Towards socially acceptable food type recognition

Junjie Wang*, Jiexiong Guan*, Y.Alicia Hong†, Hong Xue†, Shuangquan Wang‡,
Zhenming Liu*, Bin Ren*, Gang Zhou*
*William & Mary, †George Mason University, ‡Salisbury University
jwang51@wm.edu*, jguan@wm.edu*, yhong22@gmu.edu†, hxue4@gmu.edu†, spwang@salisbury.edu‡,
lzhenming@gmail.com*, bren@cs.wm.edu*, gzhou@cs.wm.edu*

2022 18th International Conference on Mobility, Sensing and Networking (MSN) | 978-1-6654-6457-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/MSN57253.2022.00110

***Abstract*—Automatic food type recognition is an essential task of dietary monitoring. It helps medical professionals recognize a user's food contents, estimate the amount of energy intake, and design a personalized intervention model to prevent many chronic diseases, such as obesity and heart disease. Various wearable and mobile devices are utilized as platforms for food type recognition. However, none of them has been widely used in our daily lives and, at the same time, socially acceptable enough for continuous wear. In this paper, we propose a food type recognition method that takes advantage of Airpods Pro, a pair of widely used wireless in-ear headphones designed by Apple, to recognize 20 different types of food. As far as we know, we are the first to use this socially acceptable commercial product to recognize food types. Audio and motion sensor data are collected from Airpods Pro. Then 135 representative features are extracted and selected to construct the recognition model using the lightGBM algorithm. A real-world data collection is conducted to comprehensively evaluate the performance of the proposed method for seven human subjects. The results show that the average f1-score reaches 94.4% for the ten-fold cross-validation test and 96.0% for the self-evaluation test.**

***Index Terms*—earbuds, Airpods Pro, socially acceptable, food type recognition, lightGBM**

I. Introduction

Unhealthy eating habits are one of the major causes of some chronic diseases, such as obesity, diabetes, metabolic syndrome, and heart diseases. According to the Centers for Disease Control and Prevention (CDC) report, in the USA, the obesity prevalence is nearly 40% among adults [1], the prevalences of diabetes and metabolic syndrome are 11.3% and 35% [2], [3], respectively, and heart diseases lead the 20.6% of deaths [4]. Chronic diseases have caused a great burden for both individuals and society. To solve the problem, automatic dietary monitoring (ADM) is a good solution because it can help people maintain healthy eating habits. ADM systems aim to identify: 1) when does the eating activity happen; 2) what is the food type consumed; 3) how much is consumed. This paper focuses on the second topic, i.e., food type recognition, a critical component among these three topics.

Multiple wearable devices have been developed to automatically recognize food types in recent years. For example, the necklace-based devices embedded with microphones [5], or proximity, ambient light, and an inertial motion sensor (IMU) [6]. A headband-based device equipped with an accelerometer and gyroscope [7]. These devices need to be attached to the skin or head tightly. Keum et al. present an intraoral mouth sensor containing a temperature sensor and accelerometer, which needs to be put inside the mouth while eating [8]. In addition to these devices, the earbud is another commonly used device. Many researchers customized earbuds prototypes equipped with microphones to collect audio data [9]–[13], and some may combine earbuds with several other devices. For example, an automatic dietary monitoring system, which contains customized earbuds, LG smartwatch, and Google Glass, is used to recognize food types and detect eating events [14]–[16].

Although multiple wearable devices introduced above are utilized as platforms for food type recognition, none of them has been widely used for long-term wear and is socially acceptable for daily lives. Typically, these solutions have three main shortages: 1) Some of them are too intrusive, such as the necklace, headband, and intraoral mouth sensor. As a result, people are not willing to wear or use them; 2) They are not accessible to a large number of users; 3) They are not entirely reliable. In other words, the hardware robustness is not good enough. From the view of healthcare professionals and patients with chronic diseases, these devices are not socially acceptable for long-term daily usage. To solve the problem, we propose a food type recognition method based on Airpods Pro, a commercial product designed by Apple. As far as we know, we are the first to use commercial earbuds alone for food type recognition. Airpods Pro takes 34% of the headphone market share in the USA [17], which strongly demonstrates its social acceptance. The audio and motion sensors data are collected when the earbuds are deployed in the left and right ears. However, recognizing different food types using Airpods Pro is not straightforward, and we need to answer two research questions: 1) What are the most useful features that can effectively represent the differences among chewing different food types? 2) What is the efficient classifier to recognize different food types?

To address these two research questions, we first conduct a data collection. Five male and two female human subjects participated in our experiment. Twenty food types are chosen from the United States Department of Agriculture (USDA) recommendation [18], which is the US government's guidance for healthy eating habits. These 20 food types cover six categories: meat, protein, dairy, grain, fruit, and vegetable. The data collection is done in an apartment's dining room, with a

This work is partially supported by GMU-CHHS pilot grant (Award #PR9449797), and NSF grants CCF-2047516 (CAREER), CCF-2146873, and III-2008557.

