Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.043176

ARTICLE

Recognition of Human Actions through Speech or Voice Using Machine Learning Techniques

Oscar Peña-Cáceres[1,2,*], Henry Silva-Marchan[3], Manuela Albert[4] and Miriam Gil[1]

[1]Professional School of Systems Engineering, Universidad César Vallejo, Piura, 20009, Perú

[2]Escola Tècnica Superior d'Enginyeria, Departament d'Informàtica, Universitat de València, Burjassot, Valencia, 46100, Spain

[3]Department of Mathematics, Statistics and Informatics, Universidad Nacional de Tumbes, Tumbes, 24000, Perú

[4]Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Valencia, 46022, Spain

*Corresponding Author: Oscar Peña-Cáceres. Email: ojpenac@ucvvirtual.edu.pe

Received: 24 June 2023 Accepted: 11 September 2023 Published: 29 November 2023

ABSTRACT

The development of artificial intelligence (AI) and smart home technologies has driven the need for speech recognition-based solutions. This demand stems from the quest for more intuitive and natural interaction between users and smart devices in their homes. Speech recognition allows users to control devices and perform everyday actions through spoken commands, eliminating the need for physical interfaces or touch screens and enabling specific tasks such as turning on or off the light, heating, or lowering the blinds. The purpose of this study is to develop a speech-based classification model for recognizing human actions in the smart home. It seeks to demonstrate the effectiveness and feasibility of using machine learning techniques in predicting categories, subcategories, and actions from sentences. A dataset labeled with relevant information about categories, subcategories, and actions related to human actions in the smart home is used. The methodology uses machine learning techniques implemented in Python, extracting features using CountVectorizer to convert sentences into numerical representations. The results show that the classification model is able to accurately predict categories, subcategories, and actions based on sentences, with 82.99% accuracy for category, 76.19% accuracy for subcategory, and 90.28% accuracy for action. The study concludes that using machine learning techniques is effective for recognizing and classifying human actions in the smart home, supporting its feasibility in various scenarios and opening new possibilities for advanced natural language processing systems in the field of AI and smart homes.

KEYWORDS

AI; machine learning; smart home; human action recognition

1 Introduction

In recent years, speech recognition of human actions through artificial intelligence has emerged as a promising and rapidly growing field of research. This area focuses on the development of techniques and algorithms that enable machines to understand and recognize the actions humans perform by using only auditory information captured through recording devices. Speech recognition of human actions has applications in various domains, such as virtual assistants, smart homes, security systems, human-machine interaction [1], and healthcare. As AI evolves and improves its ability to understand natural language, speech recognition of human actions has become even more relevant and challenging. Human speech action recognition research has benefited greatly from advances in audio signal processing and machine learning [2]. Combining audio signal feature extraction techniques and machine learning algorithms has enabled the development of more accurate and robust human speech action recognition models. These models can analyze acoustic and linguistic patterns in speech data to identify the actions users perform. However, most of these systems are expensive and require a complete replacement of existing equipment? [3]. Nowadays, people are eager to use and buy devices that make most tasks easier. According to Google, 27% of the world's online population uses voice search on mobile [4], which has promoted AI to be considered one of the most exciting fields in the new world with a continuous activity to produce machines endowed with intelligence that can help simplify many tasks.

A voice assistant is considered an AI application that can be deployed in various locations. The emergence of the Internet of Things (IoT) has made maximizing projects with minimal architectures and reasonable costs possible. Using the IoT protocol, machines can communicate with humans and other machines, making it possible to control them remotely. Many applications focused on the smart home environment communicate through the cloud. This is done to simplify some basic tasks, such as when a person is in the room and uses only his/her voice to perform some action. This kind of activity greatly contributes to many people, especially elderly people, people with determination, or pregnant women. In addition, a person with an injured leg or who is carrying a heavy load can use their voice to control some elements in the room without the help of other people.

The development of autonomous systems in the smart home domain has begun to be attracted by the integration of a module based on voice commands as a home automation solution for the user to transmit voice prompts. This module continuously listens to and processes environmental sounds to detect words. Once the user pronounces the word, the voice module activates real-time voice command processing, where it recognizes and performs the action expressed by the user. In some cases, the operations performed may lead to vocal responses or text messages [5]. Undoubtedly, there are offline speech recognition engines for this purpose [6] that could help in scenarios where the connectivity service is lacking.

To strengthen this type of study, it is fundamental to recognize that human beings can express their emotions in different ways depending on the occasion, which could generate some type of uncertainty or distortion at the moment of understanding what is expressed, and the system ends up performing other types of actions. Studies such as [7] pointed out that in some cases, it is difficult to detect and understand users' emotions because it depends on the context in which it is found. Among the best practices to look for the quality of this type of system is to develop machine learning models to classify human emotions using their speeches or requests previously made. Emotions can be classified into five categories: normal emotion, anger, surprise, happiness, and sadness. Other findings, such as [8], pointed out that it is possible to use powerful packages such as pywhatkit, pyttsx3, pygame, and OpenCV-based speech recognition.

A virtual or personal assistant-based AI could be a package of intelligent mechanisms to perform different tasks and services based on user queries. Smart technologies, such as natural language processing, lead to voice recognition and play an essential role in security and other areas. The most common mechanisms to achieve this type of results are associated with machine-deep learning and adhering to natural language. On the other hand, human-robot interaction is one of the areas of knowledge that relies on multidisciplinary technologies such as natural language processing and machine learning. Such tools can be integrated into various alternative intelligent systems. Voice can be much more economical than typing on a keyboard. Recent research [9] indicated that augmented reality is becoming an expanding field in research and practice due to its ability to link virtual information with the real world. Significant advances in wearable devices or smartphone technology have drawn people's attention to augmented reality and have led to the rapid development of applications, games, and environments where daily human activities can be executed [10]. Voice commands and speech technology have also been significant milestones of the 21st century by enabling human interaction with devices and applications through speech recognition. It provides users with interface solutions that do not require physically touching or pressing buttons to perform a given action.

This study is distinguished by its focus on recognizing human actions through artificial intelligence and speech processing. In a context where natural human-device interaction gains relevance, this research addresses crucial challenges in understanding vocally expressed actions that could act in practical scenarios such as virtual assistance, intelligent homes, security, and medical care. The combination of advances in audio signaling and machine learning algorithms promises the creation of increasingly accurate and robust models. The vision for this solution is for it to be deployable on both every day and inclusive devices, which extends its utility and enhances the user experience. This research could help simplify tasks and improve accessibility, advancing the use of voice as an intuitive and powerful interface in the digital environment. Its relevance lies in addressing emerging challenges in the human-IA interface and improving everyday life.

Technology has advanced beyond our wildest dreams in recent years. It helps human beings in their daily lives, and there is still much to learn about it in the future. Some problems, such as disabled people and the elderly who cannot enjoy the freedom of driving alone, have yet to be solved. An autonomous system in the smart home domain with a voice module can be helpful in solving this issue. The interaction and alternative required for this environment to operate autonomously is a voice command through a smartphone application [11]. The objective of this study is to use machine learning techniques in the smart home domain, where the user expresses terms or sentences through voice or speech that lead to the identification and recognition of patterns that infer the action requested by the user. The proposal is expected to be used in other knowledge domains such as education, health, security, and food.

