271, Dec. 2020, doi: 10.1109/TNNLS.2020.2965567.

[7] W. Han, Z. Tian, C. Zhu, Z. Huang, Y. Jia, and M. Guizani, "A topic representation model for online social networks based on hybrid human-artificial intelligence," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 1, pp. 191–200, Feb. 2021, doi: 10.1109/TCSS.2019.2959826.

[8] M. K. Muchamad, Z. Fuadi, and N. Nasaruddin, "Prototype design of deep learning-based voice control model for smart home," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2022, pp. 25–30, doi: 10.1109/IoTaIS56727.2022.9975901.

[9] E. Viglianisi, M. Ceccato, and P. Tonella, "A federated society of bots for smart contract testing," *J. Syst. Softw.*, vol. 168, Oct. 2020, Art. no. 110647, doi: 10.1016/j.jss.2020.110647.

[10] M. Wyrich and J. Bogner, "Towards an autonomous bot for automatic source code refactoring," in *Proc. IEEE/ACM 1st Int. Workshop Bots Softw. Eng.*, May 2019, pp. 24–28, doi: 10.1109/BOTSE.2019.00015.

[11] T. J. Saleem and M. A. Chishti, "Deep learning for the Internet of Things: Potential benefits and use-cases," *Digit. Commun. Netw.*, vol. 7, no. 4, pp. 526–542, Nov. 2021, doi: 10.1016/j.dcan.2020.12.002.

[12] H. Garg and M. Dave, "Securing IoT devices and SecurelyConnecting the dots using REST API and middleware," in *Proc. 4th Int. Conf. Internet Things, Smart Innov. Usages (IoT-SIU)*, Apr. 2019, pp. 1–6, doi: 10.1109/IoT-SIU.2019.8777334.

[13] D. Vlaj and A. Zgank, "Acoustic gender and age classification as an aid to human–computer interaction in a smart home environment," *Mathematics*, vol. 11, no. 1, p. 169, Dec. 2022, doi: 10.3390/math11010169.

[14] C. Lucia, G. Zhiwei, and N. Michele, "Biometrics for Industry 4.0: A survey of recent applications," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 11239–11261, Aug. 2023, doi: 10.1007/s12652-023-04632-7.

[15] A. Hosovsky, J. Pitel, M. Trojanova, and K. Zidek, "Computational intelligence in the context of Industry 4.0," in *Implementing Industry 4.0 in SMEs: Concepts, Examples and Applications*. Cham, Switzerland: Springer, 2021, pp. 27–94.

[16] M. A. Uddin, R. K. Pathan, M. S. Hossain, and M. Biswas, "Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN," *J. Inf. Telecommun.*, vol. 6, no. 1, pp. 27–42, Jan. 2022, doi: 10.1080/24751839.2021.1983318.

[17] B. B. Monson, A. J. Lotto, and B. H. Story, "Gender and vocal production mode discrimination using the high frequencies for speech and singing," *Frontiers Psychol.*, vol. 5, pp. 1–23, Oct. 2014.

[18] A. Tursunov, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors*, vol. 21, no. 17, p. 5892, Sep. 2021, doi: 10.3390/s21175892.

[19] S. Jadav, "Voice-based gender identification using machine learning," in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Dec. 2018, pp. 1–4, doi: 10.1109/CCAA.2018.8777582.

[20] V. S. Kone, A. Anagal, S. Anegundi, P. Jadhav, U. Kulkarni, and S. M. Meena, "Voice-based gender and age recognition system," in *Proc. Int. Conf. Advancement Comput. Comput. Technol. (InCACCT)*, May 2023, pp. 74–80, doi: 10.1109/InCACCT57535.2023.10141801.

[21] G. Richard, P. Smaragdis, S. Gannot, P. A. Naylor, S. Makino, W. Kellermann, and A. Sugiyama, "Audio signal processing in the 21st century: The important outcomes of the past 25 years," *IEEE Signal Process. Mag.*, vol. 40, no. 5, pp. 12–26, Jul. 2023, doi: 10.1109/MSP.2023.3276171.