978-1-6654-6457-4/22/$31.00 ©2022 IEEE
DOI 10.1109/MSN57253.2022.00110

Fig. 1: Sample of the 20 food types

regular eating environment and facilities. Each human subject needs at least 30 food intakes, and videos are recorded for manually labeling the ground truth.

In order to answer the first research question, we make a comprehensive survey of the research works about extracting features from audio and motion data. We extract spectral features from audio data, such as energy, spectral centroid, MFCC, etc. These features can reflect the difference in chewing different food types. For example, the energy between chewing crispy food and soft food is different. Then, we extract statistical and shape features from motion data to reflect the head movement. In total, we extracted 105 features from audio data and 144 features from motion data. Finally, we used lightGBM to evaluate the importance of each feature and selected the most important 135 features.

To address the second research question, we compared several commonly used classifiers, including Logistic Regression, Naive Bayes, support vector machine (SVM), k-nearest neighbors algorithm (KNN), random forest (RF), and lightGBM, which is the optimized version of RF. Based on the results, we chose lightGBM as our classifier and optimized the parameters, such as the number of estimators, the max depth, the number of leaves, etc. To compare our solution with existing earbud-based solutions, we reimplemented the algorithms of Ubicomp-16 [14] and MLHC-19 [16] and evaluated them on our dataset. The results show that the average f1-score of our solution reaches 94.3% for the ten-fold cross-validation test. Compared with these two baselines, the cross-validation performance of our solution is 8% higher than that of Ubicomp-16 and 30% higher than that of MLHC-19.

In summary, our primary contributions are:

- We are the first to use commercial earbuds alone for food type recognition. We conducted a data collection using Airdpods Pro, to recognize 20 food types for seven human subjects.
- We extracted 105 features from audio data and 144 features from motion data, and selected the most important 135 features to distinguish different food types.
- We chose lightGBM as the classifier and optimized its parameters to achieve high performance. The evaluation results show that our solution outperforms two baselines by 8% and 30%, respectively.

The rest of this paper is organized as follows. First, Section II describes the data collection. Next, the details of feature extraction and classification are given in Section III. The evaluation results are given in Section IV. Then, Section V introduces the related works. Discussions and future work are presented in Section VI. Finally, the conclusion is drawn in Section VII.

II. Data Collection

We developed an ios APP to collect data from Airpods Pro. The human subjects need to wear the Airpods Pro while eating. First, we collected audio and motion data from the embedded microphones and motion sensors. The data are wirelessly transmitted from earbuds to the paired smartphone. Then we

transferred the data from the smartphone to the computer for offline analysis. In this section, we first introduce the specifics of sensors. Then we describe the food types of our study. Finally, we present how to collect data.

A. Sensors

Microphones, accelerometer, and gyroscope are embedded in the Airpods Pro. Accelerometer and gyroscope are called motion sensors. We collected audio data from microphones, where the sampling rate is 44.1 kHz, and the data is one channel. For the motion data collection, the sampling rate is about 25 Hz. We collected the reading of pitch, roll, and yaw from the gyroscope. These three dimensions of data are used to measure the rotation of the head around the vertical, transverse, and longitudianl axes. For the accelerometer, the reading of user acceleration on X, Y, and Z axes are collected, which measures the movement on these three axes.

B. Food type

Twenty food types are selected based on the dietary guidelines for Americans presented by the USDA [18]. They come from six categories: meat, protein, dairy, grain, fruit, and vegetable. These twenty food types are shown in table I, covering all the groups and subgroups of the USDA guidelines.

TABLE I: The selected 20 food types

Category	Food Type
Meat	beef, pork, chicken, fish
Protein	almond, egg, peanut
Dairy	yogurt
Grain	bagel, bread, rice
Fruit	apple, banana, strawberry
Vegetable	broccoli, lettuce, carrot, tomato, green pea, potato

A sample of the 20 food types are shown in figure 1, and all of them are bought from Food Lion supermarket. Eight types of food need to be boiled before eating, including beef, pork, chicken, fish, egg, broccoli, green pea, and potato. The other 12 types of food are eaten raw without any cooking.

C. Groundtruth and Collected Dataset

We conducted the data collection in an apartment's dining room, with a regular eating environment and facilities. We use a camera to record the eating process to retrieve the ground truth, as shown in figure 2. The camera captures the movement of the hand, mouth, and head. We manually label the start and end times of each intake according to the video. In this paper, we define the period from putting a piece of food into the mouth to the end of the last chew as one intake.