2 Related Work

In recent times, significant advances have been achieved in the field of automated speech recognition through the application of various machine learning-based approaches. In this section, we describe some of the current state-of-the-art strategies, with special emphasis on the various techniques employed in machine learning. According to [12], controlling electronic devices remotely using voice and brain waves is one of the future techniques that can be used to control household appliances or wheelchairs remotely. This type of system is intended to be used by everyone. Its purpose is to assist and provide help for the purpose of meeting the needs of elderly and disabled people. Reference [13] indicated that recent technological advances allow the development of voice-machine interfaces, systems that can be used in a variety of potential applications, such as security, smart homes, and centers for people with disabilities. The proposal focused on a frequency domain speech recognition platform, where a spectral analysis of speech to extract its characteristics was performed. These extracted characteristics were then used to identify and perform actions such as issuing action commands, granting access to secure services, voice dialing, telephone banking, and accessing confidential databases. Reference [14] made known that exploring smart home environments, leads to generating knowledge where a system can make intelligent decisions and control end devices based on the current resident by voice.

Works such as [3] pointed out that these types of system can run on an Android operating system phone connected by Bluetooth to a local home automation node and depending on the connected devices, the node searches for the keywords in the command and performs a control action. This framework is very useful for the needs of elderly and disabled patients due to their minimal technical knowledge requirements. In [15], the authors used natural language processing techniques to build a system based on the Internet of Things with the purpose of making home devices easier to use and control. Through a command or question, the system understands the user's wishes and responds accordingly. Accompanying this study, the authors in [16] indicated that automatic speech recognition is an effective technique that can convert human speech into text format or computer actions. Signal processing and machine learning techniques are required to recognize speech.

In [17], the authors addressed the improvement of the voice assistant in the smart home by providing context-aware capabilities. Currently, voice assistants can only receive clear voice commands from users without being able to meet individual needs automatically. To address this limitation, a system that uses walking sounds to identify the user and provide personalized services automatically was proposed. The system recognizes footstep sounds and detects the presence of a user through machine learning algorithms. Also, in [18], the authors designed a voice assistant to command robotic tasks in residential environments aimed at people with mobility limitations. In this case, a convolutional neural network trained with a database of 3600 audios of keywords, such as paper, glass, or robot, was used, achieving an accuracy of 96.9% in the discrimination of categories. The results reached an accuracy of 88.75% in identifying eight different actions by voice commands, such as robot brings to glass. Reference [2] presented a speech and gesture signal translator focused on human-machine interaction. By using machine learning algorithms, they achieve 97% accuracy in speech recognition, even in noisy environments. These results show the power of machine learning algorithms and the potential that comes with creating high-fidelity platforms for effective human-machine interaction.

Reference [19] described that human-machine interfaces with voice support are essential in improving user interfaces. This work focused on employing deep learning for intelligent human-computer interaction. The results show that the combination of human-computer interaction and deep learning is deeply applied in gesture recognition, speech recognition, emotion recognition, and intelligent robot steering. On the other hand, approaches based on machine learning techniques and their integration into robotics have shown it is possible to generate scenario-based programming of voice-controlled medical robotic systems, such as [20] that reveals that human-machine voice communication is the subject of research in various fields (industry, social robotics). Having reasonable results on voice-controlled system functions can lead to a significant improvement in human-machine collaboration.

In the smart home domain, voice-based control devices have become a topic of great interest to major technology companies and researchers. The study introduced in [21] proposed a solution for detecting Vietnamese language speech because there is no current research in this field, and many previous solutions are needed to provide adequate scalability for the future. Using Machine Learning algorithms, the authors achieve an average recognition accuracy of 98.19% on 15 commands to control smart home devices. Solutions such as [22] pinpointed the possibility of early detection of Parkinson's disease through voice recordings in the smart home environment using machine learning prediction methods and metaheuristic algorithms. According to Hung et al. in [23], the recognition of emotions through speech can be achieved using machine learning and deep learning techniques. While Tanveer et al. in [24] reaffirmed that machine learning methods are widely used to process and analyze speech signals due to their performance improvements in multiple domains. Deep learning and ensemble learning are the two most commonly used techniques, which provide benchmark performance in different downstream tasks.

These types of studies provide insight into the potential and effectiveness of machine learning in the automation and intelligent control of actions performed by users in smart home environments. The application of different strategies demonstrates the flexibility of machine learning to adapt to various situations and challenges in recognizing human actions. These testimonials are the potential of machine learning. As technology advances, these advances are expected to continue to improve, providing new opportunities for creating more intuitive and personalized homes.

3 Methodology

This section describes the procedures used to develop the module for the recognition of human actions through speech or voice using automatic learning in the smart home domain. Fig. 1 outlines the workflow from the stage of understanding the problem to the refinement of the proposal.

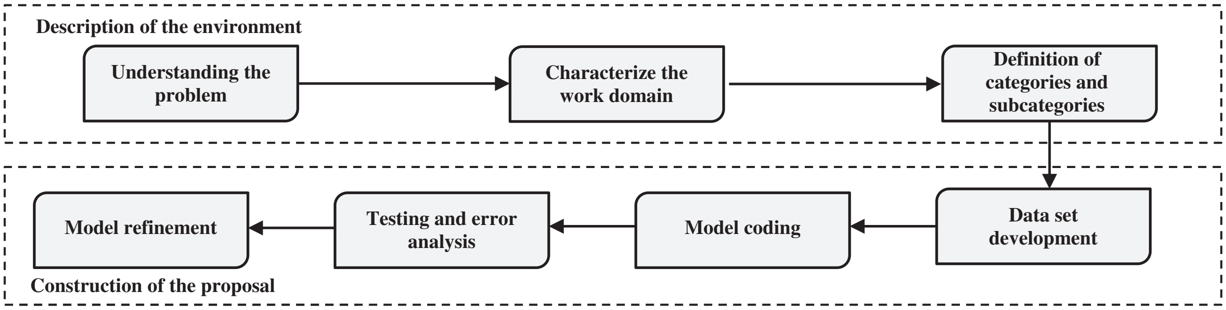

Description of the environment
Understanding the problem
Characterize the work domain
Definition of categories and subcategories
Data set development
Model coding
Testing and error analysis
Model refinement
Construction of the proposal

Figure 1: Methodological strategy of the proposal

3.1 Understanding the Problem

As technological services have evolved, human beings have learned to adapt and take advantage of the benefits offered by these digital media. However, a fragment of the population still distrusts and is not empathetic with this type of solution. These hesitations are due to the fact that many technological proposals leave aside the main actor, which in this case is characterized as the user. AI and cloud services based on IoT have sought to make disruptive changes in people's daily activities. For example, in recent years, special attention has been paid to the voice remote control of driverless vehicles for the future intelligent transportation system. In these vehicles, the remote intelligent car receives the user's commands and controls the engine rotation forward, backward, turn left or right, and stop [25]. Attempting to overcome the user's abilities still represents a significant gap and therefore the community must reformulate the principle of an autonomous system because such a system will always depend on textual, verbal, or gestural indications to perform some action.

It is necessary to address more empathetic solutions where the user trusts the technologies in their mode, form, and quality. A good start to promote autonomous solutions is to develop projects in the smart home environment. The ease of use and simplicity of a system are factors that build the safety of users during the stage of interaction with such systems, whose purpose is to assist in the development of everyday tasks such as turning on or off the light, opening the blinds, and turning on the air conditioning, among other household activities. Other recent findings used augmented reality with voice recognition capabilities where a student and teacher [26] performed joint tasks that complement the student's knowledge [27].

