applied sciences

MDPI

Article

AUDD: Audio Urdu Digits Dataset for Automatic Audio Urdu Digit Recognition

Aisha Chandio [1], Yao Shen [1,*], Malika Bendechache [2], Irum Inayat [3] and Teerath Kumar [3]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; aishabatool5512@gmail.com
[2] ADAPT & Lero Research Centres, School of Computing, Dublin City University, Dublin 9, Ireland; malika.bendechache@dcu.ie
[3] Department of Software Engineering, School of Computing, National University of Computer and Emerging Sciences, Islamabad 44000, Pakistan; irum.inayat@nu.edu.pk (I.I.); tchandani142@gmail.com (T.K.)
* Correspondence: yshen@cs.sjtu.edu.cn

Abstract: The ongoing development of audio datasets for numerous languages has spurred research activities towards designing smart speech recognition systems. A typical speech recognition system can be applied in many emerging applications, such as smartphone dialing, airline reservations, and automatic wheelchairs, among others. Urdu is a national language of Pakistan and is also widely spoken in many other South Asian countries (e.g., India, Afghanistan). Therefore, we present a comprehensive dataset of spoken Urdu digits ranging from 0 to 9. Our dataset has 25,518 sound samples that are collected from 740 participants. To test the proposed dataset, we apply different existing classification algorithms on the datasets including Support Vector Machine (SVM), Multilayer Perceptron (MLP), and flavors of the EfficientNet. These algorithms serve as a baseline. Furthermore, we propose a convolutional neural network (CNN) for audio digit classification. We conduct the experiment using these networks, and the results show that the proposed CNN is efficient and outperforms the baseline algorithms in terms of classification accuracy.

Keywords: audio classification; baseline classification accuracy; digit recognition; speech processing; Urdu dataset classification; Urdu digit dataset

check for updates

Citation: Chandio, A.; Shen, Y.; Bendechache, M.; Inayat, I.; Kumar, T. AUDD: Audio Urdu Digits Dataset for Automatic Audio Urdu Digit Recognition. *Appl. Sci.* **2021**, *11*, 8842. https://doi.org/10.3390/app11198842

Academic Editor: Javier Hernando

Received: 28 July 2021
Accepted: 15 September 2021
Published: 23 September 2021

1. Introduction

Deep Learning has been successful in multiple domains including image classification [1–3], text classification [4,5], speech recognition [6–8], and many more [9]. In speech recognition, there has been a tremendous amount of work done ranging from speech recognition to speech mimicry for multiple languages. However, in all languages, data have been a key challenge for any deep learning task, as data are difficult to find, and data collection is a time-consuming and tedious task to do. Without data, research and development for any language is not possible. To continue the robust and reliable development of speech recognition, it is necessary to have publicly available data so researchers can continue to develop speech-related systems. Similarly, for Urdu audio digit classification, there is no dataset available publicly. Inspired by the need for audio digit recognition, we release the Urdu audio digit dataset. Typically, a framework for automatic speech recognition [10] in Urdu can assist voice recognition in other languages that are similar [11].

A few works [12,13] considered speech processing schemes in the Urdu language. The field of computer science associated with the interactions with computers and natural languages is known as Natural Language Processing (NLP). This work is related to Automatic Speech Recognition (ASR), which is the advanced research domain of the NLP [12]. Automatic Speech Recognition (ASR) is the transformation of an acoustical waveform into the text equivalent of the information conveyed by the spoken word. The main objective of ASR research is to enable the computers that can recognize continuous speech with limitless vocabulary. This aim remains unfulfilled, as yet, although substantial progress has been achieved. Artificial neural networks (ANNs) were created to mimic the human brain. ANNs perform excellent classification and are used in a variety of pattern classification applications. ANNs are a suitable selection for Automatic Speech Recognition (ASR) systems because of their exceptional classification abilities. Most of the modern speech recognition systems that have been created use ANNs for speech signals classification [13]. Despite the importance of ASR, the research for Urdu ASR is a quite new domain because of two reasons: (1) lack of Urdu language studies research and (2) lack of a balanced speech dataset.

Currently, some work studied ASR for Urdu [13–20]. Additionally, numerous works [12,17,21–23] have been done on spoken digit recognition for Urdu. The above-mentioned works have shown significant results. However, the considered datasets were not comprehensive; thus, trained machine learning models can be further improved by using large datasets. To address this issue, we propose a comprehensive Audio Urdu Digits Dataset (AUDD). The main contributions of our work are as follows.

- We present an AUDD for Urdu digits that comprises 0–9. We have collected these voice samples from 740 people. These people were of different age groups ranging from 5 to 89 years.
- We perform extensive experiments for different networks to provide classification accuracy that will serve as baseline accuracy.
- We propose a convolutional neural network (CNN) for classification that shows impressive performance being a simple CNN compared to complex flavors of efficientNet.

The remaining work is organized as follows: Section 2 gives an overview of relevant work that has been done previously, Section 3 describes data collection in detail, Section 5 describes the details of the experiment and baseline results, and finally Section 6 concludes the overall contribution.

