applied sciences

MDPI

Article

Lip-Reading Classification of Turkish Digits Using Ensemble Learning Architecture Based on 3DCNN

Ali Erbey [1,2,*] and Necaattin Barışçı [3]

[1] Department of Computer Programming, Distance Education Vocational School, Usak University, Usak 64200, Türkiye
[2] Information Systems, Informatics Institute, Gazi University, Ankara 06560, Türkiye
[3] Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara 06560, Türkiye; nbarisci@gazi.edu.tr
* Correspondence: ali.erbey@usak.edu.tr

Abstract: Understanding others correctly is of great importance for maintaining effective communication. Factors such as hearing difficulties or environmental noise can disrupt this process. Lip reading offers an effective solution to these challenges. With the growing success of deep learning architectures, research on lip reading has gained momentum. The aim of this study is to create a lip reading dataset for Turkish digit recognition and to conduct predictive analyses. The dataset has divided into two subsets: the face region and the lip region. CNN, LSTM, and 3DCNN-based models, including C3D, I3D, and 3DCNN+BiLSTM, were used. While LSTM models are effective in processing temporal data, 3DCNN-based models, which can process both spatial and temporal information, achieved higher accuracy in this study. Experimental results showed that the dataset containing only the lip region performed better; accuracy rates for CNN, LSTM, C3D, and I3D on the lip region were 67.12%, 75.53%, 86.32%, and 93.24%, respectively. The 3DCNN-based models achieved higher accuracy due to their ability to process spatio-temporal data. Furthermore, an additional 1.23% improvement was achieved through ensemble learning, with the best result reaching 94.53% accuracy. Ensemble learning, by combining the strengths of different models, provided a meaningful improvement in overall performance. These results demonstrate that 3DCNN architectures and ensemble learning methods yield high success in addressing the problem of lip reading in the Turkish language. While our study focuses on Turkish digit recognition, the proposed methods have the potential to be successful in other languages or broader lip reading applications.

Keywords: lip-reading; ensemble learning; 3DCNN

check for updates

Academic Editor: Andrea Prati

Received: 6 December 2024
Revised: 1 January 2025
Accepted: 6 January 2025
Published: 8 January 2025

Citation: Erbey, A.; Barışçı, N. Lip-Reading Classification of Turkish Digits Using Ensemble Learning Architecture Based on 3DCNN. *Appl. Sci.* **2025**, *15*, 563. https://doi.org/10.3390/app15020563

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Lip reading, also known as visual speech recognition, is the ability to understand a word or phrase in one's lip movements [1]. With the lip-reading method, individuals with hearing difficulties can better understand the other side. Lip reading is one of the elements that strengthen communication not only for individuals with hearing impairment, but also for individuals without hearing impairment [2]. Individuals unconsciously approve of the correctness of the word heard from the lip movements of people, especially in noisy environments. In this context, lip reading is important for the continuity of communication for both hearing impaired and normal hearing people. At the same time, lip reading is one of the human-computer interaction tools [3]. The realization of language-specific lip-reading studies emerges as a necessary element in the communication phase with the

machines with the advancement of technology. It can be seen as a necessary element for the processes of communication in human computer interaction.

People unconsciously tend to extract words from the movements of the lips. This approach is purely due to a visual implication. It is seen in the studies that this process, which is dependent on visual extraction, can be analyzed with image processing techniques. In the analysis phase, more and more successful results are obtained with deep learning methods [4]. When lip-reading studies using deep learning methods are examined, it is seen that the studies focus mainly on the English language [5–8]. At the same time, it is seen that there are many data sets for lip reading in English in order to be able to do these studies. A review of the literature reveals that while there are numerous datasets available in English, the number of datasets in Turkish is significantly lower. This highlights the need for a dataset specifically designed for Turkish lip-reading studies.

Lip-reading, or visual speech recognition, is crucial for individuals with hearing impairments and in noisy environments where auditory information is insufficient. Despite advancements in deep learning, there is a significant gap in lip-reading research for non-English languages, particularly Turkish. This study aims to address this gap by creating a novel dataset for Turkish digit recognition and utilizing various deep learning architectures as benchmarks for evaluation.

In this study, first of all, we present a novel dataset specifically designed for lip-reading in Turkish, a language with unique phonetic characteristics. This dataset, focusing on digit recognition, fills a significant gap in the current literature which predominantly concentrates on English. The different structures of languages make it difficult for a model that achieves good in one language to achieve similar results when applied to another language. The unique phonetic characteristics of the Turkish language pose additional challenges, making this study particularly significant. Turkish, on the other hand, is even more different from commonly used languages such as English as an agglutinative language. In this context, deep learning models, which have achieved successful results in the literature in recent years, have been experimentally studied on the Turkish data set created in this study. The successes of the 2DCNN, LSTM and 3DCNN models, which are widely used in other languages, for the Turkish language were analyzed comparatively. Additionally, one of the novelties of the article is the use of ensemble learning, aiming to enhance performance by combining various models. In this study, the models were combined using a majority voting method, implementing an ensemble learning approach that has not been explored in previous literature. This innovation represents a significant step in Turkish lip reading and serves as a potential guide for future studies on different languages and datasets. By evaluating various deep learning architectures, we aim to establish a benchmark for future research in Turkish lip-reading. In the next part of this study, lip reading studies in the literature are included, and in the third part, the stages of creating a data set consisting of numbers for Turkish, the preprocessing stage of the data set and its training are explained. In the Section 4, the results are discussed.

2. Related Works

With technological advances, cameras have become smaller and have been integrated into even mobile phones. This integration allows us to record whenever we want. In line with these developments, more and more data are obtained day by day. In this way, the ever-increasing volume of visual data drives more research in image processing and better results are obtained. In the field of image processing, these studies drive advancements in areas such as autonomous vehicles and robotic systems. At the same time, with the emergence of deep learning architectures in the field of image processing, especially in recent years, the successes in these areas are increasing exponentially.

The studies on image processing not only pioneer new technologies but also serve the purpose of improving human life, which is the fundamental meaning of technology. In this context, studies such as sign language recognition [9] and lip reading [10,11] aim to enhance the lives of individuals with disabilities. Lip-reading studies, which aim to enhance human life, have emerged as a crucial area of research in image processing, especially since hearing aids are less effective in noisy environments. Although a large number of data sets have been created in lip reading studies that have been studied for many years, it is seen that more studies have been done on certain data sets. The data sets created for lip reading can have different contents. It is seen that the content of the datasets is structured as letters, words, numbers, sentences or a mixture of these.

The OuluVs2 dataset, contains 10 digit sequences repeated 3 times by 52 speakers [12]. The dataset has only mouth regions. GRID, one of the most commonly used datasets, consists of sentences. The sentences in the GRID dataset are formed according to a certain structure. This sentence structure, in which 34 speakers repeat 1000 times, consists of command, color, letter, number and adverb. One of the example sentences is 'lay green by A 4 now' [13]. Table 1 shows the content, class numbers, sample numbers, resolutions, shooting angle and language details of the datasets, which are seen to be created in different structures. Fields marked with + in the table represent the content of the relevant column in the content of that dataset, while fields marked with - indicate that the content of the relevant column is not represented in the content of that dataset. For example, for the AvLetters dataset created in 2002, the Alphabet was used, but word and sentence structures were not used in the dataset.

Table 1. Data Sets [14].

Year	Dataset	Alphabet	Word	Sentence	Class	Resolution	Angle	Language
2002	AvLetters	+	-	-	26	376 × 288	Front	English
2004	CUAVE	+	-	-	10	720 × 480	Front + Side	English
2004	AVICAR	+	-	+		720 × 480	Front + Side	English
2006	Grid	-	-	+	51	720 × 576	Front	English
2009	OuluVs	-	-	+	10	720 × 576	Front	English
2015	OuluVs2	-	+	+	540	1920 × 1080	Front	English
2016	LRW	-	+	-	500	256 × 256	Front + Side	English
2018	LRS2-BBC	-	-	+		224 × 224	Front + Side	English
2019	LRW-100	-	-	-	1000	256 × 256	Front + Side	English

When the data sets in Table 1 are examined, it is seen that the data sets are generally in English. Although most of the studies are on English, it is seen that there are studies in other languages. Fu et al. [15] created their own dataset on Chinese and worked with ShuffleNet and CBAM. In another lip-reading study on Chinese Mandarin, ref. [16] used a data set called CMLR collected on Chinese television. Ma et al. [17], on the other hand, reached state of the art on the Chinese Mandarin dialect with a Transformer-based model. Shin et al. [18] achieved a success rate of 92.67% with a system they proposed with machine learning algorithms on the Korean language. Data sets in other languages such as Spanish have also been created [19,20] and related lip-reading studies have been conducted.