It is common for people in advanced stages of life to face a greater propensity to develop chronic diseases and to experience a decline in their physical and cognitive abilities. These factors can trigger a progressive decline in their ability to lead an autonomous and independent life. In this context, voice recognition of actions in smart home systems emerges as an innovative technological intervention that can make a significant contribution. These systems allow older adults to control various functions and devices in their homes simply by voice commands, providing greater independence and autonomy in their daily lives. The recognition of voice actions in smart home systems is positioned as an innovative and adapted intervention to improve the quality of life of older adults. By allowing them to control their environment and access support services in a simple and natural way, this technology helps to strengthen their independence and general well-being. These approaches could help the user to have another look at autonomous solutions based on voice or speech recognition that contribute to domestic, academic, or work activities that improve the quality of life of the human being.

On the other hand, this study is based on the realization that, despite technological advances, there remains a huge gap between human capabilities and the expectations of autonomous systems. By enabling more natural interactions with technology, speech recognition technologies hold promise for bridging this gap. Furthermore, by improving the accuracy and efficiency of recognizing human actions through speech, machine learning research has the potential to improve people's daily lives. It is clear that we need technical solutions that are more comprehensive and flexible. From this perspective, research on human action recognition through speech is justified as a means to meet the growing demand for inclusive and accessible technology that improves the quality of life of users in different domains.

3.2 Characterize the Work Domain

In a smart home environment, a variety of activities can be performed through voice command, providing convenience and control without the need to interact directly with devices. Virtual assistants, such as Amazon's Alexa or Google Assistant, provide facilities for users to control home lighting, adjust room temperature, play music, open and close curtains or blinds, turn on and off appliances such as the coffee maker or vacuum cleaner, and manage home security such as activating and deactivating alarms and surveillance cameras [28]. Other more complex tasks that can be performed by these types of systems focus on programming customized routines, such as simulating presence at home during vacations by turning lights on and off and playing ambient sounds. It is also possible to perform online information queries to obtain weather forecasts, the latest news, or cooking recipes, as mentioned.

3.3 Definition of Categories, Subcategories and Actions

This section sets out the categories, subcategories, and actions that the speech recognition module must infer in order to perform the actions. Table 1 describes them. Categories are characterized as the main elements of the smart home environment and subcategories as the elements that are associated with the categories or represent compatibility in their operation. The actions are based on the speech or voice terms used by the user. For example, "Illuminate the kitchen today", in this case, the system must recognize the category, subcategory, and action to be performed to fulfill the user's request which corresponds to turning on the lights right now. Another situation could be, "It is cold in the library", in this case, the user needs to be cold in the library environment, the system should infer the context and activate the heating. This system could collaborate with elderly people, who develop multiple chronic diseases and experience a decline in some of their physical and cognitive functions, leading to a decrease in their ability to live independently [29]. This need is considered imperative because of the great contribution it could make as a holistic technological ecosystem to improve people's quality of life.

Table 1: Categories, subcategories, and actions

N	Categories	Subcategories		Actions
1	Lights	Kitchen Diningroom Bathroom Livingroom Basement Attic Outside	Backyard Toilet Library Cellar Random Place Myplace Home All	On Off Open Down Up None
2	Camera			
3	Heating			
4	Shutters			
5	Toaster			
6	Garagedoor			
7	Coffeemachine			
8	Direction			
9	Information			
10	Openapp			
11	Other			
12	Time			
13	Weather			

3.4 Data Set Development

The dataset[1] consists of 663 records and required labeling of several variables, as shown in Table 2. These variables were "Category" to which the action belongs, "Action needed", whether the action is necessary or not for the user, "Question", whether what is expressed by the user is a question or an instruction, the specific "Subcategory" to which the action belongs, the "Action" itself to be carried out, the "Time" or moment at which the request is made and the "Sentence" which describes precisely and concisely the requests indicated through speech or voice by the user.

The dataset was built through the collection of samples of real interactions between users and the smart home system. Each sample was labeled with the aforementioned variables, which allowed the training and development of a speech recognition model to interpret and understand user requests efficiently and accurately. It is essential to recognize that a good model will always depend on the data set to achieve an acceptable accuracy that leads to optimizing the needs and preferences of users, providing a more personalized and satisfactory experience in the control and management of a smart home.

[1] https://github.com/oscarp-caceres/speech-recognition

Table 2: Vector of the dataset

N	Category	Action needed	Question	Subcategory	Action	Time	Sentence
1	Lights	1	0	Kitchen	On	Today	Illuminate the kitchen today
2	Camera	1	1	Bathroom	On	Yesterday	Was the camera in the bathroom on yesterday?
3	Heating	1	1	Library	On	Hour	Can you turn on the heating in the library in an hour?
4	Shutters	1	0	Basement	Down	Hour	Let the shutters down in the basement in an hour
5	Toaster	1	0	Kitchen	Off	Now	Turn off the toaster for me please
...	...	...	...	...	...	...	...
663	Garagedoor	1	1	Garage	None	Now	Can you please stop doing something to the garage door?

3.5 Model Coding

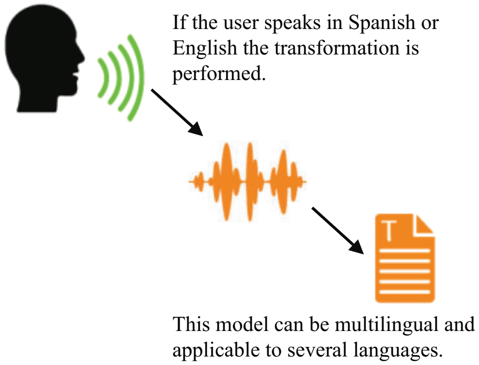
The model's functionality focuses on predicting the category, subcategory, and action based on the variable-label text "Sentence" (Table 2 shows some examples). In this case, machine learning and natural language processing techniques were used. Naive Bayes and Random Forest algorithms were used to building the classification models through the scikit-learn library in Python. The Multinomial Naive Bayes (MultinomialNB) algorithm is employed for text or data classification problems with a multinomial distribution, such as document classification, sentiment analysis, or topic categorization. At the same time, the RandomForestClassifier algorithm is very versatile for classification problems such as regression. It is beneficial because it does not require complex hyperparameter tuning and is less prone to overfitting than the decision tree algorithm, among the main limitations of the algorithms. MultinomialNB may have difficulties capturing complex relationships between features or words in a text, resulting in reduced accuracy in tasks requiring more sophisticated natural language processing. While RandomForest may face difficulties in handling data sets with high dimensionality, which may affect its performance and increase the risk of overfitting [30].

Table 3 illustrates the coding in Python to transform speech into text. The code starts with the import of the "speech_recognition" library by assigning its properties to the variable "sr". It creates an instance of the "speech_recognition" object by giving it to "r", which allows performing speech recognition operations. It then defines a function called "transcribe_microphone()" to perform speech recognition and translation. Within this function, the context "with" is used to establish a connection to the microphone, and the message "Say something" is displayed to guide the user. The function captures audio from the microphone using the "listen()" method of the "r" object. An exception handling block addresses possible problems, printing "Speech could not be recognized" in case the recognition fails "(sr.UnknownValueError)" or an error message if there are problems with the Google service "(sr.RequestError)". If there are no errors, the "recognize_google()" method of "r" is used to translate the recognized text into the desired language, storing the translated text in the variable "text". The translated text is returned, and outside the function, "transcribe_microphone()" is called to perform the recognition and translation process, storing the result in "transcribed_text," which is then used as input data for the model.