2. Related Work

There have been several studies performed on spoken digits recognition in different languages. Messer et al. proposed [24] the first complete audio visual database of spoken digits, known as Extended XM2VTS, which is made up of recordings from 295 people of all ages and genders. Each speaker's audio recording consists of two continuous digit strings. Bailly et al. proposed the BANCA database [25], which has utterances in English, Spanish, Italian, and French, among other languages. The goal of this database was to put person verification to the test in a variety of settings, including controlled, degraded, and unfavorable. Jain et al. [26] proposed Discrete Cosine Transform and Local Binary Patterns to conduct an audiovisual digit recognition challenge. Brahme et al. [27] created an in-house database for Marathi speech digits to explore lip reading movement and digit classification.Wazir et al. [28] proposed LSTM to accomplish the recognition task, and MFCCs were used to extract features from audio files to develop a speech recognition solution for Arabic numerals using RNN. This model is capable of classifying digits from noisy spoken digits. The results are promising, with a 69 percent of overall recognition accuracy. Dalsaniya et al. [29] presented a novel, publicly available audio dataset of spoken digits in the Gujarati language, as well as some preliminary results. A comparison of the categorization algorithm with an English language database was also performed. Many works [12,17,22] considered Urdu language voice data for various applications. Hasnain et al. in [17] described the frequency analysis of spoken Urdu numbers from 0 to 9. In this work, the authors collected voice samples from 15 speakers who spoke the same digit set from 0 to 9. The voice recordings were curtailed for 0.9 min. The data were recorded using a microphone connected to a Windows-based PC and saved into .wav file format. The initial processing of data was done using Simulink and MATLAB. Fourier descriptions and correlation calculations were performed using these same tools. The same words were delivered by the same and different people, and words were distinguished using correlation. A feed-forward neural network developed in MATLAB was used to classify the voices with 100 percent accuracy. Although authors obtained maximum accuracy and found a trade-off between goal, epochs, and learning rate, the dataset used in this work was too small.

All of the above works [12,17,21,22] have used Urdu datasets with limited data (e.g., [17] used a dataset of 150 samples). In contrast, our dataset is comprehensive and diverse and consists of 25,518 samples, which is helpful for classification.

3. Data Collection

Urdu is widely spoken in South Asia and is the national language of Pakistan. Predicting a digit from speech and making automatic speech digit classification efficient and robust is the basic reason for our motivation. Secondly, there is no big dataset available publicly. To the best of our knowledge, this is the largest publicly available dataset that can be used for many research purposes.

In our data collection, different age groups of people participated. There was a total of 740 participants, the age ranging from 5 to 89. Most participants were 5 to 14 years old, as shown in Figure 1. Comparing male and female participants, male participants were slightly more numerous than female participants, as shown in Figure 2. We also checked the number of samples in each class; most classes sampled more than 2000 samples, overall samples were 25,518, and the class-wise distribution is shown in Figure 3. For diversity purposes, we recorded each sample in different environments, i.e., participants record samples in a normal environment, noisy environment, closed room environment, at different heights, and different distances. Furthermore, we asked the participant to record samples with different characteristics, such as slow voice, fast voice, low loudness, high loudness, etc.

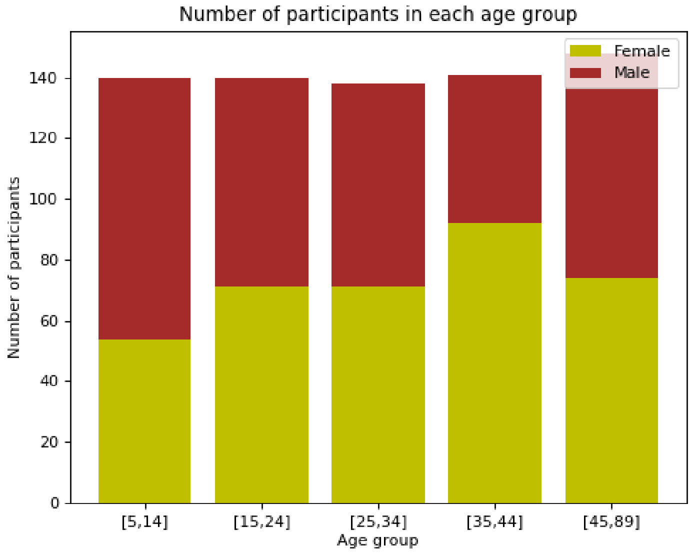

Figure 1. Gender and Age of participants.

All voice samples were collected using these mobile devices: (i) VIVO mobile, model V204321 with 64 GB storage, 2.3 GHz Octacore processor, and 4 GB RAM. (ii) VIVO mobile, model 190,719 with 128 GB storage, 2.0 GHz Octacore, and 4 GB RAM. (iii) Galaxy J5, model SM-G570F. (iv) XIAOMI mobile, model Redmi 6 Pro wiith 64 GB storage, Octacore Max 2.0 GHz, and 4 GB RAM. All these devices were used for recording voice samples with different voice-recording applications, i.e., Easy Voice Recorder Pro, Smart Recorder—High-quality voice recorder. Each sample was collected at a sample rate of 48 kHz and was a mono channel with a minimum length of 1 s and a maximum length of 2 s. All samples were stored in opus format.