When lip-reading studies are examined, it is seen that many studies have been carried out on the OuluVs2 data set consisting of numbers [21,22]. Petridis et al. [22] used OuluVs2 and different datasets and achieved very high performance using Bidirectional LSTM. Mesbah et al. [21] achieved 93.72% classification success in the OuluVs2 dataset. When Assael et al. [23] worked on the GRID dataset, it was seen that they reached 95.2% accuracy.

It is seen that more than one convolutional neural network is used in the models in lip reading studies [24]. Spatio-temporal data between video frames with multiple

uses of convolutional neural networks can also be included in the evaluation. One of the spatio-temporal evaluation methods is the application of the 3D convolution operation. In the 3D convolution operation, each frame can be evaluated with the frame in front of it.

A review of lip-reading studies using 3DCNN models in the literature reveals high success rates [25,26]. It is seen that [27] achieved state-of-the-art in their proposed model by combining the Transformer model and the 3DCNN structure. At the same time, it is seen that 3DCNNs used extensively in hybrid models in lip reading studies [25].

3. Material and Method

3.1. Dataset

In this study, we developed a dataset specifically for Turkish digit recognition through lip reading. The dataset consists of videos where participants spoke digits from 0 to 9, with each digit repeated five times. This structure results in a dataset containing 10 distinct classes. Additionally, the dataset includes an extended subset featuring words beyond digits, enabling broader experimentation. The data was collected from 78 participants, representing a diverse set of individuals. The data collection procedures were carried out with the approval of the Usak University Social and Humanities Scientific Research and Publication Ethics Committee with the approval number 2024-208. Signed informed consent was obtained from all participants during the data collection phase.

Participants ranged in age from 20 to 55 years, and each participant repeated each digit five times, yielding a total of 3900 video clips. The videos were recorded using an iPhone at a resolution of 1080 × 1920 and a frame rate of 30 frames per second (fps). The videos were saved in MOV format. Preprocessing steps included frame extraction and isolating the lip regions using the RetinaFace [28] method.

In this study, rotation was applied as a data augmentation technique to increase the diversity and robustness of the dataset. The video frames were rotated 10 degrees to the left and 10 degrees to the right to increase the robustness of the dataset by preserving the natural and unique features of the lip movements. Data augmentation was performed before splitting the dataset. However, since the dataset was grouped according to participants, the order of data splitting has no significance for the risk of data leakage. This preserves the integrity and reliability of the experimental process and supports the validity of the results.

Figure 1 presents demographic details, including the age and gender distribution of the participants. The data set consists of 57 male and 21 female participants. All participants are of Turkish origin, and the dataset comprises a total of 3900 video files, representing the 78 participants each reciting 10 digits five times.

As can be seen in Figure 1, while selecting the sample, on the one hand, generalization was considered, but in order for the system to achieve high performance of the model, we aimed to distribute the age limit between 20 and 55. In the study of Wong et al. it was stated that lip movements weaken with advancing age [29]. For this reason, the data were obtained in a right-skewed distribution. This situation creates a difference in age distribution in sample selection. Similarly, Schötz et al. examined the effect of variability in lip movements on motor control skills and found a negative correlation between age and these skills, which is in parallel with our sample selection [30].

Figure 2 illustrates sample frames from the videos, demonstrating of participants' lip movements during the pronunciation of the digits. The videos were recorded from the front angle, ensuring a consistent view of the participants' faces.

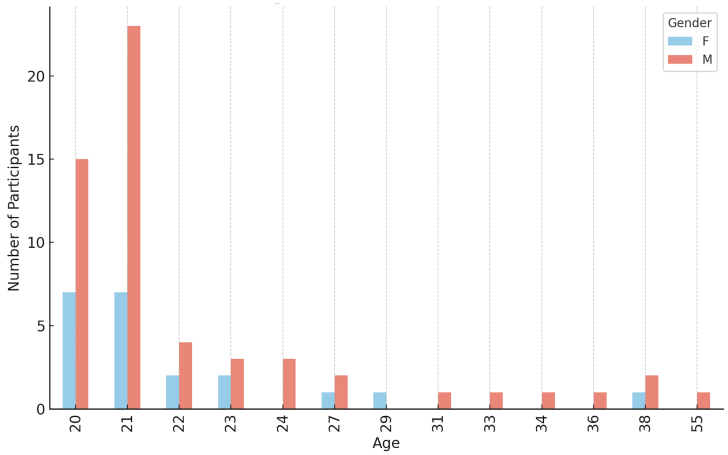


Figure 1. Demographic details.

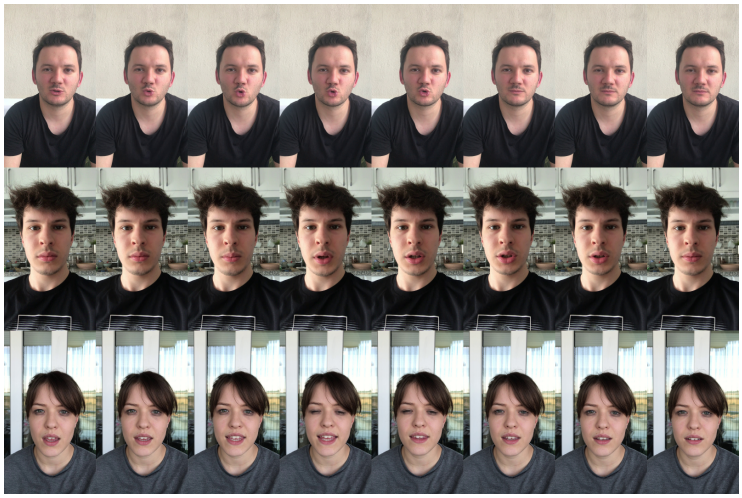


Figure 2. Frames of people at the top, middle and bottom saying the numbers 1 (bir), 7 (yedi), 3 (üç) respectively.

The dataset was further processed by cutting 420 pixels from the top and bottom of each frame to focus specifically on the facial and lip regions. The final frame resolution for each video was standardized at 1080 × 1080 pixels.

Table 2 provides detailed specifications of the dataset, including participant demographics, video resolution, and frame rate.

Table 2. Dataset Information.

Feature	Details
Dataset Type	Front face and lip region
Number of Participants	78
Age Range of Participants	20–55 years
Spoken Digits	Digits (0–9)
Repetitions per Digit	5
Total Video Clips	3900
Recording Device	iPhone
Resolution	1080 × 1920
Frame Rate	30 fps
Top and Bottom Cut (pixels)	420 pixels
Final Frame Resolution	1080 × 1080

This dataset provides a valuable resource for the development of deep learning models focused on lip reading for Turkish, filling a gap in the current literature that is heavily

dominated by English-language datasets. The use of data augmentation and preprocessing ensures variability and enhances the dataset's applicability for advanced experimentation.

3.2. Dataset Preprocessing

Preprocessing involves preparing data for use in a machine learning method. This stage includes cleaning and editing the data and formatting the data for the model. Preprocessing is a part of machine learning that is usually performed before the data is analyzed, and includes the preparation of the data.

In this study, two subsets was created, consisting of the front face and the lip region. The dataset was collected by recording 78 participants, aged between 20 and 55, speaking digits from 0 to 9 in Turkish. Each participant repeated each digit five times, resulting in a total of 3900 video clips. The recordings were made using an iPhone with a resolution of 1080 × 1920 and a frame rate of 30 fps.

During the preprocessing phase of Dataset1 preparation, the videos were converted to frame format for convolution by making top and bottom cuts. Specifically, 420 pixels were cut from the top and bottom of the videos to create frames with a resolution of 1080 × 1080. The cutting area is shown in Figure 3.

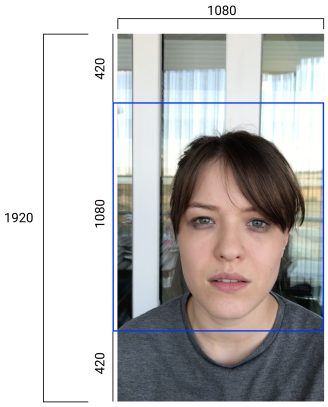


Figure 3. Trimming during data preprocessing.

As shown in Figure 3, after cropping, the videos were resized to a frame format of 1080 × 1080. After this stage, the data was resized to 224 × 224 to reduce training time and computational costs associated with training of models. During this sampling process, the total number of frames of the video was sampled at equal intervals. By preserving the temporal context, multiple video data can be extracted by taking certain segments from long videos in a sequential order. However, this was not done for the digit dataset. This is due to the fact that the average length of each video for spoken digits is typically less than 1 s and the pronunciation of digits in Turkish is naturally short.