Table 3: Class to receive the user's speech or voice and transform it into text

Encoding to transform speech to text

```
import speech_recognition as sr
r = sr.Recognizer()
#It is possible to translate the user's speech into more
than one language.
def transcribe_microphone():
  with sr.Microphone() as source:
    print("Say something...")
    audio = r.listen(source)
        try:
      text = r.recognize_google(audio,
      language='es-ES')
      return text
    except sr.UnknownValueError:
      print("Speech could not be recognized")
    except sr.RequestError as e:
      print(f"Error requesting Google speech
      recognition results; {e}")
texto_transcrito = transcribe_microphone()
print(texto_transcrito)
```



If the user speaks in Spanish or English the transformation is performed.

This model can be multilingual and applicable to several languages.

Table 4 describes the employment of the Multinomial Naive Bayes algorithm for category and subcategory classification and the Random Forest algorithm for action classification. The data set was divided into training and test data. Then, "CountVectorizer" was used to create a matrix of text features in the label "*Sentence*". Subsequently, classification models were built and trained by dividing each vector label used (category, subcategory, and action). Finally, the model's accuracy was evaluated on the test set, and predictions were made on new data achieving acceptable findings.

Table 4: Model building in Python

Coding Models

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer

from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import joblib


datos = pd.read_csv(‘dataset.csv’)

X = datos[‘Sentence’]
y_category = datos[‘Category’]
y_subcategory = datos[‘Subcategory’]
y_action = datos[‘Action’]
X_train, X_test, y_category_train, y_category_test, y_subcategory_train, y_subcategory_test, y_action_train, y_action_test = train_test_split(X, y_category, y_subcategory, y_action, test_size=0.2, random_state=42)

vectorizer = CountVectorizer()

X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test)
joblib.dump(vectorizer, ‘vectorizer.joblib’)
category_classifier = MultinomialNB()

category_classifier.fit(X_train_vectorized, y_category_train)
#continue code ( . . . )
nuevas_oraciones = [“Illuminate the kitchen today.”]
nuevas_oraciones_vectorized = vectorizer.transform(nuevas_oraciones)
category_predictions = category_classifier.predict(nuevas_oraciones_vectorized)
subcategory_predictions = subcategory_classifier.predict(nuevas_oraciones_vectorized)
action_predictions = action_classifier.predict(nuevas_oraciones_vectorized)
print(“Category Predictions:”, category_predictions)
print(“Subcategory Predictions:”, subcategory_predictions)
print(“Action Predictions:”, action_predictions)
```

The parameters set in Table 4, in particular, the y_action_test tags, define the target actions to be recognized, test_size controls how the data split to train and test the model, and random_state ensures that the data splitting is reproducible, which facilitates comparison of results and consistency across different runs. In this case, the value of 0.2 indicates that 20% of the data set was used as the test set, while the remaining 80% was considered for training the model. This division has been essential to assess the generalizability of the model. However, this ratio may vary depending on the size of the data set and the balance between the number of samples in each class. If the data set is small, a higher ratio can be considered for the test set to have enough evaluation data. It is also noted that in some cases, experimenting with different values and evaluating the model’s performance can help find the optimal combination to obtain satisfactory results.

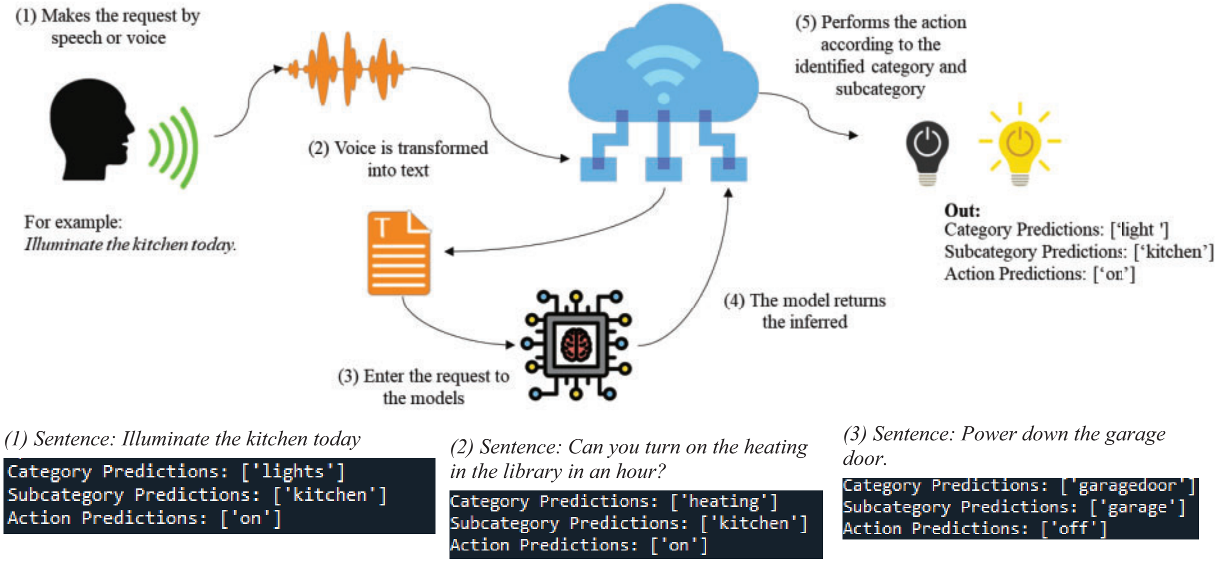
3.6 Testing and Error Analysis

Testing plays a key role in machine learning models because of its importance in assessing performance, efficiency, and robustness. Testing provides a clear understanding of how the model behaves. For the case study, a test dataset consisting of 25 records was constructed, and of these, 23 have been correctly classified and recognized. The tests were carried out in different environments. According to the categories offered by the model, mainly on and off lights, heating, opening and closing blinds, and the use of household appliances such as the toaster, interactions were simulated at different times according to questions such as, "Illuminate the dining room today", "Can you turn on the heating in the bathroom when it freezes outside?" and "Raise the shutters in the basement". In this sense, testing and error analysis are essential components in developing and continuously improving such systems. As systems integrate services based on machine learning, they will facilitate the daily lives of the residents of a home. On the other hand, it is necessary to recognize that many activities in smart homes need preparation time before they can be carried out [31].

These evaluations not only allow us to discern the depth of the model's behavior but also provide a comprehensive picture of its adaptive and predictive capabilities. The consistency and reliability of these results accentuate the inescapable importance of testing and error analysis as intrinsic components for the iterative advancement and ongoing optimization of systems of this nature. In summary, testing, as the backbone of validation and continuous improvement, plays an imperative role in the consolidation and effectiveness of machine learning systems for performance optimization in intelligent residential contexts.

Fig. 2 outlines the procedure used to simulate the proposal. The illustration starts from the user's speech or voice to the inferred outputs that determine the category, subcategory, and action to be performed by the system. The tests were simulated using the Anaconda platform working environment through the Spyder tool.

This type of testing has provided valuable feedback that has helped strengthen the optics for developing future models in the smart home environment. This approach can become a powerful solution to improve people's lives in smart home environments and provide assistance and autonomy to those with physical limitations or disabilities. This technology can make homes more accessible, intuitive, and personalized to meet the individual needs of users. As machine and deep learning techniques are known, they seek to make predictions whose purpose is focused on inferring activities that could occur in the future. The model obtained an accuracy of 79.70% in category prediction, which indicates that it can classify correctly. As for subcategory, the model achieved an accuracy of 71.43%, which implies that it is correct in the subcategorization of the instances. The action recognition model achieved an accuracy of 89.47%, correctly identifying the actions associated with the texts in most cases. These results show promising performance for each classification and prediction model for the different labels. Fig. 3 represents the confusion matrix of the three models. The confusion matrix unnormalized categories shows that the model correctly assigned most instances to their original category, although some distortions have occurred slightly between similar categories. While the confusion matrix does not normalize subcategories, it can be seen how the model successfully assigned most of the instances to their corresponding subcategories. However, some confusion between similar subcategories also occurred, slightly affecting overall accuracy. As for the unnormalized action confusion matrix, many instances have been correctly classified into the corresponding actions. This indicates that the model has learned patterns and distinctive features to accurately identify actions.


(1) Makes the request by speech or voice
For example:
Illuminate the kitchen today.
(2) Voice is transformed into text
(3) Enter the request to the models
(4) The model returns the inferred
(5) Performs the action according to the identified category and subcategory
Out:
Category Predictions: ['light ']
Subcategory Predictions: ['kitchen']
Action Predictions: ['on']

(1) Sentence: Illuminate the kitchen today

