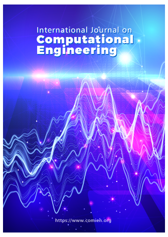
comien

International Journal on Computational Engineering

Journal Homepage www.comien.org/index.php/comien



Development of Yoruba Dialects Classification Model for Automatic Speech Recognition Systems Using KNN

Adejumobi O.K [a,*], Adenowo A.A [a,*], Yussuff A.I.O [a]

[a] *Department of Electronic and Computer Engineering, Lagos State University, Nigeria*

Corresponding author: [*] kolastar32@gmail.com

***Abstract*—This research presents, the development of Yoruba dialects classification Model for automatic speech recognition systems (ASRs) using K-Nearest Neighbor (K-NN). Research had revealed that ASRs perform better with correct dialects classification. Therefore, a non-parametric (i.e K-NN) model was developed and implemented on a Matlab 2021 platform to classify three (3) dialects (Ijebu, Ibadan and Ondo) from Ogun, Oyo and Ondo states respectively of Nigeria. The dialects were recorded at different environments, data sizes and at "opus file" format. They were later converted to ".wav" using the EZ CD Audio Converter Software. The Program4Pc Video Converter Pro was used to trim the converted audio waveforms to the same size and converted them to image signals suitable for model training, validation and testing. The results showed that the developed K-NN Classifier worked with an average performance accuracy of 91.11% and Recall {Sensitivity) of 86.67%. These results indicated that the model can be used to classify dialects of the same language hence, can help to improve the performance of robust ASR systems. However, for further improvement, better Classifiers that can handle large volumes of data should be employer.**

***Keywords*— Classification; dialects; K-Nearness neighbor; signal converter; speech recognition.**

Manuscript received 6 Jan. 2024; revised 27 Feb. 2024; accepted 15 Mar. 2024. Date of publication 31 Mar. 2024.

I. Introduction

Speech is the most natural natural means of human communications. Although ASR technologies has recorded considerable progress and improved comfort to developed countries, African languages are still at infancy. This degradable performance of ASR is attributed to non-cognizance of variability factors in its designs (Yusofet al 2013). This work is recommended to be extended to other accents and accuracy can also be increased in the future [1].

Wendy et al [2] revealed that, 'region of origin and amount of experience of listeners have great effects on dialect identification showing how well listeners are able to distinguish between Utah and non-Utah speakers of Western or non-Western States'. 'A two-stage language Chinese dialect identification system based on a shallow ResNet14 followed by a simple two-layer recurrent neural network (RNN) architecture' was presented by Zongze et al [3]. The results showed that the system can achieve high accuracy for Chinese dialects recognition under both short and long utterances conditions with less training time. Chittaragi et al [4] proposed an 'automatic dialect identification system for the Kannada language'. Spectral and prosodic features have captured the most prominent features for recognition of Kannada dialects. 'Support Vector Machine (SVM) and neural networks algorithms' are used for modeling text-independent recognition system. A neural network model that attempts for 'identification dialects based on sentence level cues' had also been built.

Kethireddy et al [5] proposed 'the use of frequency domain linear prediction cepstral coefficients (FDLPCCs) for dialect classification inspired by its long temporal summarization' during pole estimation. The results showed that there exists a complementary information between the proposed and baseline (MFCC's) Also, its performances are better than previous studies.

'A comparative study of different classifiers to recognize Malayalam language dialects' was presented by Sunija et al [6]. MFCC energy and pitch are the features extracted from both 'Thrissur and Kozhikode dialects' used for the recognition task. 'Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Naive Bayes classifiers' were used. The results showed ANN performed better than other classifiers. For further investigation the authors recommended that temporal differences in the dialect features of the dialects should be captured with small frames in the front end.

Bo Li et al. [7] presented 'a sequence-to-sequence model using listen, attend and spell (LAS)'. The authors explored the possibility of training a single model to serve different English dialects. Experimental results showed that the presented model is more effective in modeling dialect variations within a single LAS.

The study on testing the hypothesis that 'dialect differences in lexical processing reflects differences in lexical encoding strength across dialects' was carried out by Clopper et al. [8]. The authors carried out the experiments with 'Midland and Northern listeners in the Midland region' and the results showed that lexical information is more strongly encoded for the contextually-local Midland dialect than for the non-local Northern dialect. However, lexical processing is slower and less accurate for unfamiliar dialects than familiar dialects. Mohamed and Aly [9] presented 'a deep learning emotional recognition model for Arabic speech dialogues'. Here, 'audio representations - based wav2vec2.0 and HuBERT' were used [10]. The performance of the model overcome the previous known results

II. Material and Method

The developed Model was divided into four (4) main stages namely; dialects data acquisition, data pre-processing, dialects classification and Model evaluation (see Figure 1).

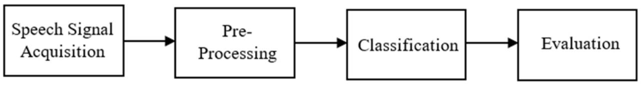


Fig. 1 Dialects Classification Model.