In deep learning studies, data augmentation techniques are employed to increase the amount of data. One of the data augmentation methods for two-dimensional images is affine transformation methods, which include operations such as rotation, scaling, flipping, and transformation [31]. However, for video data, sampling methods from video frames are used instead. In this study, specific frames were sampled from each video file. The frames in the videos were rotated 10 degrees to the right and 10 degrees to the left as part of the data augmentation process, resulting in a threefold increase in the total amount of data. The number of data after data augmentation is shown in Table 3.

Table 3. Number of videos in original and augmented datasets.

Phase	Video Count
Original Dataset	3900
After Data Augmentation (Training)	11,700

Detection and Extraction of Lip Regions

In the data set created for DataSet2, the lip region was extracted. The dataset was then extracted with RetinaFace [28] to focus only on the lip area. RetinaFace is a deep learning-based object detection algorithm for face detection and landmarks identification. Beyond verifying the existence of a face, the method is capable of detecting critical regions on the face (e.g., eyes, nose, mouth and lips) with high accuracy. RetinaFace provides a powerful solution for the precise detection of lip regions in particular.

Using a ResNet-50 based architecture, RetinaFace is able to perform both face detection and detection of key points on the face at the same time. In this study, the center coordinates of the lips are determined using keypoints generated by RetinaFace to extract lip regions during the face detection process. During this process, the central coordinates of the lips of the face are taken as the basis and a square area is extracted around this central point, 112 pixels in both directions horizontally and 112 pixels in both directions vertically. Isolation of the lip region was performed by standardizing the frame sizes to 224 × 224 pixels, so that the original size of 1080 × 1080 pixels was reduced to 224 × 224 pixels. This dimensioning reduces the computational burden of the model while at the same time ensuring that the details of the lip movements are not lost. Figure 4 shows the extraction process from the lip region.

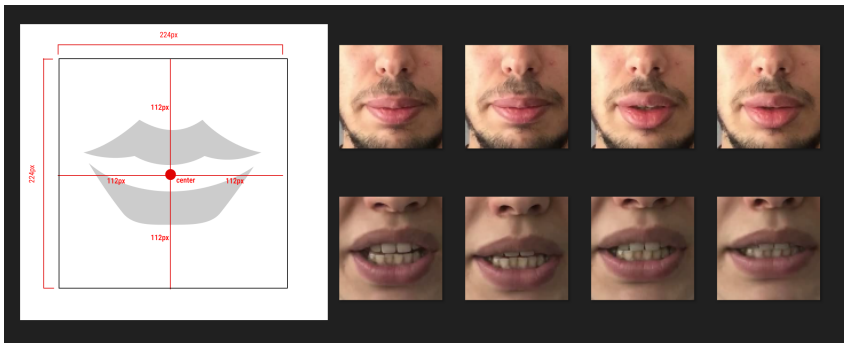


Figure 4. Extracting the lip area during data preprocessing.

After extracting data from the videos, data augmentation was performed by rotating the data 10 degrees to the left and 10 degrees to the right. In this study, data augmentation was applied using rotation only to increase the variability and robustness of the dataset. Rotation was chosen as the augmentation technique because it effectively increases the variability of the dataset without compromising the natural features of lip movements. With these operations, Dataset1 and Dataset2 were ready for deep learning models before the training, testing and validation of the data.

Grouped 5-fold cross-validation was used to divide the data into training, validation and test sets. Grouped 5-fold cross-validation ensures a fair evaluation for problems such as lip reading by preventing the same data from being present in both training and test sets. The dataset is divided into five groups according to their folder structure and in each iteration, one folder is assigned as the test set, while the remaining folders are used for training and validation. The dataset is first divided into a training set of 80% and a test set of 20%. A validation set of 20% was separated from the training set to be used in processes such as hyperparameter optimization and early stopping of the model. As a result of this

splitting strategy, the total data set is organized as 64% training, 16% validation and 20% test set. The validation set was used only in the training process to evaluate and optimize the hyperparameter settings. The test set is reserved to measure the final performance of the model, completely independent of the validation process.

3.3. Deep Learning Architectures

The subject of lip reading, which has been studied for many years in the field of image processing, was initially tried to be realized with machine learning algorithms, but today it is tried to be realized with deep learning architectures [32]. When we look at the deep learning architectures used in lip reading studies, it is seen that convolutional neural networks are used extensively. In lip-reading studies, it is seen that both 2DCNN-based methods, 3DCNN-based methods, and 2DCNN and 3DCNN-based models are used together [33]. Models based on 3DCNN seem to handle spatio-temporal problems in a more qualified manner. On the other hand, 2DCNN models are seen to be better in fine-grained operations in feature extraction. In the use of 2DCNN models for video data, concatenating the frames into a single frame is one of the methods applied in order not to lose the time context.The choice of models (CNN, LSTM, 3DCNN) was based on their proven effectiveness in previous lip-reading studies [4,32]. While CNNs are effective for spatial feature extraction and LSTMs handle temporal dependencies well, 3DCNNs capture spatio-temporal features better which makes them suitable for lip reading task. The selection of models was based on successful examples in the literature, with consideration given to the suitability of each model (CNN, LSTM, C3D, I3D) for the data type.

3.3.1. Convolutional Neural Networks

CNNs are generally used in computer vision applications and are widely used in deep learning studies such as face recognition and image recognition. Especially in the field of image processing, the most used model is CNN models. CNN outperforms previous models in image processing tasks through its effective image inference capabilities [34]. CNNs use more than one convolutional layer for feature extraction and use a combined set of operations such as pooling layer and activation layer for inference [35]. Convolution layer is the main feature of CNNs. The advantage of CNN-based architectures is that it does not require manual feature extraction [36].

3.3.2. Long Short Term Memory

Long-Short term memory is seen as an advanced model of recurrent neural networks. Recurrent neural networks process the incoming data at once and remember the previous step [37]. RNNs are also an effective model for sequential tasks, and can give good results even in non-linear structures thanks to remembering previous information [38]. However, Hochreiter and Schmidhuber [39] suggested LSTM to eliminate the existing Vanishing Gradient problem and deficiencies in RNNs. Thus, the existing problems in RNNs were resolved with LSTM.

3.3.3. 3 Dimensional Convolutional Neural Networks

3DCNN stands for 3D Convolutional Neural Network. It is a deep learning structure created using layers of neurons and used to recognize patterns from data. Although CNNs are successful in image processing, it is seen that 3DCNN models are more successful for spatio-temporal data. The main function of the 3DCNN architecture is to extract deep-level features from the given data [40]. 3DCNN exhibits higher performance than classical convolutional neural network models [41,42], evaluates spatio-temporal data [43], and has better performance compared to long-short-term memory architectures. It is seen that it is especially useful for video data in terms of low processing and memory requirements.

In the 3DCNN architecture, another dimension is added to the filters and step values such as 3 × 3 × 3 or other variations are applied as 1 × 1 × 1 or different variations [43]. 3DCNNs also receive spatio-temporal information in time-dependent images [27]. Thus, 3DCNNs can also deal with temporal data on video data [33,44]. Tran et al. [45] showed that spatio-temporal information in video data can be extracted using 3DCNN architectures more successfully than 2DCNN architectures. In this study, models containing both 2DCNN and 3DCNN were used. One of the 3DCNN models used is the C3D [45] model. The C3D model can obtain more successful results in video analysis, as it can extract spatio-temporal information better.

Another 3DCNN-containing model used in this study is the Two-Stream Inflated 3D ConvNet (I3D) model.The I3D model is presented by researchers from DeepMind and Oxford University in a paper entitled "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset" [46].

3.3.4. Ensemble Learning

One of the most widely used machine learning methods in recent years is ensemble learning. The idea of ensemble methodology, which is a predictive model created by integrating multiple models, is known as a method used to improve prediction performance [47]. Commonly used ensemble learning methods are bagging, boosting and stacking [48]. In bagging, the predictions of different trained models are processed using different methods such as averaging the predictions if it is classification, or averaging the predictions if it is majority voting or regression. In the boosting method, the models focus on making up for each other's shortcomings. In this way, a strong model can be created and difficult problems can be focused on. In the stacking method, the prediction of each trained model is used to train the meta-model. Since the basic models of machine learning are also involved in the stacking method, it can achieve better performance by combining the strengths of these models.

The basic philosophy underlying the methods used in ensemble learning is to use more models. As the variety of models trained from ensemble learning increases, the classification performance of the system also increases. This study aims to improve the performance of the system by using ensemble learning methodology.

4. Results and Discussion

In this study, CNN, LSTM, C3D, I3D, 3DCNN + BiLSTM models were used in the training phase of the data set. The performance of these models in the Turkish numeral dataset was compared. Experiments were performed on DataSet1 and DataSet2 separately. Experiments were performed on NVidia RTX 3060 graphics card using Pytorch deep learning library.