Five male and two female human subjects participated in our experiment. Each user needs to eat at least 30 intakes for each type of food. They usually took two or three days to finish the data collection of all the food types. Except for collecting eating data, the data during some non-eating periods are also collected. Each non-eating segment is about 20 seconds. The user was free to do any normal activities while seated, with only background noise.

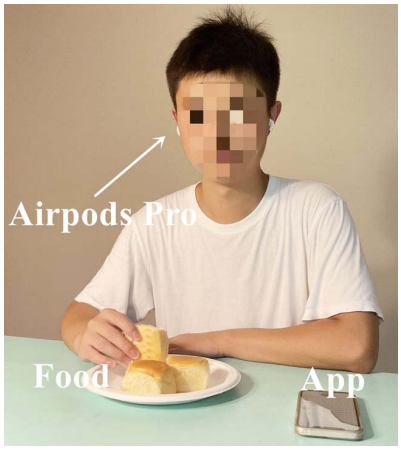


Fig. 2: Screenshot of the groundtruth

In total, there are 4805 segments in our dataset, including 180 non-eating segments and 4625 intake segments for 20 food types. The detail of the collected data for each user and each food type is shown in table II. The food type number 0 represents non-eating, and the rest correspond to the numbers shown in figure 1. For instance, 1 is beef, 2 is pork, and 20 is potato.

III. Feature Extraction and Classifier Selection

Before feature extraction, the collected audio and motion data are segmented by each food intake. The segments are labeled manually according to the video ground truth. For each segment, we extracted features from audio and motion data separately. Then, the audio features and motion features are merged to form the feature vector of a segment. In this section, we first describe how to extract features from audio and motion data. Then, we introduce feature and classifier selection.

A. Audio feature extraction

To characterize the detail of each chew in an intake, we divided each audio segment into many frames. Each frame takes $200ms$, with 50% overlap with its prior frame. The frame length is set to $200ms$ to avoid including multiple chews within a single frame [14]. Then we extracted 21 frame features from each frame. The details of them are shown below:

- *ZCR:* Zero crossing rate (ZCR) is the rate of the data changes from negative to positive and the reverse.
- *energy:* The sum of the square of all data values, normalized by the number of the data within the frame.
- *energy entropy:* The raw frame data is divided into sub-frames, each containing ten samples. This feature is the entropy of all sub-frames energy, which measures the abrupt change of energy.
- *spectral centroid:* The raw data are transformed into spectral signals after fast fourier transform (FFT). Spectral centroid indicates which frequency is the center of mass among the spectral signals.

TABLE II: Number of segments of each user for each food type

Food type	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
User1	47	33	35	33	35	37	30	38	32	36	34	32	38	36	37	37	31	35	33	30	32	731
User2	22	33	33	33	35	33	33	32	33	33	33	33	30	33	34	33	33	32	31	33	32	677
User3	21	33	33	33	33	33	35	33	33	33	33	33	33	33	34	33	34	33	33	33	36	688
User4	22	33	33	33	33	33	31	33	33	32	33	33	33	33	33	33	33	32	33	33	33	678
User5	22	31	33	33	32	34	33	33	33	33	32	33	33	33	33	31	32	34	33	33	33	676
User6	22	33	33	33	32	33	33	33	33	33	32	32	33	32	33	33	33	33	33	33	30	675
User7	24	29	33	34	32	33	33	32	33	33	31	33	33	34	33	34	33	33	33	33	34	680
Total	**180**	**225**	**233**	**232**	**232**	**236**	**228**	**234**	**230**	**233**	**228**	**229**	**233**	**234**	**237**	**234**	**229**	**232**	**228**	**228**	**230**	**4805**

- *spectral spread:* It is the variance of the spectral signals. If the audio has too much noise, usually the spectral spread would be large. On the contrary, this feature would be low if the audio only has isolated peaks.
- *spectral entropy:* The spectral signals are divided into sub-frames with a size of ten samples. This feature is the entropy of these spectral sub-frames energy, which measures the spectral power distribution.
- *spectral flux:* It is "the squared difference between the normalized magnitudes of the spectra [19]" of the current frame and previous frame. If the current frame is the first frame, then the previous one would be itself. This feauture is used to distinguish whether the spectrum changes quickly or not.
- *spectral rolloff:* It is the frequency that 90% of the magnitude distribution is centered among spectral signals. This feature is efficient to distinguish voiced and unvoiced audio signals.
- *MFCC:* Mel Frequency Cepstral Coefficients (MFCC) are widely used in audio signal processing, and we extracted 13 coefficients to distinguish the sounds of chewing different types of food.

For each of the 21 features from each frame, we computed five statistics to form the feature vector of one segment. The five statistics are: $mean$, standard deviation (std), max, min, and $median$. In total, we got $21 \times 5 = 105$ audio features for a segment.