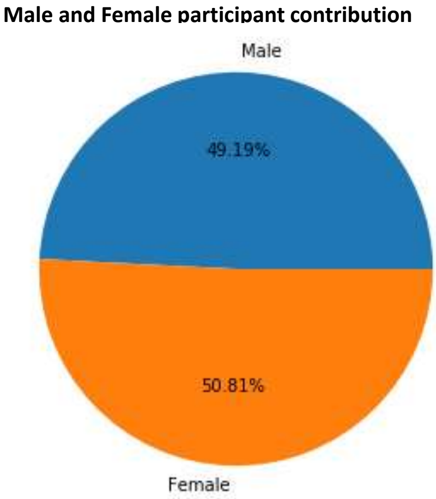

Figure 2. Gender of participants, percentage.

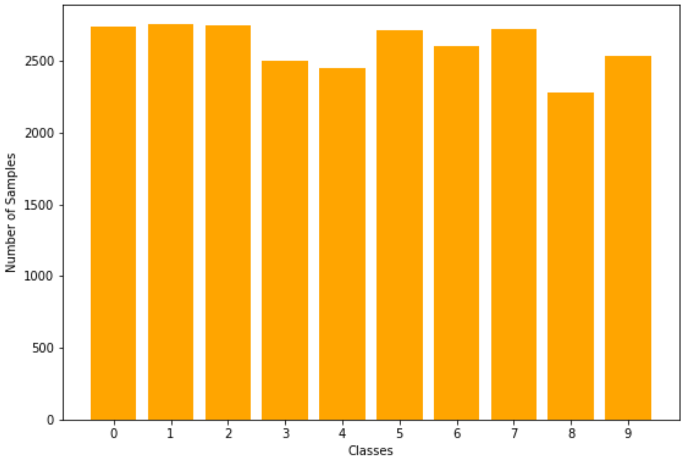

Figure 3. Class numbers of samples.

Visualization

We visualized the samples in different ways, i.e., wave form, spectrogram, and Mel spectrogram. To visualize the wave form of the samples of each class, first we loaded the samples using Librosa library [30], and we show each digit audio in the waveform, as in Figure 4. As the samples were already loaded, we just calculated the Mel spectrometer using Librosa and showed each sample in Mel spectrogram, as shown in Figure 5. Finally, we calculated the log of Mel spectrogram using the library and visualized it, as shown in Figure 6. From this visualization, we can understand that any representation (waveform, spectrogram, or Mel spectrogram) for single digits is unique. Based on this uniqueness, it is easy for CNN to learn these unique features. Thus, we applied CNN and other neural network models.

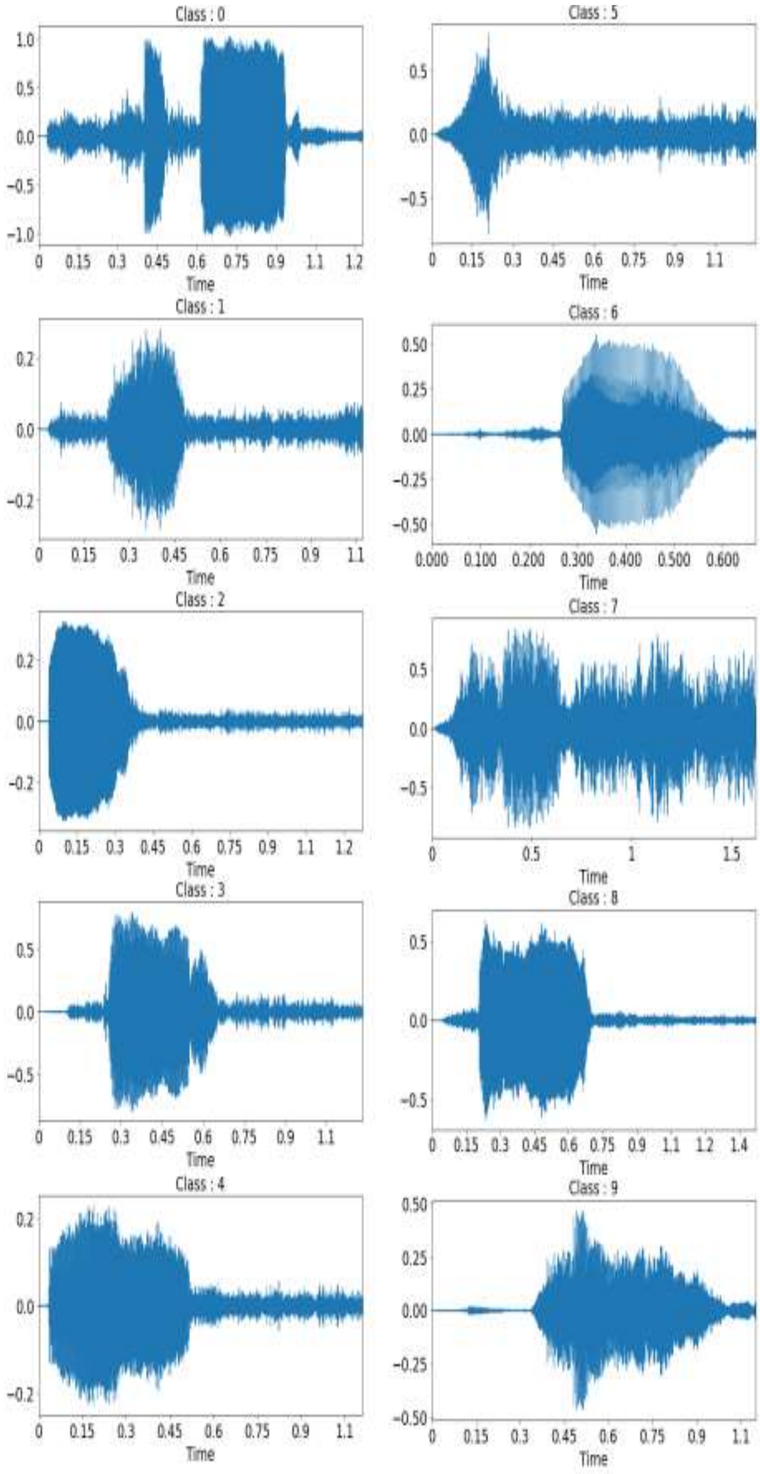

Figure 4. Wave form of each class sample.