[22] H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019, doi: 10.1109/JSTSP.2019.2908700.

[23] M. Almutairi, L. A. Gabralla, S. Abubakar, and H. Chiroma, "Detecting elderly behaviors based on deep learning for healthcare: Recent advances, methods, real-world applications and challenges," *IEEE Access*, vol. 10, pp. 69802–69821, 2022, doi: 10.1109/ACCESS.2022.3186701.

[24] M. Mohamed, A. El-Kilany, and N. El-Tazi, "Future activities prediction framework in smart homes environment," *IEEE Access*, vol. 10, pp. 85154–85169, 2022, doi: 10.1109/ACCESS.2022.3197618.

[25] L. Filipe, R. S. Peres, and R. M. Tavares, "Voice-activated smart home controller using machine learning," *IEEE Access*, vol. 9, pp. 66852–66863, 2021, doi: 10.1109/ACCESS.2021.3076750.

[26] A. V. Kuchebo, V. V. Bazanov, I. Kondratev, and A. M. Kataeva, "Convolution neural network efficiency research in gender and age classification from speech," in *Proc. IEEE Conf. Russian Young Researchers Electr. Electron. Eng.*, Jan. 2021, pp. 2145–2149, doi: 10.1109/ElConRus51938.2021.9396365.

[27] M. Yani, S. S. M. T. B. Irawan, and S. T. M. T. C. Setiningsih, "Application of transfer learning using convolutional neural network method for early detection of Terry's nail," *J. Phys. Conf. Ser.*, vol. 1201, no. 1, May 2019, Art. no. 012052, doi: 10.1088/1742-6596/1201/1/012052.

[28] Y. H. Tsai, N. Y. Lyu, S. Y. Jung, K. H. Chang, J. Y. Chang, and C. T. Sun, "Deep learning based AOI system with equivalent convolutional layers transformed from fully connected layers," in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jul. 2019, pp. 103–107, doi: 10.1109/AIM.2019.8868602.

[29] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[30] M. Varsha and C. S. Nair, "Indian sign language gesture recognition using deep convolutional neural network," in *Proc. 8th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jul. 2021, pp. 193–197, doi: 10.1109/ICSCC51209.2021.9528246.

[31] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer," *Proc. Comput. Sci.*, vol. 179, pp. 423–431, Jan. 2021, doi: 10.1016/j.procs.2021.01.025.

[32] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.

[33] A. A. Alnuaim, M. Zakariah, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, "Speaker gender recognition based on deep neural networks and ResNet50," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–13, Mar. 2022, doi: 10.1155/2022/4444388.

[34] T. Vaiyapuri, E. L. Lydia, M. Y. Sikkandar, V. G. Díaz, I. V. Pustokhina, and D. A. Pustokhin, "Internet of Things and deep learning enabled elderly fall detection model for smart homecare," *IEEE Access*, vol. 9, pp. 113879–113888, 2021, doi: 10.1109/ACCESS.2021.3094243.

[35] S. Showkat and S. Qureshi, "Efficacy of transfer learning-based ResNet models in chest X-ray image classification for detecting COVID-19 pneumonia," *Chemometric Intell. Lab. Syst.*, vol. 224, May 2022, Art. no. 104534, doi: 10.1016/j.chemolab.2022.104534.

[36] L. Ma, R. Shuai, X. Ran, W. Liu, and C. Ye, "Combining DC-GAN with ResNet for blood cell image classification," *Med. Biol. Eng. Comput.*, vol. 58, no. 6, pp. 1251–1264, Jun. 2020, doi: 10.1007/s11517-020-02163-3.

[37] A. Mahajan and S. Chaudhary, "Categorical image classification based on representational deep network (RESNET)," in *Proc. 3rd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Jun. 2019, pp. 327–330, doi: 10.1109/ICECA.2019.8822133.

[38] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021, doi: 10.1109/TGRS.2020.3043267.

[39] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018, *arXiv:1811.03378*.