A. Dialects Data Acquisition

The following factors were taken into consideration when acquiring samples of the selected dialects:

Participants:
Samples of three (3) classes of Yorubadialects (i.e Ijebu, Ibadan and Ondo) were collected. 24 dialect samples of each of the classes were recorded making a total of 72 samples in the datastore (see Table 1).

Data Type:
Sentences, dialogues etc of participants were recorded.

Recording Environment:
The datasets were recorded at different offices, quiet rooms, telephone and radio programmes.

Speech Styles:
Reading, conversation etc.

Sampling Rate:
Duration of recording at different sample rates and data sizes.

Speech Format:
".opus file"

TABLE I
Dialects Dataset

S/N	Title	Participant	Data Type	Recording Environment	Speech Style	Sample Rate (s)	Data Size (kb)
1	Ibadan1	Ibadan	Recording	Room	Scripted	2	6
2	Ibadan2	Ibadan	Recording	Room	Scripted	12	31
3	Ibadan3	Ibadan	Voice note	Office	Scripted	2	6
4	Ibadan4	Ibadan	Voice note	Office	Scripted	11	27
5	Ibadan 5	Ibadan	Voice note	Office	Scripted	2	6
6	Ibadan6	Ibadan	Recording	Room	Scripted	10	24
7	Ibadan7	Ibadan	Recording	Room	Scripted	2	6
8	Ibadan8	Ibadan	Recording	Room	Scripted	13	32
9	Ibadan9	Ibadan	Voice note	Room	Scripted	2	7
10	Ibadan10	Ibadan	Voice note	Office	Scripted	1	29
11	Ibadan11	Ibadan	Voice note	Office	Scripted	2	7
12	Ibadan12	Ibadan	Voice note	Office	Scripted	12	31
13	Ibadan13	Ibadan	Voice note	Room	Scripted	12	29
14	Ibadan14	Ibadan	Voice note	Room	Scripted	13	154
15	Ibadan15	Ibadan	Voice note	Room	Scripted	2	25
16	Ibadan 16	Ibadan	Voice note	Room	Scripted	15	186
17	Ibadan17	Ibadan	Voice note	Room	Scripted	11	27
18	Ibadan 18	Ibadan	Voice note	Room	Scripted	15	185
19	Ibadan 19	Ibadan	Voice note	Room	Scripted	13	30
20	Ibadan 20	Ibadan	Voice note	Room	Scripted	2	6
21	Ibadan 21	Ibadan	Voice note	Office	Scripted	2	6
22	Ibadan 22	Ibadan	Voice note	Office	Scripted	16	37
23	Ibadan 23	Ibadan	Voice note	Office	Scripted	2	7
24	Ibadan24	Ibadan	Voice note	Office	Scripted	1	43
25	Ijebu 1	Ijebu	Voice note	Room	Scripted	2	6.61
26	Ijebu 2	Ijebu	Voice note	Room	Scripted	14	33.7
27	Ijebu 3	Ijebu	Voice note	Room	Scripted	3	6.81
28	Ijebu 4	Ijebu	Voice note	Room	Scripted	14	32.2
29	Ijebu 5	Ijebu	Recording	Room	Scripted	2	29.9
30	Ijebu 6	Ijebu	Recording	Room	Scripted	2	153
31	Ijebu 7	Ijebu	Voice note	Room	Scripted	2	29
32	Ijebu 8	Ijebu	Voice note	Room	Scripted	15	184
33	Ijebu 9	Ijebu	Voice note	Room	Scripted	s	5.33
34	Ijebu 10	Ijebu	Voice note	Room	Scripted	15	35.4
35	Ijebu 11	Ijebu	Voice note	Office	Scripted	2	5.69
36	Ijebu 12	Ijebu	Voice note	Office	Scripted	15	32.2
37	Ijebu 13	Ijebu	Voice note	Office	Scripted	3	35.3
38	Ijebu 14	Ijebu	Voice note	Office	Scripted	23	273