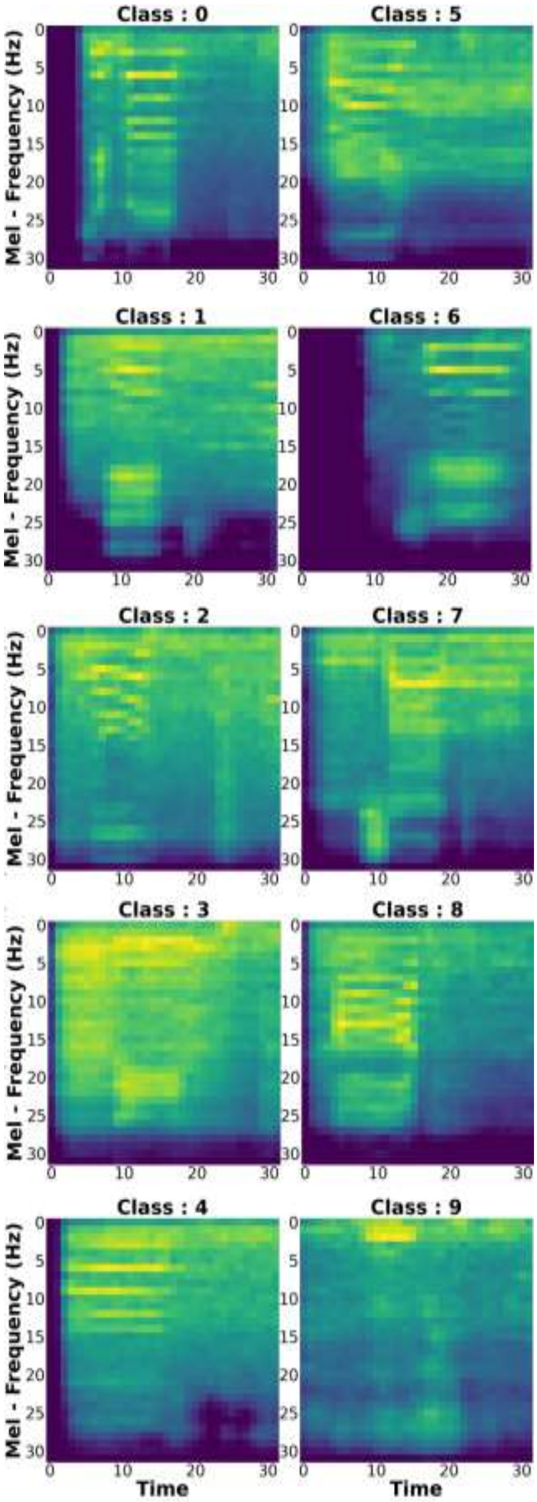

Figure 5. Mel Spectrogram of each class sample.