B. Motion feature extraction

Similar to the audio feature extraction, we first divided each motion segment into many frames. However, the sampling rate of the motion sensors are relatively low, i.e., about $25Hz$. Therefore, we set the frame length as $5s$, with 50% overlap with its prior frame. The motion sensors data has six dimensions: pitch, roll, and yaw from the gyroscope and X, Y, and Z from the accelerometer. First, we extracted ten features from each of these six dimensions for each frame, including:

- *basic statistics:* We extracted six basic statistical features, including: $mean$, max, min, $median$, $variance$, and std.
- *ZCR:* ZCR is the rate of the data changes from negative to positive and the reverse.
- *IQR:* Interquartile range (IQR) is the difference between the upper and lower quartiles. This feature measures the spread of the data.
- *skewness:* After getting the probability distribution of the data, this feature measures the asymmetry of the distribution.
- *kurtosis:* Similar to skewness, kurtosis is a feature to measure the distribution. Kurtosis indicates whether the distribution is heavy-tailed or light-tailed.

In addition, we extracted four shape features only from accelerometer data.

- *shape features:* After getting the polynomial fit of the data with the degree of three, then the four coefficients of the polynomial formula are used as shape features.

In total, for each frame, we collected $10 \times 6 = 60$ frame features from both the gyroscope and accelerometer data, and $4 \times 3 = 12$ features only from accelerometer data. For each of them, we computed the $mean$, std value to form the motion feature vector of a segment. There are $72 \times 2 = 144$ motion features for each segment.

C. Feature selection

We extracted 105 features from audio data and 144 features from motion data. In total, there are 249 features for each segment. To eliminate redundant and useless features, a feature selection algorithm is applied to select the most important features for the following classifier construction.

We chose lightGBM to evaluate the feature importance. LightGBM is a gradient boosting framework, which uses tree-based learning algorithms. For the tree-based algorithms, the data is split by a selected feature at each non-leaf node of the tree. A feature would be more important if it is used in more nodes. In lightGBM, the feature importance is represented by the number of times that the feature is used. The larger the number, the more important the feature is. Finally, we selected the most important 135 features. The details of the feature selection process is shown in section IV-C.

D. Classifier selection

To figure out which classifier is most appropriate in our application scenario, we compared multiple widely used classifiers, including Logistic Regression, Naive Bayes, SVM, KNN, RF, and lightGBM. The performance comparison results without parameter optimization are shown in section IV-D, which indicates that lightGBM outperforms the other classifiers.

Ultimately, we decided to use lightGBM as our classifier. To get the best performance, we optimized the model's parameters, such as the number of estimators, learning rate, the max depth, and the number of leaves.

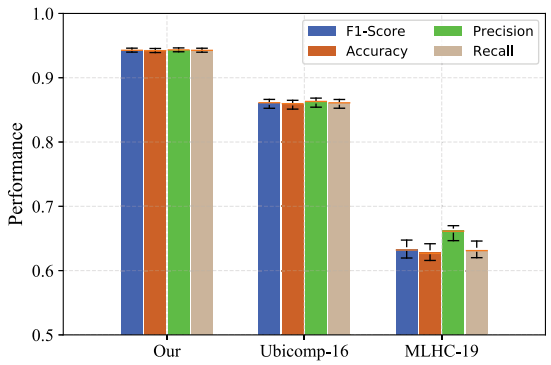


Fig. 3: Results of cross-evaluation

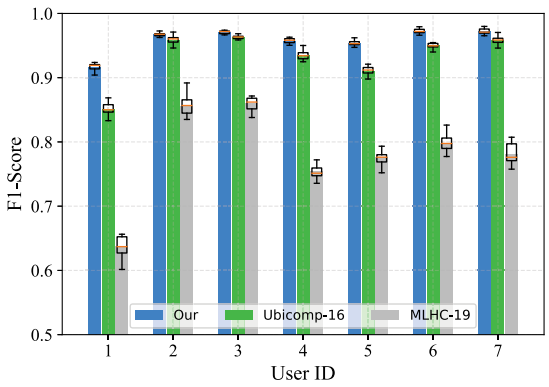


Fig. 4: Results of self-evaluation

IV. Evaluation

In this section, we first present the evaluation results of the cross-evaluation and self-evaluation. Next, we introduce the impact of sensor fusion. Finally, we describe how to select features and classifier.

A. Performance Evaluation

The first question we would like to answer is about the performance of our solution. To answer it, we present the evaluation results of cross-evaluation and self-evaluation. The cross-evaluation is for all users, which means the training and test data come from all users. The self-evaluation is for each individual, where the training and test are evaluated on the same user's data.