```
Category Predictions: ['lights']
Subcategory Predictions: ['kitchen']
Action Predictions: ['on']
```

(2) Sentence: Can you turn on the heating in the library in an hour?

```
Category Predictions: ['heating']
Subcategory Predictions: ['kitchen']
Action Predictions: ['on']
```

(3) Sentence: Power down the garage door.

```
Category Predictions: ['garagedoor']
Subcategory Predictions: ['garage']
Action Predictions: ['off']
```

Figure 2: Simulation of model operation

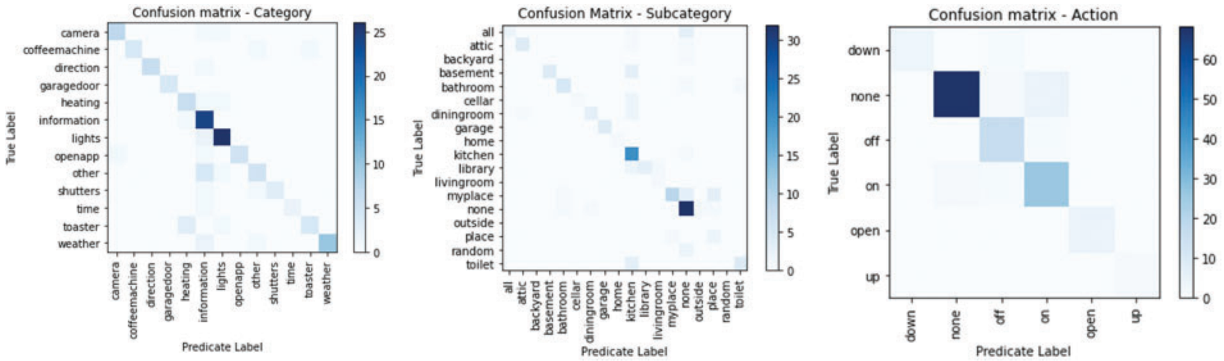


Figure 3: Model confusion matrix

3.7 *Model Refinement*

In data science and machine learning, refining classification models is fundamental. After analyzing the confusion matrix and evaluating the accuracy of the models in classifying categories, subcategories, and actions, the need to improve the performance of the models and address areas where confusion occurs arises. To achieve greater accuracy and discrimination capability, it is possible to consider collecting additional, high-quality data to enrich the training set. This could help to improve the representation of categories, subcategories, and actions in the model, allowing better generalization and distinction between them. However, this process will be considered a working milestone for a future study to extend the capacity and robustness of the models. In addition, with the advancement of AI and substance increase in machine learning research [32], new techniques are expected to be available and used to help this type of proposal [33], to reduce the burden and increase performance and robustness [34].

4 Results and Discussion

The development of machine learning models has led to recognizing human actions through voice in the smart home domain. This approach has been possible thanks to the quality of the dataset and the application of machine learning best practices. The Naive Bayes multinomial classifier proved suitable for classification with discrete features. Using the CountVectorizer, it was possible to recount each word in the text and construct a numerical vector representing the presence or absence of each word. This text processing technique is used in the field of natural language processing. It is convenient to use CountVectorizer in classification and topic modeling tasks, where numerical representation of texts is required to apply machine learning algorithms. By converting text documents into numerical vectors, one can use this representation to train classification models or perform topic analysis based on the frequency of words in the texts. As mentioned above, the classification models were evaluated using precision and confusion matrix metrics. In the accuracy evaluation, it is reaffirmed that the category model achieved an accuracy of 79.70%, an accuracy of 71.43% for the subcategory, and an accuracy of 89.47% for the action. These results indicate a promising performance on each of the classification models using the different labels. However, when analyzing the confusion matrix, areas of confusion were identified. In the category model, some instances were observed to be incorrectly classified in similar categories. In the subcategory model, there was also confusion between related subcategories, indicating the need for further distinction in specific themes or characteristics. On the other hand, the action confusion matrix revealed high accuracy in most instances.

It should be noted that the models are supported by the confusion matrix analysis. While the models have achieved generally high accuracy, the confusion matrix reveals that some confounding has occurred between similar categories or subcategories. These findings highlight the importance of conducting a more detailed analysis and looking for opportunities for improvement in future iterations. Having assessed the confusion matrix, it is relevant to note that the normalized confusion matrix also provides a more detailed view of how the misclassifications are distributed in relation to the true classes. Fig. 4 represents the normalized confusion matrix on the label "category". It can be seen that, although the overall accuracy of the model is acceptable, the data set on categories that represent noise, for example, still needs to be improved, as other, shutters, times, and toaster.

Compared to the abstract representation in Fig. 5, the normality of the "subcategory" criterion, there is a little more noise in the inference. This is because these criteria or subcategories are associated with a category. However, the results obtained have represented a valid proximity in the identification of categories and subcategories, without generating a high rate of distortion that affects the action that the system should perform by the user through speech or voice.

Fig. 6 illustrates the ability of the output variable "action", which is one of the best results of the three classification models. The bias rate is minimal and is only affected by the down, none, and off Labels. Although the model's overall accuracy is high, there is room for improvement.

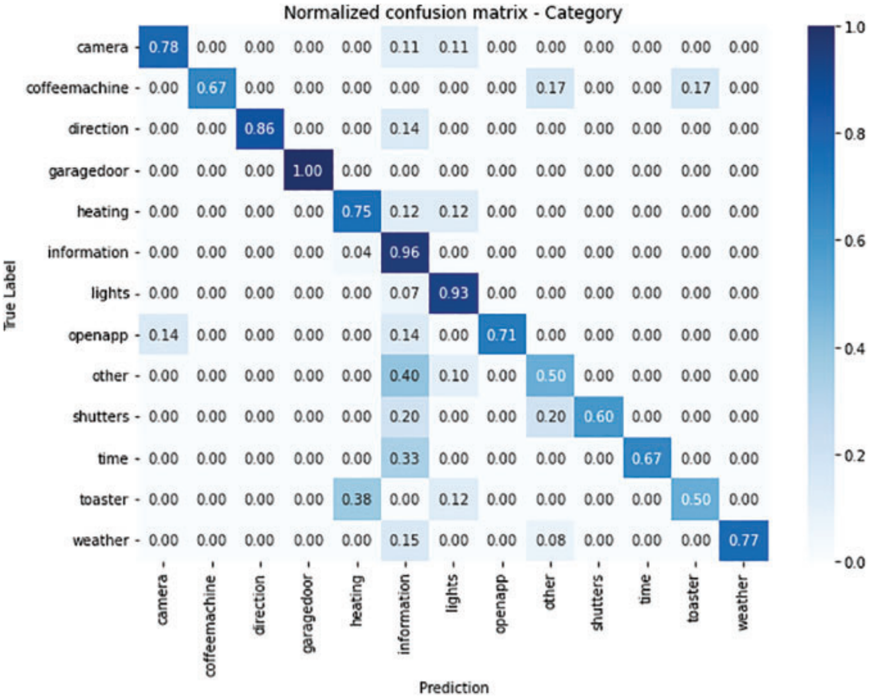


Figure 4: Normalized confusion matrix of the category model

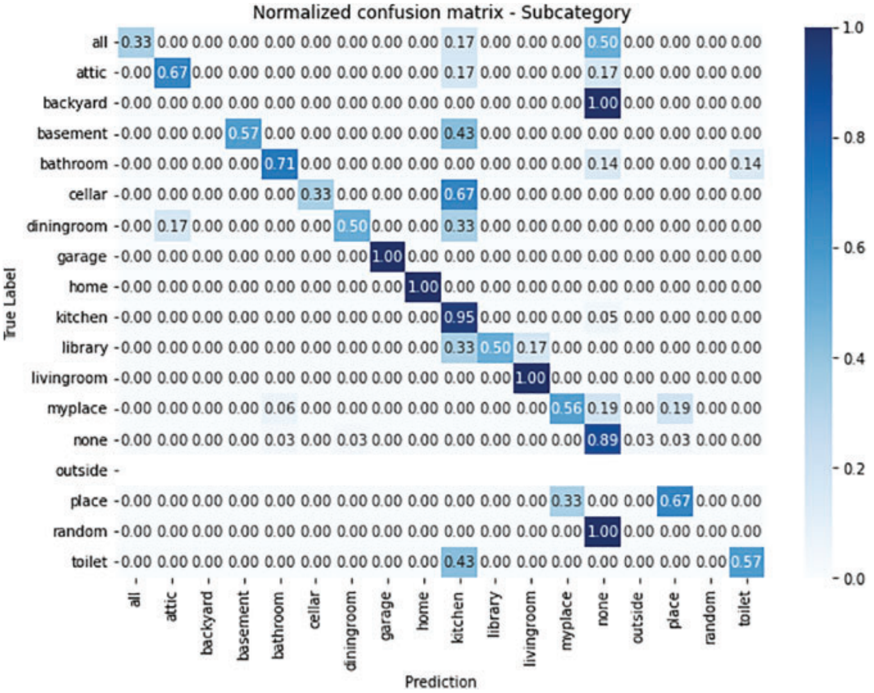


Figure 5: Normalized confusion matrix of the subcategory model

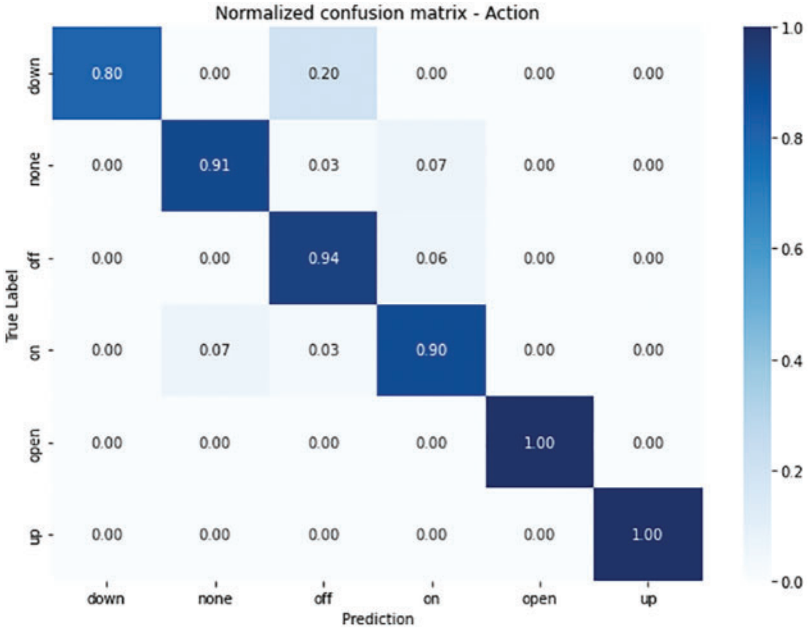


Figure 6: Normalized confusion matrix of the action model