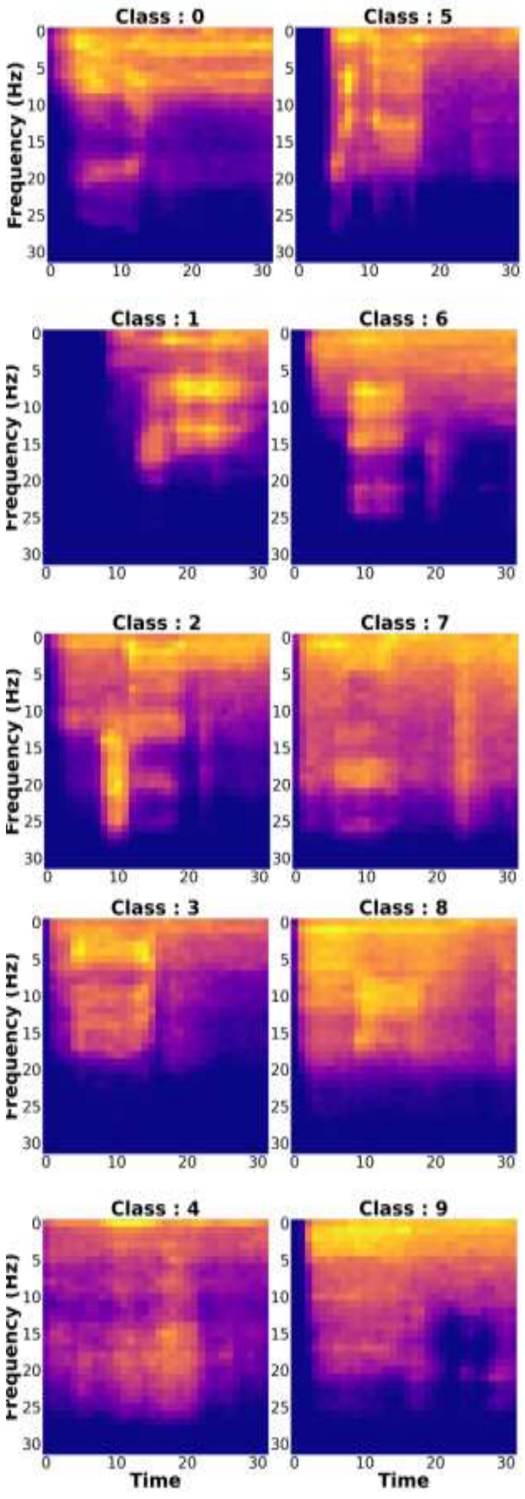

Figure 6. Spectrogram of each class sample.

4. Models

In this section, we explain various models used for the classification of audio Urdu digits.

4.1. Support Vector Machine

SVM is a supervised machine learning approach, which can be applied to classification and regression problems. It is, however, mostly utilized in classification problems. In the SVM algorithm, each data item is plotted as a point in an n-dimensional space (where n is the number of features), and each feature value corresponds to a particular coordinate. The classification is performed by searching the hyper-plane that differentiates the different classes. The coordinates of individual observations are known as support vectors [31].

4.2. Multilayer Perceptron

MLP is an artificial neural network that comprises multiple layers of neurons in a feed-forward fashion. A multilayer perceptron has three or more layers, each with an input, output, and one or more hidden layers. The neurons in MLP have a nonlinear activation function, and each layer is fully connected to the next. Using non-linear/linear activation functions, several perceptrons are merged to generate a decision boundary. A non-linear mapping to a new dimension is provided by each perceptron [32].

4.3. EfficientNet

EfficientNet is a new model scaling method recently developed by Google [33] for scaling up CNNs. It uses a simple, greatly effective compound coefficient. EfficientNet works differently from traditional methods that scale dimensions of networks, such as width, depth, and resolution; it scales each dimension with a fixed set of scaling coefficients uniformly. Practically, scaling individual dimensions improves model performance. However, balancing all dimensions of the network concerning the available resources effectively improves the whole performance.

5. Experiments

In this section, we explain preprocessing, training setup, and results as a baseline by different models.

5.1. Preprocessing

We perform a single-step preprocessing on audio samples that is to calculate Log-Mel Spectrogram [34]. To calculate the Log-Mel Spectrogram, first Fast Fourier Transform (FFT) is calculated on an audio signal using Equation (1).

$$S_i(\mathrm{p}) = \sum_{n=1}^{N} s_i(n)h(n)\mathrm{e}^{-\mathrm{j}2\pi\mathrm{pn/N}}, \mathrm{p} = 0,\ldots,\mathrm{N}-1 \tag{1}$$

In Equation (1), $h(n)$ and s_i are the N sample long analysis window and time-domain samples, respectively, and $S_i(\mathrm{p})$ and N are frequency-domain samples and Fast Fourier Transform size, respectively. After frequency domain samples, the next task is to map its amplitude to the Mel scale of perceptual excitation using a filter bank. A Mel filter bank is used to convert the spectrum to a Mel spectrum. The Mel-scale is calculated based on human hearing frequencies [35]. Thus, the Mel-scale is used for tone measuring, and Mel-frequency scale is calculated as shown in Equation (2).

$$\mathrm{mel}\ (\mathrm{frequency}) = \ \mathrm{scale}\ \times \ln\left(1 + \frac{frequency}{700}\right) \tag{2}$$

We performed the above calculation using Librosa library. After Mel-Spectrogram calculation, we resized it to 32 × 32 using the resize feature of OpenCV [36] and saved it as an image.

5.2. Training Setup

We used multiple deep learning models. For SVM and MLP, we used Scikit learn [37,38]. SVM was used with default parameters, MLP was used with two hidden layers having dimensions of 512 and 256, and the maximum iterations were 200.

We used the Keras library [39] for built-in deep learning models; the last layer of each model was removed, and the new last layer was added with 10 classes, as in our dataset we had only ten classes. During loading the model weight is set as none since we did not want a pretrained model. For each model, we used a batch size 256 and a learning rate of 0.01. We used Adam optimizer [40].

We devised a simple 3-layer convolutional neural network (CNN). Interestingly, it has shown impressive performance being a simple network. Further architecture detail is given in Table 1.

Table 1. CNN Model.

Layer Type	Dimensions	Comments
input	input layer	$(32, 32, 1)$
CNN	$(30, 30, 64)$	kernel 3×3; stride 1; relu activation
Max Pool	$(15, 15, 64)$	N.A
BN	$(15, 15, 64)$	default value as given in Keras [39]
CNN	$(13, 13, 64)$	kernel 3×3; stride 1; relu activation
Max Pool	$(6, 6, 64)$	N.A
BN	$(6, 6, 64)$	default value as given in Keras [39]
CNN	$(4, 4, 64)$	kernel 3×3; stride 1; relu activation
Max Pool	$(2, 2, 64)$	N.A
BN	$(2, 2, 64)$	default value as given in Keras [39]
Dropout	$(2, 2, 64)$	dropout rate = 0.1
Flatten	256	N.A
Fully connected	512	N.A
Dropout	512	dropout rate = 0.1
Fully connected	128	N.A
Dropout	512	dropout rate = 0.1
fully connected	10	softmax activation

5.3. Classification Results

We performed experiments for different models. Each experiment was repeated three times, and average accuracy with variance is reported as shown in Table 2. In Table 2, Support Vector Machine [31] and Multilayer perceptron [13] are multi class classifiers, CNN is our proposed architecture, and other models are the flavors of efficient Net that range from B0 to B7. Accuracy, defined as below, was used for evaluating the performance:

$$A = \frac{C}{T} \tag{3}$$

where A is accuracy, C is the number of samples recognized correctly, and T is the number of all samples.