While we can evaluate the time dimension for 3DCNN models during the training phase, the frames are concatenated into a single frame in order to evaluate the time dimension using 2DCNN models. Figure 5 shows the diagram in which the frames are given as input to the model.

In this study, since lip reading involves both spatial (lip positions) and temporal (movement flow) information, 3DCNN architectures (C3D, I3D) were preferred. 3DCNN provides stronger spatio-temporal analysis by preserving connections between frames. This capability is crucial for accurately modeling lip movements.The input of the data and the schematization of the model in 3DCNN models are shown in Figure 6.

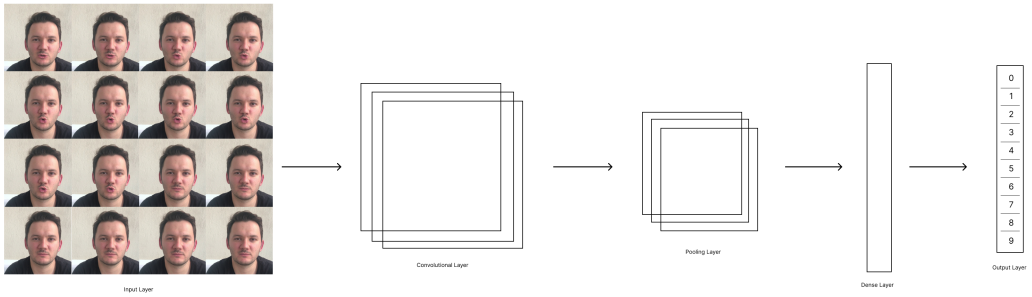


Figure 5. Frame input and model for CNN.

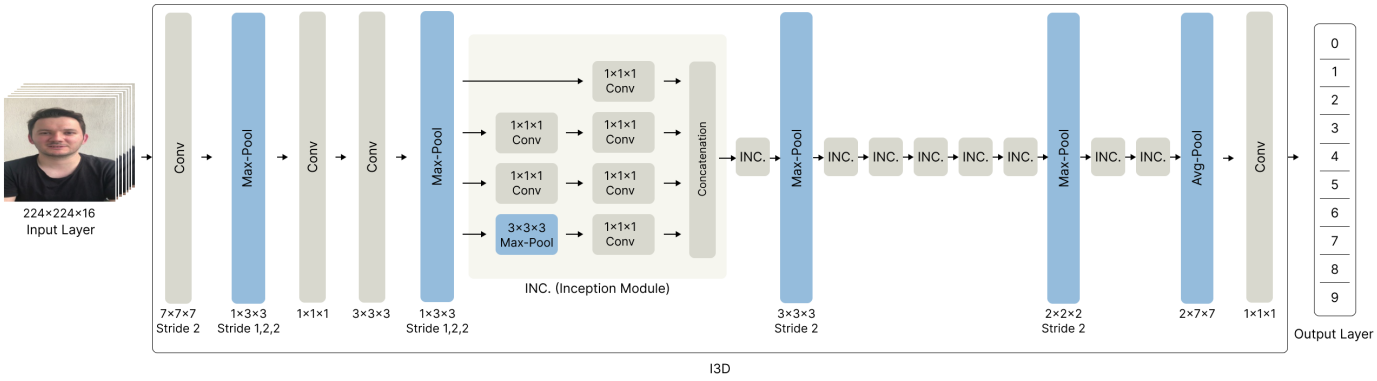


Figure 6. Input for 3DCNN models.

During the training phase of the models, we employed the grouped 5-fold cross-validation method. This approach entails dividing the dataset into five subsets, with training performed on four subsets and validation on the remaining one, rotating through all subsets to ensure a thorough evaluation. The data obtained in the study were collected from 78 people and subjected to training and testing using grouped 5-fold cross-validation. 5-fold cross-validation is a widely used technique in machine learning for evaluating model performance. This method typically involves dividing the dataset into five equal parts (folds), where each fold is successively used as the test set, while the remaining folds are used for training the model. After randomly splitting the dataset, one fold is used as the test set in each iteration, and the other four folds are used for training. This process is repeated five times, with each fold serving as the test set once. The results of each iteration are recorded, and the average performance score is calculated to evaluate the model's overall effectiveness. 5-fold cross-validation helps reduce the risk of overfitting by ensuring that each part of the dataset is used for both training and testing, allowing the model's consistency across different data subsets to be measured. However, this method can become computationally expensive for large datasets due to the need for retraining the model in each iteration.

Although 5-fold cross-validation provides a balanced approach to evaluating a dataset, it may have limitations depending on the structure of the dataset. Specifically, in this study, the use of videos from different individuals may make random splitting less effective for proper evaluation. If the same individual's videos are present in both the training and test sets, the model may learn person-specific information (data leakage), leading to an unfairly inflated performance. To avoid this, the grouped 5-fold cross-validation method should be employed. In this approach, the videos of each individual are grouped within the same fold, ensuring that each iteration uses a different group of individuals for the training and test sets. This guarantees that videos from the same individual do not appear in both the training and test sets, allowing the model to be tested on completely unseen individuals,

thus providing a more accurate measure of generalization. Grouped 5-fold cross-validation is particularly beneficial for datasets like lip-reading, where person-specific data could skew the model's performance. Compared to standard 5-fold cross-validation, this method offers more reliable results. If grouped 5-fold cross-validation is not applied, there is a higher risk of over-fitting due to the potential overlap of individual-specific data across the sets, which may result in a misestimation of the model's generalization performance. Consequently, constructing validation and test sets with data from individuals not previously seen by the model is crucial for an accurate evaluation of model performance.

The impact of hyperparameter values, including the number of epochs, batch size, learning rate, and chosen optimizer, on model performance is well-established during training. The hyperparameters employed in the training of the models, as determined through experimental studies, are presented in Table 4.

Table 4. Model Architecture and Parameters.

Model Name	Epoch	Optimizer	LR	Batch S.	Act. Func.
CNN	50	Adam	0.001	4	ReLU
LSTM	50	Adam	0.001	4	Tanh
C3D	50	SGD	0.01	4	ReLU
I3D	50	Adam	0.001	4	ReLU
3DCNN + BiLSTM	50	Adam	0.001	4	ReLU

The performance of each model was evaluated across DataSet1 and DataSet2, as summarized in Table 5. Notably, the 3D convolution-based models (C3D, I3D, 3DCNN + BiLSTM) consistently outperformed their 2D counterparts (CNN, LSTM), underscoring the importance of preserving temporal information in video data processing.

Table 5. Performance Metrics for Dataset1 and Dataset2.

Model	Dataset1 Accuracy (%)	Dataset2 Accuracy (%)
CNN	52.37	67.12
LSTM	66.51	74.53
C3D	78.82	86.32
I3D	87.17	93.24
3DCNN + BiLSTM	87.72	93.30
C3D + I3D + (3DCNN + BiLSTM)	89.92	94.53

When the data in Table 5 are examined, it is seen that the results for DataSet2 obtained from the lip region are high. Resnet152 weights were used as pre-trained in the training phase of DataSet1 obtained from the face region with CNN. Batch size 4 was selected with a training period of 50 epochs. As a result of the training, 52.37% success was achieved. The training was then carried out within the LSTM and 66.51% success was found. Afterwards, DataSet1 and DataSet2 were trained with C3D and I3D models. Although LSTM architectures are effective in capturing dependencies in time-series data, they lack the capability to jointly process spatial and temporal features in video data. The effects of hyper parameter values such as epoch number, batch size, learning rate and selected optimizer on the performance of the model are known during the training of the model. In experimental studies, a training of 50 epochs was carried out for the training of the C3D model. In the training, a batch size of 4 and a learning rate of 0.01 were used. Stochastic Gradient Descent(SGD) was chosen as the optimizer. As a result of this training, an accuracy rate of 78.82% was found for DataSet1. As a result of training the C3D model with DataSet2, 86.32% accuracy rates were found. A training of 50 epochs was carried out for the training of the I3D model. In the training, a batch size of 4 and a learning rate of 0.001 were used.

Adam was chosen as the optimizer. As a result of this training, accuracy rates of 87.17% for DataSet1 and 93.24% for DataSet2 were found.

Upon analyzing the obtained results, it is observed that the 3DCNN + BiLSTM model achieved the highest performance among the I3D, C3D, and 3DCNN + BiLSTM models. The 3DCNN + BiLSTM model reached an accuracy rate of 87.72% for Dataset1 and 93.30% for Dataset2.

The success rates of the C3D, I3D, and 3DCNN + BiLSTM models can be attributed to their superiority in spatio-temporal inference. When the results obtained from both datasets are examined, it is evident that lip reading studies should focus on analyzing the lip region in the dataset.