We reimplemented the food type recognition algorithms presented in Ubicomp-16 [14] and MLHC-19 [16], respectively, and set them as baselines. We made some modifications to the classification algorithm of MLHC-19 because it does not fully meet our evaluation objective. The original algorithm is a hierarchical classification, where the classification result is a food type or category. However, in our evaluation, the result should be a food type. We revised the algorithm of MLHC-19 as follows: if the classification result is a category, we selected the food type with the highest probability as the classification result. These two baselines are evaluated on our dataset and compared with our solution.

We choose four metrics to evaluate the performance, including f1-score, accuracy, precision, and recall. For accuracy, similar to binary classification, it is the ratio of the number of correctly predicted samples to the number of all samples. As our solution is multi-class classification, our experiment has non-eating and 20 food types. Therefore, for each of the rest three metrics, we first calculated the value of each class separately, thus non-eating and each food type. Then we computed the mean value of all classes to get the overall result.

1) Cross-Evaluation: For cross-evaluation, we conducted ten-fold cross-validation 20 times with different random seeds of splitting data. The evaluation results are shown in figure 3. From this figure, we can see that our solution outperforms the two baselines. The average f1-score of our solution is 94.3%, which is 8% higher than that of Ubicomp-16 and 30% higher than that of MLHC-19. Moreover, our solution has a much smaller standard deviation. These results demonstrate that our solution is more efficient in food type recognition and more stable for data splitting.

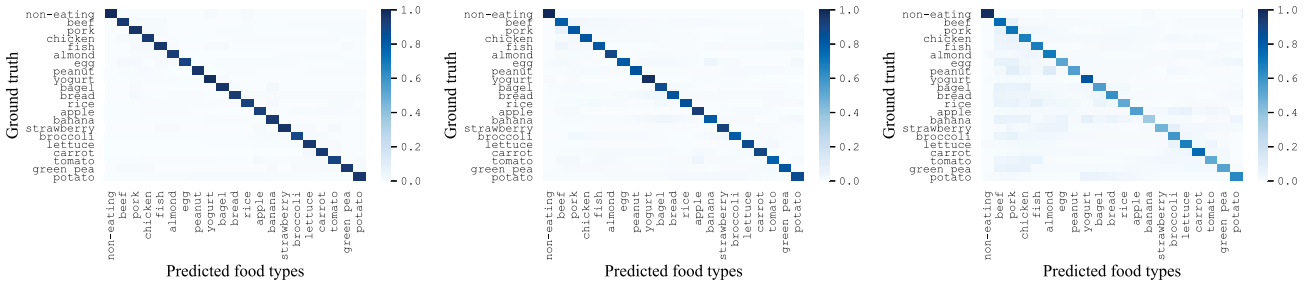
To show the classification results in more detail, we plot the confusion matrixes of these three algorithms and show them in figure 5. From this figure, we observed that our solution achieves good performance on every type, and there is much less misclassified samples than the other two solutions. On the contrary, Ubicomp-16 has more misclassified samples in beef, pork, bagel, apple, and banana. Some pork and fish are recognized as beef, and some eggs are misclassified as banana. For MLHC-19, we can find many misclassified samples among all the food types. In summary, our solution performs much better in food type recognition compared to the two baselines.

2) Self-Evaluation: We trained and tested the model on each user's data in the self-evaluation. Similar to cross-validation, we conducted ten-fold cross-validation 20 times, and the results are shown in figure 4. In this figure, as the data variance is larger, we also showed the first and third quartiles in the boxplot. We observed the performance of our solution is higher than the other two baselines for every user. For the average f1-score of every user, our solution can reach 95.9%. Ubicomp-16 has a similar performance, 93.2%, and MLHC-19 only has 78.1%.

Specifically, for user 1, the classification is more complex because there is more background noise. Our solution achieves an f1-score of 91.7%, while Ubicomp-16 and MLHC-19 only have 85.1% and 63.8%, respectively. Compared with the average performance of all users, the performance of our solution is still good, while the other two solutions dropped too much. It indicates that our solution is more reliable in a natural environment than the two baselines.