Furthermore, we also showed the accuracy at each epoch for all models, as in Figure 7. Among them, our proposed CNN model showed better performance. Additionally, our proposed CNN showed faster convergence, as in Figure 7. For evaluating the effectiveness of our proposed CNN on different languages, we performed the experiments on Gujarati [29] and English [41,42] spoken digits. Our proposed CNN also outperformed the Gujarati Digit Model by absolute 22% accuracy. We compared the proposed model's accuracy on the English digit dataset with multiple baseline models; the proposed CNN outperformed CNNDigit Reco [43], SVM [44], random forest [44], and English model results by absolute 19.3%, 7.3%, and 0.3%, respectively. Gujarati digit model and English digit model performances are compared in Table 3 and in Table 4. Furthermore, we checked the effectiveness of the proposed CNN on the Urdu Corpus [11] dataset. It is found that the proposed CNN outperformed LDA, SVM, and RF by absolute accuracy margins of 34.53%, 24.53%, and 34.53%, respectively, as it is shown in Table 5.

Table 2. Baseline accuracies of different models.

Model Name	Accuracy
Support Vector Machine	0.65 ± 0.0
Multilayer Perceptron	0.73 ± 0.02
CNN	0.86 ± 0.02
EfficientNetB0	0.84 ± 0.05
EfficientNetB1	0.82 ± 0.02
EfficientNetB2	0.83 ± 0.04
EfficientNetB3	0.84 ± 0.06
EfficientNetB4	0.82 ± 0.03
EfficientNetB5	0.84 ± 0.04
EfficientNetB6	0.81 ± 0.06
EfficientNetB7	0.56 ± 0.07

Table 3. Comparison of CNN on Gujarati Digits [29].

Model Name	Accuracy
Gujarati Digits Model [29]	0.75
Our CNN	0.97

Table 4. Comparison of CNN on English Digits [41].

Model Name	Accuracy
CNNDigitReco-speakerindependent [43]	0.78
Support Vector Machines [44]	0.90
Random Forest [44]	0.96
English Digit Model [41]	0.97
Our CNN	0.973

Table 5. Comparison of CNN on Urdu Corpus [11].

Model Name	Accuracy
Linear Discriminant Analysis (LDA) [11]	0.63
Support Vector Machines (SVM) [11]	0.73
Random Forest (RF) [11]	0.63
Our CNN	0.9753

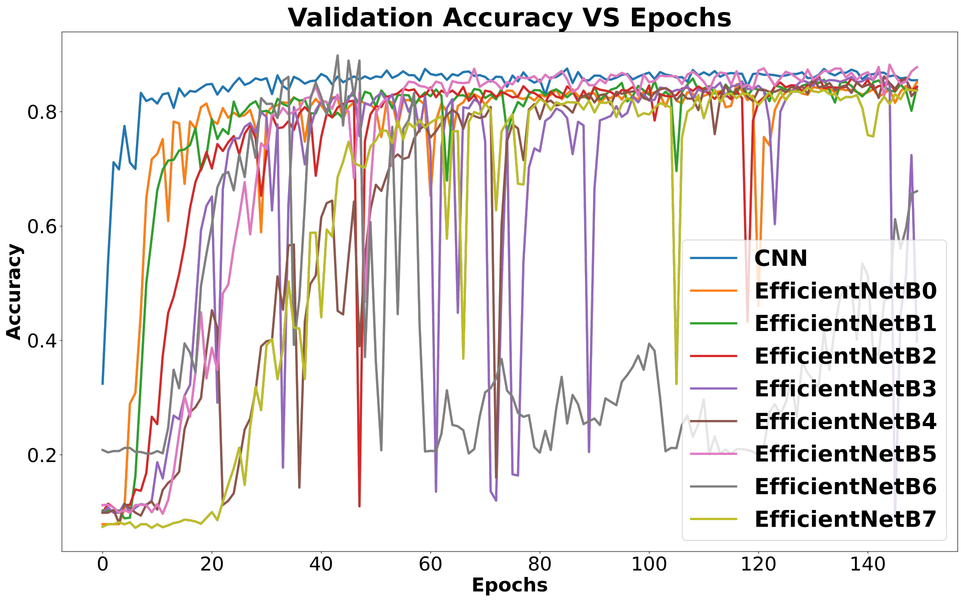

Figure 7. Validation accuracy of each model.