When the results of comparisons of all models are examined in Table 5, it is seen that higher accuracy rates can be achieved with the use of models that perform 3D convolution for time-dependent studies such as lip reading. The results obtained in both datasets may indicate the superiority of the 3DCNN + BiLSTM model in integrating spatial and temporal features. For models with 2DCNN, even concataneting frames to avoid losing time-dependent context does not achieve the success of 3DCNN models.

Moreover, the significant improvement in accuracy from CNN to LSTM models underscores the importance of considering temporal dynamics in video analysis tasks. However, the transition from LSTM to C3D and I3D models suggests that incorporating these dynamics through 3D convolutions directly within the network architecture is even more beneficial. This is particularly evident in the substantial leap in performance between C3D and I3D models, implying that I3D's approach to simultaneously modeling motion and appearance features is highly effective for complex, time-sensitive tasks.

After reviewing the results of individual models, it becomes evident that 3D convolution-based models (C3D, I3D, 3DCNN + BiLSTM) demonstrate clear superiority in integrating temporal information with spatial features. In particular, the I3D and 3DCNN + BiLSTM models achieve higher accuracy in tasks requiring temporal dependencies, such as lip reading. This performance advantage stems from their ability to simultaneously model both motion and spatial characteristics. However, each model exhibits varying levels of success across different classes, highlighting that while some models perform exceptionally well for certain classes, they may fall short for others. To leverage the strengths of each model, ensemble learning techniques were employed, aiming to combine the diverse classification capabilities of these models. By incorporating ensemble methods, the goal is to enhance overall performance through the aggregation of predictions from models that individually excel in different aspects of the task.

We extend our experimental work with ensemble learning. In this context, we use the 3DCNN + BiLSTM model in addition to C3D and I3D, the two models with the highest rates. We vote on the prediction results from the models using majority learning. In ensemble learning studies, different methods such as averaging and majority voting can be adopted at this stage. At the same time, the weights assigned to the models can be used to change their impact value in voting. In this phase of the study, the weights of the models are taken as equal in the voting method we applied. Figure 7 shows the schematic of the ensemble learning method used.

In this study, majority voting was chosen as the ensemble method due to its simplicity and effectiveness in combining the strengths of different models. Majority voting is particularly advantageous when individual models perform well in different areas, as it allows for a more balanced decision-making process by giving equal weight to each model's prediction. This method helps mitigate the risk of overfitting to any particular model's biases, as it relies on the consensus of multiple models rather than the dominance of a single one. Additionally, compared to more complex ensemble methods, such as weighted voting

or stacking, majority voting is computationally less intensive and easier to implement. Since the performance of the individual models in this study (C3D, I3D, 3DCNN + BiLSTM) is relatively close, giving each model equal importance ensures that the ensemble benefits from their combined strengths without the need for fine-tuning weight parameters, making it an efficient and robust choice for this specific application.

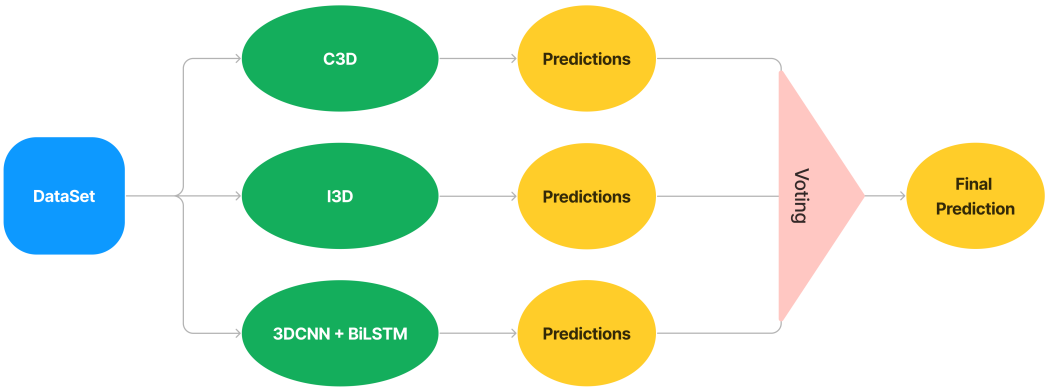


Figure 7. Ensemble learning scheme.

The performance of the models was assessed using key metrics such as accuracy, precision, recall, and F1 score, allowing for a comprehensive and detailed evaluation of the models' effectiveness.The results of the evaluations obtained are given in Table 6.

Table 6. Performance Metrics for Dataset2 Precision, Recall, F1-Score, Accuracy.

Model	Precision	Recall	F1-Score	Accuracy
CNN	55.47	60.13	57.48	63.12
LSTM	70.71	71.18	70.82	74.53
C3D	83.91	85.16	84.78	86.32
I3D	91.77	90.12	89.23	93.24
3DCNN + BiLSTM	91.34	91.92	90.32	93.30
C3D + I3D + (3DCNN + BiLSTM)	92.93	90.63	91.66	94.53

It is known that different techniques are used with ensemble learning. The ensemble learning technique used in this study is majority voting. After the experimental runs, 2.20% improvement was achieved in DataSet1 and 1.23% improvement was achieved in DataSet2.

Figure 8 shows the accuracy values graphs of the Ensemble model during training for DataSet2.

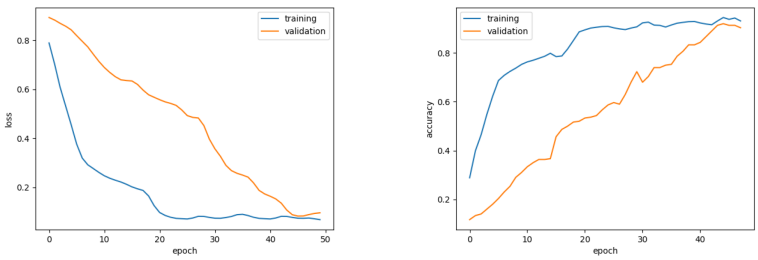


Figure 8. Accuracy and loss plots for DataSet2.

When examining the graphs in Figure 8, it is evident that the ensemble model's training and validation losses, as well as accuracy rates on DataSet2, change significantly over time. The training loss drops rapidly within the first 10 epochs and stabilizes, while the validation loss decreases more slowly and shows some fluctuations after the 20th epoch, indicating an improvement in the model's generalization ability. The training accuracy

steadily increases, while the validation accuracy rises more gradually and starts to converge after the 40th epoch. These trends suggest that the model is learning effectively without signs of overfitting. Furthermore, through the use of majority voting in the ensemble learning process, a performance improvement of 2.17% was achieved for DataSet1 and 1.23% for DataSet2.

The evaluation metrics provide crucial insights into the model's classification effectiveness. By analyzing these metrics, we can determine how well the model distinguishes between different classes. In Table 7, the precision, recall, and F1 score values are provided.

Table 7. Model performance for Ensemble Learning.

Numbers	Precision	Recall	F1 Score
0	0.8830	0.9651	0.9222
1	0.9004	0.9457	0.9225
2	0.9179	0.8120	0.8617
3	0.8986	0.8333	0.8647
4	0.8767	0.8205	0.8477
5	0.8995	0.8034	0.8488
6	0.8996	0.9188	0.9091
7	0.9291	0.9651	0.9468
8	0.8971	0.9457	0.9208
9	0.8913	0.9535	0.9213

The results presented in Table 7 evaluate the performance of the Ensemble Learning method. The metrics of Precision, Recall, and F1-Score indicate the accuracy of classifying Turkish numerals (0–9).The model demonstrates notable performance across most digits, with Precision values ranging from 0.8767 for digit 4 to 0.9291 for digit 7. Recall scores highlight the model's strong ability to correctly predict digits 0, 1, 7, 8, and 9, with values exceeding 0.94 for these digits. The F1-Scores, which provide a balanced measure of precision and recall, show the highest performance for digit 7 (0.9468), closely followed by digits 1 and 0, which yield F1-Scores of 0.9225 and 0.9222, respectively.

The model encounters some challenges with digits 2, 3, 4, and 5, where the Recall scores are relatively lower (ranging from 0.8034 to 0.8333), leading to slightly lower F1-Scores, indicating room for improvement in detecting these specific digits. The lower recognition accuracy of the digits 2, 3, 4 and 5 can be attributed to the phonetic and linguistic features of the Turkish language. In Turkish, the number "2" (iki) is a two-syllable word, while the numbers "3" (üç), "4" (dört) and "5" (beş) are monosyllabic words. Monosyllabic words often contain more subtle and less pronounced lip movements, making them less visually distinct and difficult for lip-reading models to distinguish. The difficulty of discriminating words with few syllables is seen in English [49] as well as in Turkish. Also, the transition to the second syllable in the word "two" does not require a significant lip movement, which may reduce the model's ability to capture the distinctive features of this word. These observations underscore the impact of language-specific phonetic and grammatical features on the performance of lipreading models and highlight the need for further adaptation and refinement to address these nuances in Turkish.