B. Impact of sensor fusion

To determine the impact of different types of sensors, we evaluated our proposed method with varying fusions of sensor data, thus the features were extracted from audio data, motion data, or both. The cross-evaluation results are shown in figure 6. The average f1-score of audio data is only 79.2%, while



(a) Our (b) Ubicomp-16 (c) MLHC-19

Fig. 5: Confusion matrixs of cross-evaluation

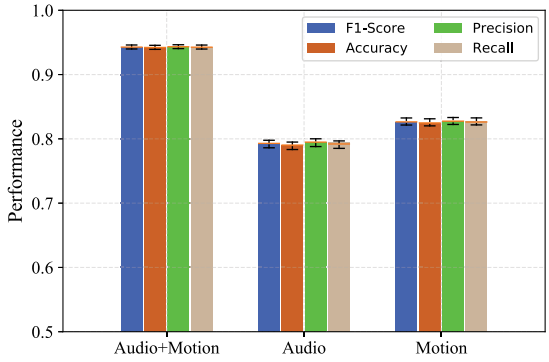


Fig. 6: Results of different sensor's data

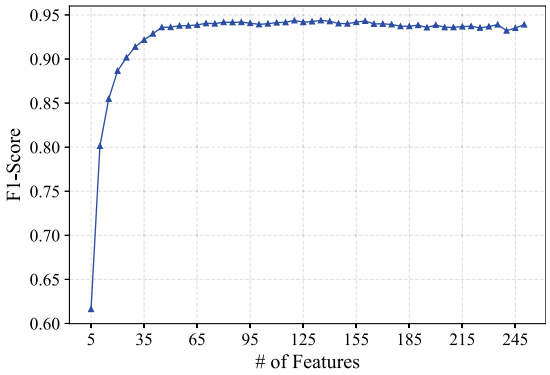


Fig. 7: Performance with different number of features

that of motion data achieves 83.7%. A possible reason is that the participants have relatively similar movement when eating the same type of food. Compared with solely using audio or motion data, the combination of both data achieves an f1-score of 94.3%, which has at least 10% improvement. The result demonstrates that the combination of audio and motion sensor data does enhance the performance. On the other hand, the f1-score of only using motion data is also good, 83.7%. The energy consumption of motion sensors is relatively small due to the low sampling rate. Therefore, only using motion sensors for food type recognition also is a good choice if the device can not provide too much energy.

C. Feature selection analysis

The second question we would like to answer is how many features could achieve the best performance. First, we sort the 249 features by importance. Then, we gradually select the features from the most important to the least important. When the number of selected features increases, the performance of our proposed method is shown in figure 7. The performance increases sharply from no more than 65% to higher than 90% when the number of features increases from 5 to 35. Then the performance fluctuates within a small range when the number is larger than 35. The performance achieves the highest when the number of features is 135. Therefore, we select the most important 135 features for classification.

D. Classifier selection analysis

The last question we would like to answer is which classifier is the best for food type recognition. We present the evaluation results of multiple widely used classifiers, including Logistic Regression, Naive Bayes, SVM, KNN, RF, and lightGBM. The evaluation is based on the selected 135 features, and the results are shown in table III. From this table, we can see that the tree-based classifiers (RF & lightGBM) perform better than others. Specifically, LightGBM outperforms all other algorithms. Therefore, we choose lightGBM as the classifier for our evaluation.

TABLE III: Results of different classifiers

	f1-score
Logistic Regression	26.9%
Naive Bayes	45.5%
SVM	80.9%
KNN	89.1%
Random Forest	92.1%
lightGBM	94.3%

V. Related Work

There have been many research works on food type recognition using wearable devices. These methods can be categorized by the sensors, including microphones, accelerometers, gyroscopes, etc.

The microphone is the most commonly used sensor for food type recognition. Amft et al. placed a microphone in the inner ear to record the sound while eating to classify four different food types, including apple, chips, pasta, and lettuce [9]. The accuracy of each food type is between 80% to 100%, which demonstrates the chewing sound could be used to recognize food types. Later, they used a customized earpad device embeded with a microphone to recognize three food types with an accuracy of 94% [11]. By using the same device, they achieved an accuracy of 86.6% for four food types [12] and 80% for 19 food types [10]. PÍaSSler et al. presented a hearing aid package that integrates two microphones, an in-ear microphone, and a reference microphone, which are used to record the acoustic signals from bone and environmental sounds. It can classify seven food types and one drink with an accuracy of 79%. In addition to the ear, microphones could also be placed in other places. For example, it is placed near the mouth to record the sound of chewing [20], which is used to classify six types of food. Bi et al. developed a prototype that attaches the microphones in the neck to record the sound from the throat area [5], which achieves an accuracy of 84.9% for classifying seven types of food, including apple, carrot, chips, cookies, peanut, walnut, and water. Similarly, another prototype that places the microphones in the neck is used to classify 12 activities, such as eating, drinking, speaking, etc [21]. Although the main objective of this work is not for food type recognition, it can indirectly recognize two food types, cookies and bread. Kalantarian et al. used a Samsung smartwatch to identify four activities from background noise, including eating apples, eating chips, drinking water, and speaking [22]. Except for specific food types, the earbuds with microphones are also used to classify whether the subject is eating hard food, soft food, drinking, or speaking [23], [24].