6. Conclusions

To the best of our knowledge, this is the first study that provides the largest and publicly available Audio Urdu Digits dataset having diverse characteristics. The dataset comprised 25,518 samples of 10 classes (0–9) that were collected from 740 participants of diverse age groups in different environmental conditions.The comprehensive data analysis in the form of waveform, spectrogram, and Mel spectrogram shows that a CNN with a small receptive size, more numbers of filters, and a small max pooling window can improve the results. We have also provided the baseline results of this novel study for the research community. Furthermore, to evaluate the effectiveness of the proposed CNN, it was also tested with two different language digit datasets, i.e., Gujarati and English, and promising results were obtained.

Author Contributions: Conceptualization, Y.S. and A.C.; methodology, A.C. and M.B.; software, A.C. and T.K.; validation, Y.S. and I.I.; formal analysis, M.B. and T.K.; investigation, T.K. and Y.S.; resources, Y.S.; data curation, A.C. and I.I.; writing—original draft preparation, A.C.; writing—review and editing, I.I. and M.B.; visualization, A.C.; supervision, Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Research and Development Projects of Applied Technology of Inner Mongolia Autonomous Region, China under Grant No. 201802005, the Key Program of the National Natural Science Foundation of China under Grant No. 61932014, and Pudong New Area Science & Technology Development Fundation under Grant No. PKX2019-R02. Yao Shen is the corresponding author.

Data Availability Statement: Our released dataset can be found at Urdu Audio Dataset (https://www.kaggle.com/zeroaishazero/urduaudiodigit). Code is available: github Code.

Acknowledgments: Apart from my own efforts, the success of this research work depends mainly on the encouragement and guidelines of my supervisor Yao Shen. His positive outlook and confidence in my research inspired me and gave me confidence. I also want to express my most significant appreciation to Rodrigo Gantier and Hassam Ahmad. I can not say thank you enough for their tremendous support and help.

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

References

1. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
2. Khan, L.; Saad, W.; Han, Z.; Hong, C. Dispersed federated learning: Vision, taxonomy, and future directions. *arXiv* **2020**, arXiv:2008.05189.
3. Khan, L.; Saad, W.; Han, Z.; Hossain, E.; Hong, C. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1759–1799. [CrossRef]
4. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]
5. Jindal, R.; Malhotra, R.; Jain, A. Techniques for text classification: Literature review and current trends. *Webology* **2015**, *12*, 2.
6. Piczak, K. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
7. Diez, Gaspon, I.; Saratxaga, I.; Ipiña, K. Deep Learning For Natural Sound Classification. In Proceedings of the INTER-NOISE and NOISE-CON Congress and Conference Proceedings, Madrid, Spain, 16–19 June 2019; Volume 259, pp. 5683–5692.
8. Lu, H.; Zhang, H.; Nayak, A. A deep neural network for audio classification with a classifier attention mechanism. *arXiv* **2020**, arXiv:2006.09815.
9. Heaton, J. Applications of deep neural networks. *arXiv* **2020**, arXiv:2009.05673.
10. Meng, J.; Zhang, J.; Zhao, H. Overview of the speech recognition technology. In Proceedings of the 2012 Fourth International Conference on Computational And Information Sciences, Chongqing, China, 17–19 August 2020; pp. 199–202.
11. Ali, H.; Ahmad, N.; Hafeez, A. Urdu speech corpus and preliminary results on speech recognition. In Proceedings of the International Conference on Engineering Applications of Neural Networks, Aberdeen, UK, 2–5 September 2016; pp. 317–325.
12. Ashraf, J.; Iqbal, N.; Khattak, N.; Zaidi, A. Speaker independent Urdu speech recognition using HMM. In Proceedings of the 2010 the 7th International Conference on Informatics And Systems (INFOS), Cairo, Egypt, 28–30 March 2010; pp. 1–5.
13. Ahad, A.; Fayyaz, A.; Mehmood, T. Speech recognition using multilayer perceptron. In Proceedings of the IEEE Students Conference (ISCON'02 Proceedings), Lahore, Pakistan, 16–17 August 2002; Volume 1, pp. 103–109.
14. Sarfraz, H.; Hussain, S.; Bokhari, R.; Raza, A.; Ullah, I.; Sarfraz, Z.; Pervez, S.; Mustafa, A.; Javed, I.; Parveen, R. Speech corpus development for a speaker independent spontaneous Urdu speech recognition system. In Proceedings of the O-COCOSDA, Kathmandu, Nepal, 24 November 2010.
15. Raza, A.; Hussain, S.; Sarfraz, H.; Ullah, I.; Sarfraz, Z. Design and development of phonetically rich Urdu speech corpus. In Proceedings of the 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, Urumqi, China, 10–12 August 2009; pp. 38–43.
16. Akram, M.; Arif, M. Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach. In Proceedings of the 8th International Multitopic Conference, (Proceedings of INMIC 2004), Lahore, Pakistan, 24–26 December 2004; pp. 91–96.
17. Hasnain, S.; Awan, M. Recognizing spoken Urdu numbers using fourier descriptor and neural networks with Matlab. In Proceedings of the 2008 Second International Conference on Electrical Engineering, Lahore, Pakistan, 25–26 March 2008; pp. 1–6.
18. Ittoo, A.; Bouma, G.; Maruster, L.; Wortmann, H. Extracting meronymy relationships from domain-specific, textual corporate databases. In Proceedings of the International Conference on Application of Natural Language to Information Systems, Cardiff, UK, 23–25 June 2010; pp. 48–59.
19. Ali, H.; Ahmad, N.; Zhou, X.; Iqbal, K.; Ali, S. DWT features performance analysis for automatic speech recognition of Urdu. *SpringerPlus* **2014**, *3*, 1–10. [CrossRef] [PubMed]
20. Faisal, M.; Manzoor, S. Deep learning for lip reading using audio-visual information for urdu language. *arXiv* **2018**, arXiv:1802.05521.
21. Ali, H.; Jianwei, A.; Iqbal, K. Automatic speech recognition of Urdu digits with optimal classification approach. *Int. J. Comput. Appl.* **2015**, *118*, 1–5. [CrossRef]
22. Azam, S.; Mansoor, Z.; Mughal, M.; Mohsin, S. Urdu spoken digits recognition using classified MFCC and backpropgation neural network. In Proceedings of the Computer Graphics, Imaging and Visualisation (CGIV 2007), Bangkok, Thailand, 14–17 August 2007; pp. 414–418.
23. Zia, T.; Zahid, U. Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. *Int. J. Speech Technol.* **2019**, *22*, 21–30. [CrossRef]
24. Messer, K.; Matas, J.; Kittler, J.; Jonsson, K.; Luettin, J.; Maître, G. XM2VTSDB: The extended M2VTS database. In Proceedings of the Second International Conference on Audio and Video-Based Biometric Person Authentication, Washington, DC, USA, 22–23 March 1999; pp. 965–966.
25. Bailly-Bailliére, E.; Bengio, S.; Bimbot, F.; Hamouz, M.; Kittler, J.; Mariéthoz, J.; Matas, J.; Messer, K.; Popovici, V.; Porée, F.; et al. The BANCA database and evaluation protocol. In Proceedings of the International Conference on Audio-and Video-Based Biometric Person Authentication, Guildford, UK, 9–11 June 2003; pp. 625–638.
26. Abhilash, J.; Rathna, G.N. Visual speech recognition for isolated digits using discrete cosine transform and local binary pattern features. In Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14–16 November 2019; pp. 368–372.