Overall, the ensemble model exhibits consistent high performance, particularly for digits 0, 1, 7, 8, and 9. Notably, digit 6 achieves a strong F1-Score of 0.9091, suggesting a balanced performance across both precision and recall, even though the precision and recall do not peak compared to other digits.These results highlight the effectiveness of the model across different classes. The consistency in performance metrics indicates that the model is reliable and performs well in distinguishing between all numeral categories. Confusion matrix of the results is given in Figure 9.

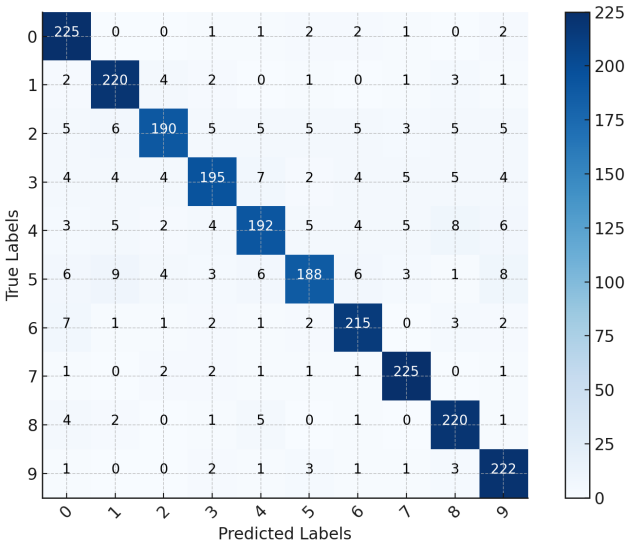


Figure 9. Confusion matrix of the proposed ensemble model.

The confusion matrix depicted in Figure 9 illustrates the ensemble learning model's performance in classifying the digits 0 through 9. The diagonal elements, representing the correct predictions for each class, are relatively high, indicating strong overall accuracy. Notably, class 0 and class 7 exhibit the highest accuracy, with 225 correctly predicted samples. Class 9 also performs well with 222 correctly classified instances.

Some misclassifications can be observed off the diagonal. For example, class 2 shows misclassifications across several neighboring classes, including 5, 6, and 3, with a noticeable number of samples being incorrectly classified as these digits. Similarly, class 5 demonstrates some difficulty, with multiple samples being misclassified as class 1 and class 6. This suggests that there may be a higher level of ambiguity or feature overlap among these classes, potentially making them harder for the model to distinguish. To address this issue and reduce the misclassification rate between these classes, a more comprehensive dataset could be created. In particular, the learning capacity of the model on these classes can be increased by including more examples of these classes. In addition, data obtained from different angles and under varying lighting conditions can be added to help the model learn wider variations. Another method is to use more sophisticated architectures such as attention mechanisms or feature extraction layers to help the model better discriminate overlapping features between classes. In particular, solutions such as language model integration for monosyllabic digits with similar lip movements can reduce ambiguity between classes by providing context information. These approaches can be effective for improving the performance of the model and minimizing the misclassification rate between the classes in question.

The confusion matrix highlights that while the ensemble learning model performs well overall, further refinement may be necessary to improve performance on classes with higher misclassification rates. Addressing these areas of confusion, particularly between digits like 2, 5, and 6, could further enhance the model's robustness and accuracy in distinguishing among all digit classes. When evaluating the digits with high accuracy rates, it can be observed that digits with multiple syllables are more easily detected.

Word Dataset

In order to evaluate the performance of the ensemble learning model on a different dataset, a new dataset composed of Turkish words was prepared, and experimental studies were conducted. During the creation of this dataset, existing datasets in the literature were

reviewed, and the process of obtaining the dataset was initiated. When examining datasets such as GRID, OuluVs, and AvLetter, it was observed that the number of repetitions ranged between 3 and 5 on average. The number of repetitions in the datasets is shown in Table 8.

Table 8. Number of repetitions of data sets.

Dataset	Repetitions
AvLetters	3
AvLetters2	3–7
OuluVS	5
OuluVS2	3
M2VTS [50]	5

In this context, the dataset was created with 4 repetitions for each sample. A total of 18 speakers participated in the process of creating the dataset, which consists of 150 words. By multiplying the number of speakers (18) with the number of repetitions (4) and the total number of words (150), a total of 10,800 videos were collected.

During the selection of the words, a process was carried out to identify the most frequently used words on news websites. This process was designed for the purpose of word recognition, and the categories of news articles used in the dataset were used only as a reference to identify frequent words. This approach to word selection aims to contribute to making lip-reading systems more practical and useful, especially for hearing impaired individuals. In the process of identifying the most frequently occurring words in news websites, a dataset of 41,992 news articles in 13 categories published by the Kemik Natural Language Processing Group was utilized. Table 9 shows the categories in the dataset.

Table 9. Distribution of News Articles Across Various Categories.

Class Name	Number of Articles	Class Name	Number of Articles
World	3724	Health	1383
Economy	3265	Politics	1849
General	6673	Sports	9997
Current Affairs	5847	Technology	771
Culture & Arts	1155	Turkey	1939
Entertainment	2792	Life	644

The Natural Language Toolkit (NLTK), a natural language processing library, was used to determine the words to be included in the dataset. During the preprocessing stage, tokenization, removal of punctuation and numbers, normalization, and the elimination of stopwords were performed. Through tokenization, the text was split into words and stored as an array. As a result of this process, the total word count was calculated as 9,855,099. Subsequently, punctuation marks and numbers were removed, allowing for more accurate word matching. In the normalization step, the distinction between uppercase and lowercase letters was eliminated to ensure correct matching of the words. The removal of stopwords refers to the process of eliminating words in the text that do not affect its meaning. During preprocessing, stopwords were removed. NLTK is an open-source natural language processing library developed for the Python programming language. Some of the Turkish stopwords included in the NLTK library are as follows: bu (this), gibi (like), daha (more). After the words were determined, the video recording process began. A studio environment was set up for the video recordings using a Canon 6D Mark II camera. Once the environment was prepared, the recording process commenced, and data collection started. Figure 10 shows the video recording environment.



Figure 10. Video shooting environment.

The data was collected by having 18 individuals repeat 150 words 4 times. The dataset was recorded at a resolution of 1920 × 1080 and a frame rate of 50 frames per second (fps).

Following these processes, the lip region was detected during the preprocessing stage, and frames were extracted from the videos by centering on the lip regions. The central point used during the data augmentation process was determined by considering the center of the lip region. While generating the frames, the right and left coordinates of the lip region as well as the nose region coordinates were identified. Based on these coordinates, frames were cropped from the videos by centering the lip region. This approach enabled more focused and clearer analysis of lip movements during the data augmentation phase. In videos where lip movements are critical, this method aimed to enhance spatial accuracy.

During data augmentation, each image was rotated 10 degrees to the right and 10 degrees to the left, using the central point as a reference. This process aims to enhance the model's spatial generalization ability by allowing objects in the images to be viewed from different angles. It is particularly beneficial in situations where objects are observed from various perspectives, helping the model make accurate predictions. The tilting process was implemented to diversify the dataset and improve the model's generalization capacity. Additionally, data augmentation helps prevent the model from overfitting during the training phase, enabling more accurate predictions.

As an additional data augmentation method, a multi-step preprocessing and augmentation strategy was implemented to standardize input dimensions and enhance the dataset for robust training. In this method, for videos with fewer than 16 frames, missing frames are replaced with a video by iteratively repeating the first and last frames until the required number of frames is reached. This approach minimizes the disruption of temporal flow as the first and last frames often contain critical spatial and temporal information. By using these key frames, the method ensures that the model receives a sequence that remains consistent with the structure of the original video. For videos with between 16 and 32 frames, the number of frames was reduced to 16. The frames were sampled at equal intervals throughout the video duration, ensuring that all temporal segments of the video contribute equally to the reduced sequence. This method avoids over-representation of certain segments, such as the beginning or end of the video, and ensures that salient features spanning the entire temporal range are preserved. For videos with more than 32 frames, a segmentation strategy was applied. To maximize the variability of the dataset while maintaining temporal coherence, the videos were divided into multiple segments of 16 frames each. Specifically, odd-numbered frames were assigned to the first segment and even-numbered frames to the second segment. In cases where the number of frames was large, equal intervals were set and the odd and even numbered frames were split into separate videos, resulting in two videos. For videos with more than 48 frames, the same process was performed to obtain 3 videos. This method effectively utilizes the full temporal resolution of the original video while capturing the dynamics of continuous lip movement and ensures that the temporal sequence in each segment remains intact.