[40] M. B. Hossain, S. M. H. S. Iqbal, M. M. Islam, M. N. Akhtar, and I. H. Sarker, "Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images," *Informat. Med. Unlocked*, vol. 30, Jan. 2022, Art. no. 100916, doi: 10.1016/j.imu.2022.100916.

[41] M. Noura, S. Heil, and M. Gaedke, "Natural language goal understanding for smart home environments," in *Proc. 10th Int. Conf. Internet Things*, 2020, pp. 1–25, doi: 10.1145/3410992.3410996.

[42] I. S. Areni, A. Waridi, I. Amirullah, C. Yohannes, A. Lawi, and A. Bustamin, "IoT-based of automatic electrical appliance for smart home," *Int. J. Interact. Mobile Technol. (iJIM)*, vol. 14, no. 18, p. 204, Nov. 2020, doi: 10.3991/ijim.v14i18.15649.

NASARUDDIN NASARUDDIN (Member, IEEE) received the B.Eng. degree in electrical engineering from the Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, in 1997, and the M.Eng. and D.Eng. degrees in physical electronics and informatics from the Graduate School of Engineering, Osaka City University, Japan, in 2006 and 2009, respectively. He is currently a Full Professor with the Electrical Engineering Department, Universitas Syiah Kuala. He is also the Head of the Distribution System and High-Performance Computing Laboratory, Universitas Syiah Kuala. He has published numerous articles on cooperative communication networks, communication technologies, and applications of deep learning. His research interests include digital communications, information theory, and deep learning applications, in addition to computer and communication networks. He is a member of the IEEE Systems, Man, and Cybernetics Society and ACM.

MUHAMMAD AGUNG P. PRATAMA TRESMA was born in Kotabangun, Pekanbaru, Riau, Indonesia, in 2001. He received the B.Eng. degree in computer engineering from Universitas Syiah Kuala, Banda Aceh, in 2023. From 2021 to 2023, he held the role of a Laboratory Assistant of the Python basic programming course. This role aims to share his knowledge with other students, help them understand basic programming concepts, and develop their skills in the Python programming language. In addition, he is also involved in valuable community activities. In 2022, he was active in the student program, an initiative that aims to empower the community through various service activities. One of the successful projects in this program is the implementation of the Internet of Things (IoT) to control the watering of chili plants. By using the IoT technology, he was developing an innovative solution that enables automatic, efficient, and sustainable control of plant watering.

MASDUKI KHAMDAN MUCHAMAD received the degree in informatic engineering from Universitas Dian Nuswantoro, Semarang, in 2016, and the master's degree in software engineering and intelligence from the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia, in 2017. He started his career as a Programmer with Indonesian Embassy, Kuala Lumpur, in 2017. In July 2019, he joined Universitas Syiah Kuala, Banda Aceh, and was placed at the Computer Engineering Department. His research interests include the development of intelligent agents, natural language processing, and speech technology to facilitate graceful human–computer interactions, conversational robot, and web services.

ZAHRUL FUADI (Member, IEEE) received the B.Eng. degree in mechanical engineering from the Mechanical Engineering Department, Universitas Syiah Kuala, Banda Aceh, Indonesia, in 1996, the M.Sc. degree in mechanical engineering from the School of Mechanical Engineering, Universiti Sains Malaysia, in 2004, and the Ph.D. degree from the Tribology Laboratory, Tohoku University, Japan, in 2009. Subsequently, he was a Postdoctoral Research Fellow with the Institute of Fluid Science, Tohoku University, and the Nanointerface Engineering Laboratory, Graduate School of Engineering, Tohoku University. He is currently a Full Professor with the Mechanical and Industrial Engineering Department, Universitas Syiah Kuala, where he is the Head of the Mechanical Construction and Design Laboratory. His research interests include brake noises, friction sound, surface texture, tribological and dynamic behavior of soft material surfaces, friction and wear of carbon-based coatings, and tribology of renewable lubricants. In 2005, he was awarded the "Monbukagakusho" Scholarship from Japanese Government for the Ph.D. degree.