27. Brahme, A.; Bhadade, U. Marathi digit recognition using lip geometric shape features and dynamic time warping. In Proceedings of the TENCON 2017—2017 IEEE Region 10 Conference, Penang, Malaysia, 5–8 November 2017; pp. 974–979.
28. Wazir, A.S.M.B.A.; Chuah, H.J. Spoken arabic digits recognition using deep learning. In Proceedings of the 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Selangor, Malaysia, 29 June 2019; pp. 339–344.
29. Dalsaniya, N.; Mankad, S.H.; Garg, S.; Shrivastava, D. Development of a novel database in Gujarati language for spoken digits classification. In Proceedings of the International Symposium on Signal Processing and Intelligent Recognition Systems, Trivandrum, India, 18–21 December 2019; pp. 208–219.
30. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25
31. Mathur, A.; Foody, G. Multiclass and binary SVM classification: Implications for training and classification users. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 241–245. [CrossRef]
32. Zhai, X.; Ali, A.; Amira, A.; Bensaali, F. MLP neural network based gas classification system on Zynq SoC. *IEEE Access* **2016**, *4*, 8138–8146. [CrossRef]
33. Tan, M.; Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach Convention & Entertainment Center, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
34. Oo, M.; Oo, L. Fusion of Log-Mel Spectrogram and GLCM feature in acoustic scene classification. In Proceedings of the International Conference on Software Engineering Research, Management and Applications, Honolulu, HI, USA, 29–31 May 2019; pp. 175–187.
35. Majeed, S.; Husain, H.; Samad, S.; Idbeaa, T. Mel frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: A comparison study. *J. Theor. Appl. Inf. Technol.* **2015**, *79*, 38.
36. Bradski, G. The OpenCV Library. *Dr. Dobb'S J. Softw. Tools* **2000**, *25*, 120–123.
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* 2825–2830.
38. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API design for machine learning software: Experiences from the scikit-learn project. In Proceedings of the ECML PKDD Workshop: Languages for Data Mining and Machine Learning, Prague, Czech Republic, 23–27 September 2013; pp. 108–122.
39. Chollet, F.; Rahman, F.; Zhu, Q.S.; Lee, T.; Marmiesse, G.D.; Zabluda, O.; Pumperla, M.; Santana, E.; McColgan, T.; Snelgrove, X.; et al. Available online: https://github.com/fchollet/keras (accessed on 20 July 2021).
40. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Nasr, S.; Quwaider, M.; Qureshi, R. Text-independent Speaker Recognition using Deep Neural Networks. In Proceedings of the 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 14–15 July 2021; pp. 517–521.
42. Jackson, Z.; Souza, C.; Flaks, J.; Pan, Y.; Nicolas, H.; Thite, A. Jakobovski/Free-Spoken-Digit-Dataset: V1. 0.8. 2018. Available online: https://zenodo.org/record/1342401#.YUdDMLhKjIV (accessed on 3 September 2021).
43. Oscar, CNNDigitReco-Speakerindependent 2020. Available online: https://www.kaggle.com/saztorralba/cnndigitreco-speakerindependent (accessed on 8 September 2021).
44. Inam Ul Haq, Classification on FSDD Using Spectograms. 2021. Available online: https://www.kaggle.com/iinaam/classification-on-fsdd-using-spectogramsInamurRehman (accessed on 13 September 2021).