Combining these approaches, the preprocessing pipeline addresses the challenges of variable frame counts, ensuring that the input data fed into the model is both standardized and representative. After the data augmentation process was completed, the data was stored in AVI format. The number of videos created using the data augmentation methods, such as removing specific regions from the videos and rotating them, is presented in Table 10.

Table 10. Number of videos in original and augmented for word dataset.

Phase	Video Count
Original Dataset	10,800
After Data Augmentation (Training)	34,713

After the data presented in Table 11 was obtained, the data preprocessing phase was completed, and the training phase began. The results obtained when the dataset was trained with the same architectures as the numerical dataset are presented in Table 11.

Table 11. Performance metrics of different models for word data set.

Model	Precision	Recall	F1-Score	Accuracy
CNN	63.15	55.09	58.88	65.35
LSTM	65.45	61.09	63.20	66.42
C3D	65.81	72.55	69.03	66.78
I3D	70.77	67.92	69.31	68.13
3DCNN + BiLSTM	75.12	73.21	74.15	77.69
C3D + I3D + (3DCNN + BiLSTM)	77.91	76.43	77.16	78.63

In the experiments conducted, the ensemble learning method achieved results that significantly outperformed the performance of individual models. The implementation of the ensemble learning strategy combined the strengths of different deep learning models, producing more consistent and accurate results. This approach balanced the weaknesses of each model, especially in complex datasets where video and natural language data were used together, enhancing the model's generalization ability.

The ensemble model achieved significant improvements in both precision and overall accuracy by combining the expertise of different models on spatial, temporal, and sequential data. When the advantages of I3D and C3D models in processing spatial and temporal information were combined with the BiLSTM model's ability to process sequential data, the Ensemble Learning method was observed to provide better results than other approaches. This approach not only enhanced the model's overall success but also allowed for more stable results despite the diversity and complexity of the dataset. The results clearly demonstrate that by integrating the strengths of individual models, Ensemble Learning produces more reliable and high-performance solutions for such complex datasets.

The original contribution of this study lies in presenting an ensemble approach that combines the strengths of different deep learning models, specifically increasing generalization ability in video data. In this context, this study contributes to the literature by offering a method that provides higher accuracy and generalization capacity using ensemble learning on a dataset that integrates both spatial and temporal dimensions. The strong performance of the model suggests that the use of Ensemble Learning methods will be beneficial in future studies.

5. Conclusions

In this study, a data set was created to use in lip reading Turkish digit recognition with deep learning architectures. The dataset is divided into two separate subsets in which

both the facial region and the lip region are extracted. The created data set was trained with incorporating both 2DCNN and 3DCNN architectures after data preprocessing. CNN, LSTM and 3DCNN based C3D, I3D, 3DCNN + BiLSTM models were used in the trainings. Optimal results were experimentally determined by tuning different hyperparameters. In the 2DCNN models, a single frame is generated by combining picture frames, prioritizing spatio-temporal information. As a result of the training, CNN and LSTM models achieved lower accuracy in both the data set containing the face region and the data set containing the lip region compared to C3D and I3D containing 3DCNN.

The C3D and I3D models were preferred for their superior ability to extract spatial information. In our study, the I3D model achieved a high success rate with an accuracy of 93.24%. In a study by Weng and Kitani [51] using I3D on English words, they found 84.07% success. Weng and Kitani used two-stream I3D and BiLSTM on the LRW dataset in their study and the study provided 5.3% improvement compared to previous state-arts. In this context, it is seen that I3D improves performance in lip reading studies. Similarly, ref. [52] used the C3D model in the pre-training stage in a multilayer approach, considering that it achieved higher performance than single layer in temporal tasks in their lip reading studies on LRW dataset. In this context, it can be said that models based on the 3DCNN architecture produce more accurate results in lip reading studies.

As a result of experimental studies, it is seen that 3DCNN models are more successful. In our study, the 3DCNN + BiLSTM model achieved the highest success rate with an accuracy of 93.30%. The ability of 3DCNN models to process spatio-temporal data is seen as the reason for this success. It is seen that 3DCNN models should be focused on for time-dependent video data and lip reading studies. It is seen that high performance can be achieved as a result of the correct selection of hyperparameters with good optimization techniques.

When we extended our study with the ensemble learning technique, we used the 3DCNN + BiLSTM model in addition to the C3D and I3D models that we achieved the highest successes. By using the majority voting method, which is one of the ensemble learning methods with the predictions of all models, we improved our best result by 1.23% to 94.53% and achieved a new state-of-the-art on our Dataset2 subset.

In addition to accuracy, we evaluated model performance using precision, recall, and F1 score. Precision measures the proportion of true positive predictions among all positive predictions, recall measures the proportion of true positives among all actual positives, and F1 score provides a harmonic mean of precision and recall. These metrics offer a more nuanced evaluation of the models' performance.

When the experimental results obtained are examined, when DataSet2 containing the lip region and DataSet1 containing the face region are compared, it is seen that the lip region achieves higher success in all experiments. In this context, it can be said that the extraction of the lip region is important in lip reading studies.

When the results obtained as a result of the training of the model are examined, it is seen that the digits can be determined by the Turkish lip reading method. In particular, it is seen that the accuracy of some figures is higher than others. It is thought that the reason for the difference in accuracy rates in digits is that the number of syllables of the numbers with high accuracy is higher than the number of syllables with low accuracy. For some figures, it can be said that the lip area makes sharp movements that will enable more specific inferences to be made.

While our study demonstrates promising results for Turkish digit recognition, the applicability of our approach to broader lip-reading tasks and other languages remains an open question. Future research should explore the scalability of our models by testing

them on more diverse datasets, including continuous speech and other languages with different phonetic characteristics.

While our study focuses on Turkish digit recognition, the methods and findings have potential applications in broader contexts. Future research could extend these techniques to continuous speech recognition and other languages with different phonetic and morphological structures. This would test the scalability and generalization capability of the models, providing further insights into their robustness across diverse linguistic datasets.

Our findings suggest that 3DCNN architectures combined with ensemble learning methods achieve higher accuracy in Turkish digit lip-reading. However, there is room for improvement in model performance, particularly for digits with lower recognition rates. Future work should focus on expanding the dataset, exploring other deep learning architectures, and applying our approach to continuous speech recognition tasks. Additionally, investigating the model's performance on other languages will provide valuable insights into its generalization and robustness.

In this study, a new dataset containing not only numbers but also words has been developed. This extended dataset allows the model's performance to be tested on broader lip-reading tasks. The experiments conducted on both numbers and words demonstrate the model's adaptability to different linguistic structures. By enabling training and testing processes with diverse word groups, the new dataset ensures that the scope of the study goes beyond digit recognition. Future studies can expand the scope by applying these architectures to other languages or testing them on more complex sentence structures.

In this study, when the examples that the model mislabeled after the learning process were examined, it was seen that the errors were caused by the sample. For further studies, the success rate can be increased by increasing the number of frames used in each sample during data augmentation and improving factors such as light and resolution during data collection. A Turkish lip reading system can be created by working with parameters such as different filter sizes and number of layers on different data.

One of the most significant contributions of this study to the literature is the introduction of a unique lip-reading dataset for Turkish. Furthermore, enhancing performance by combining 3DCNN and I3D models with BiLSTM using the ensemble learning method is a novel approach introduced for the first time in the literature. This study offers a methodology that not only contributes to Turkish lip-reading research but also serves as a reference for future lip-reading research in other languages.

Future research should aim to expand the dataset to include more diverse linguistic contexts and continuous speech recognition. Additionally, exploring other deep learning architectures and techniques, such as transformer models, may offer further improvements. Investigating the application of these models in real-world scenarios, including different languages and noisy conditions, will be crucial for practical deployment.