Research such as [20] specified that this type of method will facilitate the creation of voice-controlled systems, fulfilling the essential requirements of human beings and taking into account the possible scenarios of human-machine collaboration, achieving a significant improvement. Also, Kumar et al. [35] described that the proposed algorithms offer better performance during their operation than existing technologies, suggesting the support vector machine algorithm. The authors in [36] discussed and recommended using seizure model-centered machine learning algorithms for action recognition. While reference [16] pointed out that signal processing and machine learning techniques must be used to recognize speech and that traditional systems have low performance. The approach of [37] describes that using mobile applications for speech recognition is possible as an alternative that allows processing tasks remotely or in contexts of high criticality on the user's part.

Hung et al. [21] proposed a control unit in a smart home using voice commands in Vietnamese. This study uses machine learning techniques. Its main limitation is that it can only recognize voice commands in the Vietnamese language. Compared to the proposed solution, the main advantage of voice recognition is that it is multilingual, which allows a broader range of applicability and demonstrates that machine learning techniques are not limited to a particular set of commands but can be extrapolated to different contexts in smart home environments. In [38], the authors used machine learning algorithms for human recognition from audio signals; their findings show that the Naive Bayes algorithm represents an accuracy of 60%. The proposed solution adopts an average value of 80.2% accuracy. In [39], the authors used the Naive Bayes classification algorithm to classify tasks linked to an intelligent virtual assistant based on speech recognition, which offers the best accuracy among k-nearest neighbors, logistic regression, random forest, decision tree, and support vector machine. This work reaffirms that the Multinomial Naive Bayes algorithm is potentially suitable for developing activities related to speech recognition.

The mobile application of [40] mySmartCart focuses on transforming the traditional way of writing a shopping list into a digitized smart list that implements speech recognition using the

Google Speech cloud service, with costs ranging from $0.036 per minute and limited to specific languages. It uses the Naive Bayes algorithm for text classification, achieving 76% accuracy. Unlike the aforementioned study the proposed solution does not require any economic cost for the execution of tasks that transform the voice into text and identify the request requested by the user. It is essential to note that this research focused on the application of speech recognition-based classification and natural language processing in the context of smart homes. Although traditional approaches, such as regression analysis or rule-based techniques, exist to address similar issues, they often need help capturing the complexity of human interactions in such a dynamic and diversified environment as intelligent homes. Our choice to employ the Multinomial Naive Bayes and RandomForestClassifier algorithms was based on their ability to deal with nonlinear relationships and handle multiple features efficiently. Our approach is best suited to the inherent nature of the problem at hand, offering promising results both in terms of accuracy and in its ability to adjust to variations in natural language and human actions in the context of smart homes.

To improve the model's performance, it is necessary to consider expanding the dataset by collecting more examples for each label. This could help strengthen the representation of categories, subcategories, and actions, allowing for better model generalization. In addition, it is possible to explore advanced natural language processing techniques, such as word embeddings or pre-trained language models, to capture semantic and contextual relationships in texts better. Another area for improvement focuses on the tuning of model hyperparameters. A common hyperparameter is the vocabulary size, which defines the number of unique words in the text representation. A large vocabulary can improve information richness but also increase computational complexity. Another hyperparameter is the size of the hidden layer in the case of using neural networks, which controls the number of units in the inner layers of the model. A larger size may allow the model to learn more complex features but could also lead to overfitting problems. The learning rate is another crucial hyperparameter, which determines how the model weights are updated during training. A low learning rate may lead to slow convergence, while a high rate could cause the model to oscillate and not converge. Finally, the number of epochs, representing the number of times the model sees the entire training set, is also an important hyperparameter affecting model performance. An adequate number of epochs is essential to balance the fit to the training set and the ability to generalize to new data. Careful selection and adjustment of these hyperparameters are critical to obtaining well-calibrated models with optimal performance on various tasks [41]. Exhaustive search for optimal hyperparameter combinations or optimization techniques can help find a configuration that improves accuracy and reduces confounding. The obtained results demonstrate promising performance of the classification model in predicting categories, subcategories, and actions based on the "Sentence" tag texts. Although high accuracy was achieved overall, the confounding identified in the confusion matrix highlights the need for further improvement and refinement of the model.

5 Conclusion