Motion sensors are also used in classifying food types. Wang et al. developed a headband embedded with an accelerometer and gyroscope to sense the mastication dynamics while eating [7], which can recognize 20 food types with an accuracy of 82.3%. Kim et al. embedded a tri-axial accelerometer in a wrist-worn prototype to identify 29 actions while eating [25], which indirectly infers two types of food, rice and noodle. Moreover, the motion sensors can be combined with other types of sensors. A sub-centimeter scale device that integrates an accelerometer and temperature sensor is put into the mouth while eating. It can classify five food categories, which contain nine types of food [8]. Mirtchouk et al. used a customized earbud with internal and external microphones, an LG G watch, and Google glasses to recognize 40 types of food [14]. To address the challenge of collecting labeled data in free-living environments, they proposed a hierarchical classification algorithm where the classified result is a category or specific food type, and the overall accuracy is 88% [16].

The above methods have made significant progress on this research topic. However, most of the devices they used are customized prototypes, which are not socially acceptable enough for long-term daily usage. For example, even if the Google glasses and LG smartwatches have been used to recognize 40 food types, these devices need to cooperate with another customized earbud [14], and the user needs to wear three devices at the same time. In contrast, our proposed method solely uses Airpods Pro for food type recognition, which is more socially acceptable. Likewise, the Samsung smartwatch is a widely acceptable device but can only indirectly infer two food types [22], and our solution can recognize 20 types of food.

In addition to food type recognition, eating events detection is another important topic in automatic dietary monitoring. An electroglottography (EGG) device is embedded in a necklace to detect swallowing [26], where the EGG sensor is good at measuring the vocal vibration degree. A pair of eyeglasses equipped with Electromyography (EMG) is used for eating detection by monitoring the dynamics of temporalis muscles [27]. Bi et al. placed a contact microphone behind the ear to capture the chewing sound that passes through the bone [28]. By embedding in the necklace, the proximity sensor can be used to detect eating events by measuring the distance between itself and the jawbone [29], [30]. Similar to food type recognition, the above devices are customized prototypes and are not socially acceptable enough compared with our method.

VI. Discussion and Future Work

In this paper, we demonstrated that our solution could perform well in classifying 20 types of food, which uses Airpods Pro. The socially acceptable device makes long-term wear possible. Moreover, the evaluation results show that the combination of both audio and motion data outperforms using only audio or motion data.

In our study, some food types are cooked in a similar method as others. For example, all the beef, pork, chicken, and fish in our experiment are boiled. However, these meat may be cooked in different methods in our daily cuisine, such as steak, fried chicken, steamed fish, etc. Different cooking methods result in different food properties, such as hardness, fracturability, and size. Therefore, the chewing sound of the same food type would be different. Our future work will explore the impact of the variety of cooking methods.

We evaluated our study in the apartment's dining room. However, people may choose to have meals in a restaurant or dining hall, where there is much more background noise. Moreover, they may talk and drink during eating. In future works, we will investigate the solutions dealing with filtering out the chewing sound and handling these complex background noise activities.

VII. Conclusion

In this paper, we propose a food type recognition method that uses a socially acceptable device, the Airpod Pro, to recognize 20 different types of food. As far as we know, we are the first to solely use a socially acceptable commercial product to recognize food types. The data from audio and motion sensors are collected when the earbuds are deployed in the left and right ears. We extracted 105 features from audio data

and 144 features from motion data. Then we used lightGBM to evaluate the importance of each feature and selected the most important 135 features. We conducted the data collection in an apartment's dining room. The experiment includes five male and two female human subjects. We chose lightGBM as the classifier and optimized its parameters. The results show that the average f1-score reaches 94.4% for the ten-fold cross-validation test and 96.0% for the self-evaluation test.

REFERENCES

[1] C. for Disease Control and Prevention, "Adult obesity facts," 2022. [Online]. Available: https://www.cdc.gov/obesity/data/adult.html

[2] ——, "National diabetes statistics report," 2022. [Online]. Available: https://www.cdc.gov/diabetes/data/statistics-report/index.html

[3] J. X. Moore, N. Chaudhary, and T. Akinyemiju, "Peer reviewed: Metabolic syndrome prevalence by race/ethnicity and sex in the united states, national health and nutrition examination survey, 1988–2012," *Preventing chronic disease*, vol. 14, 2017.

[4] C. for Disease Control and Prevention, "Heart disease facts," 2022. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm

[5] Y. Bi, M. Lv, C. Song, W. Xu, N. Guan, and W. Yi, "Autodietary: A wearable acoustic sensor system for food intake recognition in daily life," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 806–816, 2015.

[6] S. Zhang, D. Nguyen, G. Zhang, R. Xu, N. Maglaveras, and N. Alshurafa, "Estimating caloric intake in bedridden hospital patients with audio and neck-worn sensors," in *2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2018, pp. 1–2.