Author Contributions: Conceptualization, A.E. and N.B.; methodology, A.E. and N.B.; software, A.E.; validation, A.E. and N.B.; formal analysis, A.E. and N.B.; investigation, A.E. and N.B.; resources, A.E. and N.B.; data curation, A.E. and N.B.; writing—original draft preparation, A.E. and N.B.; writing—review and editing, A.E. and N.B.; visualization, A.E.; supervision, N.B.; project administration, N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Usak University Ethics Committee (protocol code 2024-208 and date of approval: 7 November 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dweik, W.; Altorman, S.; Ashour, S. Read my lips: Artificial intelligence word-level Arabic lipreading system. *Egypt. Inform. J.* **2022**, *23*, 1–12. [CrossRef]
2. Yavuz, Z.; Nabiyev, V.V. Bilgisayarlı Dudak Okuma. In Proceedings of the 15. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Eskişehir, Turkey, 11–13 June 2007; Volume 1–3, pp. 1216–1219.
3. Sun, B.; Xie, D.; Duan, T. Variable Structure and Modeling Units for Chinese Lipreading. *Int. J. Pattern Recognit. Artif. Intell.* **2022**, *36*, 2256021. [CrossRef]
4. Fernandez-Lopez, A.; Sukno, F.M. Survey on automatic lip-reading in the era of deep learning. *Image Vis. Comput.* **2018**, *78*, 53–72. [CrossRef]
5. Bi, C.; Zhang, D.; Yang, L.; Chen, P. An Lipreading Model with DenseNet and E3D-LSTM. In Proceedings of the 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; pp. 511–515.
6. Stafylakis, T.; Tzimiropoulos, G. Combining residual networks with LSTMs for lipreading. *arXiv* **2017**, arXiv:1703.04105.
7. Matthews, I.; Cootes, T.F.; Bangham, J.A.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 198–213. [CrossRef]
8. Wand, M.; Koutnık, J.; Schmidhuber, J. Lipreading with long short-term memory. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6115–6119.
9. Gündüz, C.; Polat, H. Turkish sign language recognition based on multistream data fusion. *Turk. J. Electr. Eng. Comput. Sci.* **2021**, *29*, 1171–1186. [CrossRef]
10. Miled, M.; Messaoud, M.A.B.; Bouzid, A. Lip reading of words with lip segmentation and deep learning. *Multimed. Tools Appl.* **2023**, *82*, 551–571. [CrossRef]
11. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *44*, 8717–8727. [CrossRef]
12. Anina, I.; Zhou, Z.; Zhao, G.; Pietikäinen, M. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; pp. 1–5.
13. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2421–2424. [CrossRef]
14. Feng, J.; Long, R. Cross-language lipreading by reconstructing spatio-temporal relations in 3D convolution. *Displays* **2023**, *76*, 102357. [CrossRef]
15. Fu, Y.; Lu, Y.; Ni, R. Chinese Lip-Reading Research Based on ShuffleNet and CBAM. *Appl. Sci.* **2023**, *13*, 1106. [CrossRef]
16. Zhao, Y.; Xu, R.; Song, M. A cascade sequence-to-sequence model for Chinese Mandarin lipreading. In Proceedings of the ACM Multimedia Asia, Beijing, China, 15–18 December 2019; pp. 1–6.
17. Ma, S.; Wang, S.; Lin, X. A transformer-based model for sentence-level Chinese Mandarin lipreading. In Proceedings of the 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Hong Kong, China, 27–30 July 2020; pp. 78–81.
18. Shin, J.; Lee, J.; Kim, D. Real-time lip reading system for isolated Korean word recognition. *Pattern Recognit.* **2011**, *44*, 559–571. [CrossRef]
19. Gimeno-Gomez, D.; Martinez-Hinarejos, C.D. Analysis of Visual Features for Continuous Lipreading in Spanish. In Proceedings of the IberSPEECH, Valladolid, Spain, 24–25 March 2021.
20. Schwiebert, G.; Weber, C.; Qu, L.; Siqueira, H.; Wermter, S. A multimodal German dataset for automatic lipreading systems and transfer learning. *arXiv* **2022**, arXiv:2202.13403.
21. Mesbah, A.; Berrahou, A.; Hammouchi, H.; Berbia, H.; Qjidaa, H.; Daoudi, M. Lip reading with Hahn convolutional neural networks. *Image Vis. Comput.* **2019**, *88*, 76–83. [CrossRef]
22. Petridis, S.; Wang, Y.; Ma, P.; Li, Z.; Pantic, M. End-to-end visual speech recognition for small-scale datasets. *Pattern Recognit. Lett.* **2020**, *131*, 421–427. [CrossRef]
23. Assael, Y.M.; Shillingford, B.; Whiteson, S.; De Freitas, N. LipNet: End-to-end sentence-level lipreading. *arXiv* **2016**, arXiv:1611.01599.
24. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.

25. Chen, X.; Du, J.; Zhang, H. Lipreading with DenseNet and resBi-LSTM. *Signal Image Video Process.* **2020**, *14*, 981–989. [CrossRef]
26. Zhao, X.; Yang, S.; Shan, S.; Chen, X. Mutual information maximization for effective lip reading. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 420–427.
27. Wang, H.; Pu, G.; Chen, T. A Lip Reading Method Based on 3D Convolutional Vision Transformer. *IEEE Access* **2022**, *10*, 77205–77212. [CrossRef]
28. Serengil, S.I.; Ozpinar, A. Lightface: A hybrid deep face recognition framework. In Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 15–17 October 2020; pp. 1–5.
29. Wong, V.; Abe, T.; Spitz, R.W.; Bell, Z.W.; Yamada, Y.; Chatakondi, R.N.; Loenneke, J.P. Effects of Age, Sex, Disease, and Exercise Training on Lip Muscle Strength. *Cosmetics* **2020**, *7*, 18. [CrossRef]
30. Schötz, S.; Frid, J.; Löfqvist, A. Development of speech motor control: Lip movement variability. *J. Acoust. Soc. Am.* **2013**, *133*, 4210–4217. [CrossRef]
31. Gürkahraman, K.; Karakiş, R. Brain tumors classification with deep learning using data augmentation. *J. Fac. Eng. Archit. Gazi Univ.* **2021**, *36*, 997–1011.
32. Hao, M.; Mamut, M.; Yadikar, N.; Aysa, A.; Ubul, K. A survey of research on lipreading technology. *IEEE Access* **2020**, *8*, 204518–204544. [CrossRef]
33. Pu, G.; Wang, H. Review on research progress of machine lip reading. *Vis. Comput.* **2023**, *39*, 3041–3057. [CrossRef]
34. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Cai, J.; Chen, T. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
35. Baur, D.; Kroboth, K.; Heyde, C.-E.; Voelker, A. Convolutional Neural Networks in Spinal Magnetic Resonance Imaging: A Systematic Review. *World Neurosurg.* **2022**, *166*, 60–70. [CrossRef]
36. Sevli, O. Performance comparison of different pre-trained deep learning models in classifying brain MRI images. *Acta Infologica* **2021**, *5*, 141–154. [CrossRef]
37. Kantar, O.; Kilimci, Z.H. Derin öğrenme temelli hibrid altın endeksi (XAU/USD) yön tahmin modeli. *Gazi Üniversitesi Mühendislik Mimar. Fakültesi Derg.* **2023**, *38*, 1117–1128. [CrossRef]
38. Sutskever, I.; Martens, J.; Hinton, G.E. Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 1017–1024.
39. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
40. Hu, Z.; Si, X.; Luo, Y.; Tang, S.; Jian, F. Speaker Recognition Based on 3DCNN-LSTM. *Eng. Lett.* **2021**, *29*, 2.
41. Al-Hammadi, M.; Muhammad, G.; Abdul, W.; Alsulaiman, M.; Bencherif, M.A.; Mekhtiche, M.A. Hand gesture recognition for sign language using 3DCNN. *IEEE Access* **2020**, *8*, 79491–79509. [CrossRef]
42. Tatebe, Y.; Deguchi, D.; Kawanishi, Y.; Ide, I.; Murase, H.; Sakai, U. Pedestrian detection from sparse point-cloud using 3DCNN. In Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, Thailand, 7–9 January 2018; pp. 1–4.
43. Zhang, L.; Zhu, G.; Shen, P.; Song, J.; Shah, S.A.; Bennamoun, M. Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 3120–3128.
44. Maeda, T.; Tamura, S. Multi-view Convolution for Lipreading. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 1092–1096.
45. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision(ICCV), Santiago, Chile, 11–18 December 2015; pp. 4489–4497.
46. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
47. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, *33*, 1–39. [CrossRef]
48. Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 757–774. [CrossRef]
49. Fisher, C.G. Confusions Among Visually Perceived Consonants. *J. Speech Hear. Res.* **1968**, *11*, 796–804. [CrossRef]
50. Vanegas, O.; Tokuda, K.; Kitamura, T. Location normalization of HMM-based lip-reading: Experiments for the M2VTS database. In Proceedings of the 1999 International Conference on Image Processing (Cat. 99CH36348), Kobe, Japan, 24–28 October 1999; Volume 2, pp. 343–347. [CrossRef]

51. Weng, X.; Kitani, K. Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading. In Proceedings of the 30th British Machine Vision Conference 2019 (BMVC 2019), Cardiff, UK, 9–12 September 2019.
52. Song, Q.; Sun, B.; Li, S. Multimodal Sparse Transformer Network for Audio-Visual Speech Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 10028–10038. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.