This research proposes recognizing human actions through speech or voice using machine learning. The accuracy obtained for the category model is 82.99%, for the subcategory model 76.19%, and for the action model 90.28%. While exactitude is described as 79.70% for the category model, 71.43% for the subcategory model, and 89.47% for the action model. Each model has performed well in classifying the categories, subcategories, and actions based on the input texts. However, it is essential to consider that these results are based on specific metrics and a specific data set. It is recommended for future work that the vector integrates information on user preferences, context, and level of criticality so that the experience when interacting with the system shows expected responses and is not out of context. It is also suggested to explore models based on rules and fuzzy logic that could provide a logical and transparent structure to establish patterns of relationships between voice commands and specific actions that improve the system's ability to interpret the user's intentions coherently. Also, we plan to investigate using deep learning models, such as neural networks, to improve the ability to capture complex relationships and subtle patterns in human speech. Then, explore their adaptability in different contexts and environments, such as variations in accents, noisy environments, or users with special needs. In addition, it is necessary to mention that systems that integrate this type of services help to increase confidence in the human-system interaction and that the person is characterized by being the main actor before, during, and after using an intelligent system.

Acknowledgement: Thanks to the Universidad César Vallejo for their collaboration in this study.

Funding Statement: This work was supported by Generalitat Valenciana with HAAS (CIAICO/2021/03 9) and the Spanish Ministry of Science and Innovation under the Project AVANTIA PID2020-114480RB-I00.

Author Contributions: Oscar Peña-Cáceres: Data curation, Formal analysis, Software, Methodology, Writing, Original draft. Henry Silva-Marchan: Conceptualization, Fundraising. Manuela Albert: Conceptualization, Research, Methodology, Writing (original draft), Acquisition of funds. Miriam Gil: Conceptualization, Research, Methodology, Validation, Writing (original draft), Fundraising.

Availability of Data and Materials: The data used in this paper can be requested from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] N. S. Jaddu, S. R. S. Shashank and A. Suresh, "Voice emotion detection: Acoustic features extraction using multi-layer perceptron classifier algorithm," in *Proc. of Int. Conf. on Innovative Computing and Communications*, Delhi, India, vol. 492, pp. 593–602, 2023.

[2] H. Luo, J. Du, P. Yang, Y. Shi, Z. Liu *et al.,* "Human-machine interaction via dual modes of voice and gesture enabled by triboelectric nanogenerator and machine learning," *Applied Materials and Interfaces*, vol. 15, no. 13, pp. 17009–17018, 2023.

[3] P. Nath and U. C. Pati, "Low-cost android app based voice operated room automation system," in *Proc. of 3rd Int. Conf. for Convergence in Technology (I2CT)*, Pune, India, pp. 1–4, 2018.

[4] R. Alshamsi, F. Ahli, M. Alqassim, R. A. Alhamad and T. Bonny, "Voice controlled bedroom using aiy google kit," in *Proc. of Advances in Science and Engineering Technology Int. Conf. (ASET)*, Dubai, United Arab Emirates, pp. 1–6, 2022.

[5] N. H. Abdallah, E. Affes, Y. Bouslimani, M. Ghribi, A. Kaddouri *et al.,* "Smart assistant robot for smart home management," in *Proc. of 1st Int. Conf. on Communications, Control Systems and Signal Processing*, El Oued, Algeria, pp. 317–321, 2020.

[6] I. Stefanovic, E. Nan and B. Radin, "Implementation of the wake word for smart home automation system," in *Proc. of IEEE Int. Conf. on Consumer Electronics*, Berlin, Germany, pp. 271–272, 2017.

[7] S. Kanjanawattana, A. Jarat, P. Praneetpholkrang, G. Bhakdisongkhram and S. Weeragulpiriya, "Classification of human emotion from speech data using deep learning," in *Proc. of 5th Int. Conf. on Big Data and Artificial Intelligence*, Fuzhou, China, pp. 1–5, 2022.

[8] C. H. M. H. Saibaba, S. F. Waris, S. H. Raju, V. Sarma, V. C. Jadala *et al.,* "Intelligent voice assistant by using OpenCV approach," in *Proc. of Second Int. Conf. on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, pp. 1586–1593, 2021.

[9] S. S. Priya, R. Poongodan and D. Chellani, "Augmented reality and speech control from automobile showcasing," in *Proc. of 4th Int. Conf. on Smart Systems and Inventive Technology*, Tirunelveli, India, pp. 1703–1708, 2022.

[10] X. Li, H. Chen, S. He, X. Chen, S. Dong *et al.,* "Action recognition based on multimode fusion for VR online platform," *Virtual Real*, vol. 27, no. 3, pp. 1797–1812, 2023.

[11] A. B. Andrew, S. A. Rajiv, N. Jusat, A. A. Zainuddin, K. Subramaniam *et al.,* "Implementation of low-cost voice command robot using arduino uno platform," in *Proc. of IEEE 7th Int. Conf. on Smart Instrumentation, Measurement and Applications*, Bandung, Indonesia, pp. 134–139, 2021.

[12] D. Boucha, A. Amiri and D. Chogueur, "Controlling electronic devices remotely by voice and brain waves," in *Proc. of Int. Conf. on Mathematics and Information Technology*, Adrar, Algeria, vol. 2018, pp. 38–42, 2017.

[13] N. Almayouf, S. M. Qaisar, L. Alharbi and R. Madani, "A speech to machine interface based on the frequency domain command recognition," in *Proc. of IEEE 2nd Int. Conf. on Signal and Image Processing*, Singapore, vol. 2017, pp. 356–360, 2017.

[14] M. S. Nguyen and T. L. Vo, "Resident identification in smart home by voice biometrics," in *Proc. of Int. Conf. on Future Data and Security Engineering*, Ho Chi Minh City, Vietnam, vol. 11251, pp. 433–448, 2018.

[15] G. Alexakis, S. Panagiotakis, A. Fragkakis, E. Markakis and K. Vassilakis, "Control of smart home operations using natural language processing, voice recognition and IoT technologies in a multi-tier architecture," *Designs*, vol. 3, no. 3, pp. 1–18, 2019.