[7] S. Wang, G. Zhou, J. Guan, Y. Ma, Z. Liu, B. Ren, H. Zhao, A. Watson, and W. Jung, "Inferring food types through sensing and characterizing mastication dynamics," *Smart Health*, vol. 20, p. 100191, 2021.

[8] K. S. Chun, S. Bhattacharya, C. Dolbear, J. Kashanchi, and E. Thomaz, "Intraoral temperature and inertial sensing in automated dietary assessment: a feasibility study," in *Proceedings of the 2020 International Symposium on Wearable Computers*, 2020, pp. 27–31.

[9] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, "Analysis of chewing sounds for dietary monitoring," in *International Conference on Ubiquitous Computing*. Springer, 2005, pp. 56–72.

[10] O. Amft and G. Troster, "On-body sensing solutions for automatic dietary monitoring," *IEEE pervasive computing*, vol. 8, no. 2, pp. 62–70, 2009.

[11] O. Amft, M. Kusserow, and G. Troster, "Bite weight prediction from acoustic recognition of chewing," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 6, pp. 1663–1672, 2009.

[12] O. Amft, "A wearable earpad sensor for chewing monitoring," in *SENSORS, 2010 IEEE*. IEEE, 2010, pp. 222–227.

[13] S. Päßler, M. Wolff, and W.-J. Fischer, "Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food," *Physiological measurement*, vol. 33, no. 6, p. 1073, 2012.

[14] M. Mirtchouk, C. Merck, and S. Kleinberg, "Automated estimation of food type and amount consumed from body-worn audio and motion sensors," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 451–462.

[15] M. Mirtchouk, D. Lustig, A. Smith, I. Ching, M. Zheng, and S. Kleinberg, "Recognizing eating from body-worn sensors: Combining free-living and laboratory data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–20, 2017.

[16] M. Mirtchouk, D. L. McGuire, A. L. Deierlein, and S. Kleinberg, "Automated estimation of food type from body-worn audio and motion sensors in free-living environments," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 641–662.

[17] F. Espósito, "Apple leads the headphone market in the us with airpods and beats," 2022. [Online]. Available: https://9to5mac.com/2022/02/10/apple-leads-the-headphone-market-in-the-us-with-airpods-and-beats/

[18] U. D. of Agriculture, U. D. of Health, and H. Services, *Dietary Guidelines for Americans: 2020-2025*. DietaryGuidelines.gov, 2020.

[19] S. Basu, *Near-field radiative heat transfer across nanometer vacuum gaps: fundamentals and applications*. William Andrew, 2016.

[20] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. E.-D. Mousa, and B. Schuller, "I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on asr performance," *PloS one*, vol. 11, no. 5, p. e0154486, 2016.

[21] K. Yatani and K. N. Truong, "Bodyscope: a wearable acoustic sensor for activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 341–350.

[22] H. Kalantarian and M. Sarrafzadeh, "Audio-based detection and evaluation of eating behavior using the smartwatch platform," *Computers in biology and medicine*, vol. 65, pp. 1–9, 2015.

[23] M. Shuzo, G. Lopez, T. Takashima, S. Komori, J.-J. Delaunay, I. Yamada, S. Tatsuta, and S. Yanagimoto, "Discrimination of eating habits with a wearable bone conduction sound recorder system," in *SENSORS, 2009 IEEE*. IEEE, 2009, pp. 1666–1669.

[24] Y. Gao, N. Zhang, H. Wang, X. Ding, X. Ye, G. Chen, and Y. Cao, "ihear food: eating detection using commodity bluetooth headsets," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2016, pp. 163–172.

[25] H.-J. Kim, M. Kim, S.-J. Lee, and Y. S. Choi, "An analysis of eating activities for automatic food type recognition," in *Proceedings of the 2012 asia pacific signal and information processing association annual summit and conference*. IEEE, 2012, pp. 1–5.

[26] M. Farooq, J. M. Fontana, and E. Sazonov, "A novel approach for food intake detection using electroglottography," *Physiological measurement*, vol. 35, no. 5, p. 739, 2014.

[27] R. Zhang and O. Amft, "Monitoring chewing and eating in free-living using smart eyeglasses," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 23–32, 2017.

[28] S. Bi, T. Wang, N. Tobias, J. Nordrum, S. Wang, G. Halvorsen, S. Sen, R. Peterson, K. Odame, K. Caine *et al.*, "Auracle: Detecting eating episodes with an ear-mounted sensor," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–27, 2018.

[29] K. S. Chun, S. Bhattacharya, and E. Thomaz, "Detecting eating episodes by tracking jawbone movements with a non-contact wearable sensor," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 1, pp. 1–21, 2018.

[30] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa, "Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 2, pp. 1–26, 2020.