[16] M. H. Ali, M. M. Jaber, S. K. Abd, A. Rehman, M. J. Awan *et al.,* "Harris hawks sparse auto-encoder networks for automatic speech recognition system," *Applied Sciences*, vol. 12, no. 3, pp. 1091, 2022.

[17] L. Huang and C. Wang, "WalkID: Towards context awareness of smart home by identifying walking sounds," in *Proc. of IEEE 6th World Forum on Internet of Things (WF-IoT)*, New Orleans, LA, USA, pp. 1–6, 2020.

[18] R. Jiménez-Moreno and R. A. Castillo, "Deep learning speech recognition for residential assistant robot," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 585–592, 2023.

[19] Z. Lv, F. Poiesi, Q. Dong, J. Lloret and H. Song, "Deep learning for intelligent human-computer interaction," *Applied Sciences*, vol. 12, no. 22, pp. 11457, 2022.

[20] A. Rogowski, "Scenario-based programming of voice-controlled medical robotic systems," *Sensors*, vol. 22, no. 23, pp. 9520, 2022.

[21] P. D. Hung, T. M. Giang, L. H. Nam, P. M. Duong, H. Van Thang *et al.,* "Smarthome control unit using vietnamese speech command," *Advances in Intelligent Systems and Computing*, vol. 1072, pp. 290–300, 2020.

[22] A. M. Anter, A. W. Mohamed, M. Zhang and Z. Zhang, "A robust intelligence regression model for monitoring Parkinson's disease based on speech signals," *Future Generation Computer Systems*, vol. 147, pp. 316–327, 2023.

[23] Y. Xu, Y. Li, Y. Chen, H. Bao and Y. Zheng, "Spontaneous visual database for detecting learning-centered emotions during online learning," *Image and Vision Computing*, vol. 136, no. 4, pp. 104739, 2023.

[24] M. Tanveer, A. Rastogi, V. Paliwal, M. A. Ganaie, A. K. Malik *et al.,* "Ensemble deep learning in speech signal tasks: A review," *Neurocomputing*, vol. 550, no. 2, pp. 126436, 2023.

[25] S. Wu, S. Huang, W. Chen, F. Xiao and W. Zhang, "Design and implementation of intelligent car controlled by voice," in *Proc. of Int. Conf. on Computer Network, Electronic and Automation*, Xi'an, China, pp. 326–330, 2022.

[26] X. Wang, J. Liang, X. Xu, X. Li and Q. Zhu, "An online classroom question answering evaluation system based on voiceprint and behavior recognition," in *Proc. of 21st Int. Symp. on Distributed Computing and Applications for Business Engineering and Science*, Chizhou, China, pp. 66–69, 2022.

[27] M. Kenoui and M. A. Mehdi, "Collaborative learning environment over internet using augmented reality and voice interaction," in *Proc. of 5th Edition of the Int. Conf. on Advanced Aspects of Software Engineering*, Constantine, Algeria, pp. 1–8, 2022.

[28] R. Kumaraswamy, S. Srivastav, Saurabh, Rishabh and N. B. Ks, “The human assistant system,” in *Proc. of Int. Conf. on Futuristic Technologies (INCOFT)*, Belgaum, India, pp. 1–5, 2022.

[29] I. Dratsiou, A. Varella, E. Romanopoulou, O. Villacañas, S. Cooper *et al.,* “Assistive technologies for supporting the wellbeing of older adults,” *Technologies*, vol. 10, no. 1, pp. 8, 2022.

[30] A. N. M. Jubaer, A. Sayem and M. A. Rahman, “Bangla toxic comment classification (machine learning and deep learning approach),” in *Proc. of 8th Int. Conf. on System Modeling and Advancement in Research Trends*, Moradabad, India, pp. 62–66, 2020.

[31] W. Wang, J. Li, Y. Li and X. Dong, “Predicting activities of daily living for the coming time period in smart homes,” *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 1, pp. 228–238, 2023.

[32] P. Rajesh and R. Kavitha, “An imperceptible method to monitor human activity by using sensor data with cnn and bidirectional LSTM,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, pp. 96–105, 2023.

[33] T. C. Hsu and M. S. Chen, “The engagement of students when learning to use a personal audio classifier to control robot cars in a computational thinking board game,” *Research and Practice in Technology Enhanced Learning*, vol. 17, no. 1, pp. 27, 2022.

[34] J. V. Jeyakumar, A. Sarker, L. A. Garcia and M. Srivastava, “X-CHAR: A concept-based explainable complex human activity recognition model,” in *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, New York, USA, vol. 7, pp. 1–28, 2023.

[35] S. Kumar, M. A. Haq, A. Jain, C. Andy Jason, N. R. Moparthi *et al.,* “Multilayer neural network based speech emotion recognition for smart assistance,” *Computers, Materials & Continua*, vol. 74, no. 1, pp. 1523–1540, 2023.

[36] J. Ma and J. Han, “Value evaluation of human motion simulation based on speech recognition control,” *International Journal of Systems Assurance Engineering and Management*, vol. 14, no. 2, pp. 796–806, 2022.

[37] I. Diddeniya, I. Wanniarachchi, H. Gunasinghe, C. Premachandra and H. Kawanaka, “Human-robot communication system for an isolated environment,” *IEEE Access*, vol. 10, pp. 63258–63269, 2022.

[38] U. Sadique, M. S. Khan, S. Anwar and M. Ahmad, “Machine learning based human recognition via robust features from audio signals,” in *Proc. of 3rd IEEE Int. Conf. on Artificial Intelligence*, Islamabad, Pakistan, pp. 52–57, 2023.

[39] K. M. Bhargav, A. Bhat, S. Sen, A. V. K. Reddy and S. D. Ashrith, “Voice-based intelligent virtual assistant for windows,” in *Proc. of Int. Conf. on Innovations in Computer Science and Engineering*, Hyderabad, India, vol. 565, pp. 491–500, 2023.

[40] S. K. Nanjappa, S. Prakash, A. Burle, N. Nagabhushan and C. S. Kumar, “mySmartCart: A smart shopping list for day-to-day supplies,” *IAES International Journal of Artificial Intelligence*, vol. 12, no. 3, pp. 1484–1490, 2023.

[41] C. Alzaman, “Forecasting and optimization stock predictions: Varying asset profile, time window, and hyperparameter factors,” *Systems and Soft Computing*, vol. 5, pp. 200052, 2023.