39	Ijebu 15	Ijebu	Voice note	Office	Scripted	1	17.4
40	Ijebu 16	Ijebu	Voice note	Office	Scripted	10	125
41	Ijebu 17	Ijebu	Voice note	Office	Scripted	2	28
42	Ijebu 18	Ijebu	Recording	Office	Scripted	19	232
43	Ijebu 19	Ijebu	Recording	Room	Scripted	3	31.1
44	Ijebu 20	Ijebu	Voice note	Room	Scripted	13	157
45	Ijebu 21	Ijebu	Voice note	Room	Scripted	3	38.3
46	Ijebu 22	Ijebu	Recording	Room	Scripted	15	183
47	Ijebu 23	Ijebu	Recording	Room	Scripted	3	39.3
48	Ijebu 24	Ijebu	Voice note	Room	Scripted	16	188
49	Ondo 1	Ondo	Voice note	Room	Scripted	2	7
50	Ondo 2	Ondo	Voice note	Room	Scripted	17	40
51	Ondo 3	Ondo	Voice note	Office	Scripted	2	6
52	Ondo 4	Ondo	Voice note	Office	Scripted	12	29
53	Ondo 5	Ondo	Voice note	Office	Scripted	2	4
54	Ondo 6	Ondo	Voice note	Office	Scripted	10	25
55	Ondo 7	Ondo	Voice note	Office	Scripted	1	5
56	Ondo 8	Ondo	Voice note	Office	Scripted	16	39
57	Ondo 9	Ondo	Recording	Office	Scripted	3	11
58	Ondo 10	Ondo	Recording	Office	Scripted	14	34
59	Ondo 11	Ondo	Recording	Office	Scripted	1	4
60	Ondo 12	Ondo	Voice note	Office	Scripted	10	25
61	Ondo 13	Ondo	Voice note	Office	Scripted	2	7
62	Ondo 14	Ondo	Voice note	Office	Scripted	15	36
63	Ondo 15	Ondo	Voice note	Office	Scripted	3	8
64	Ondo 16	Ondo	Voice note	Office	Scripted	13	31
65	Ondo 17	Ondo	Voice note	Office	Scripted	2	7
66	Ondo18	Ondo	Voice note	Office	Scripted	16	39
67	Ondo 19	Ondo	Voice note	Office	Scripted	2	40
68	Ondo 20	Ondo	Voice note	Office	Scripted	15	247
69	Ondo 21	Ondo	Voice note	Office	Scripted	3	9
70	Ondo 22	Ondo	Voice note	Office	Scripted	15	35
71	Ondo 23	Ondo	Voice note	Office	Scripted	3	36
72	Ondo 24	Ondo	Voice note	Office	Scripted	1	186

B. Data pre-processing

EZ CD Audio Converter Software was used to convert the input waveforms of audio samples recorded from “.opus file” format to “.wav” format (see Figure 2). The Program4Pc Video Converter Pro was used to trim the converted audio waveforms to the same size and converting them to image signals (see Figure 3).

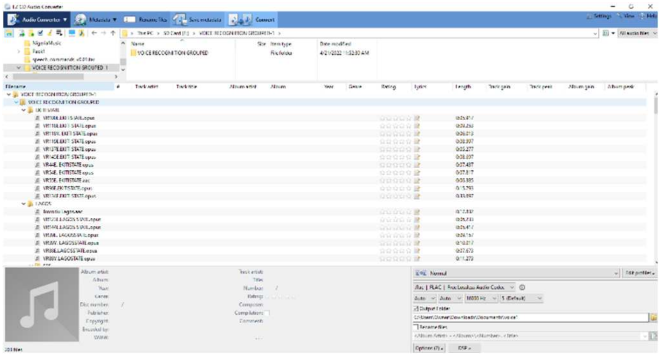


Fig. 2 EZ CD Audio Converter

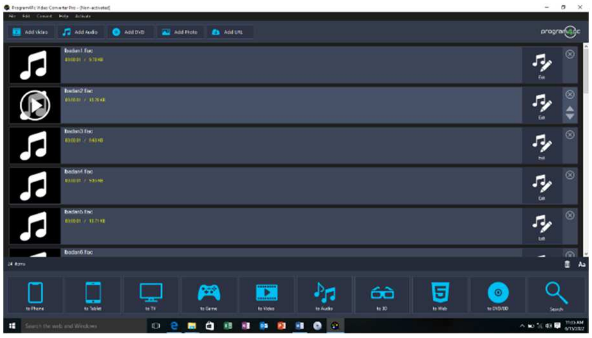
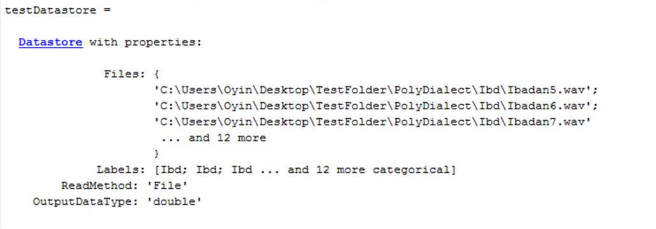


Fig. 3 Program4Pc Video Converter Pro

Datastore

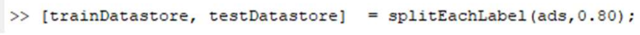
Datastore object was created to manage the database for training, validating and testing the Model as follows;



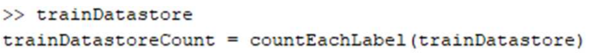
```
testDatastore =

  Datastore with properties:

             Files: {
                    'C:\Users\Oyin\Desktop\TestFolder\PolyDialect\Ibd\Ibadan5.wav';
                    'C:\Users\Oyin\Desktop\TestFolder\PolyDialect\Ibd\Ibadan6.wav';
                    'C:\Users\Oyin\Desktop\TestFolder\PolyDialect\Ibd\Ibadan7.wav'
                     ... and 12 more
                    }
            Labels: [Ibd; Ibd; Ibd ... and 12 more categorical]
        ReadMethod: 'File'
    OutputDataType: 'double'
```

Split Each Label Method

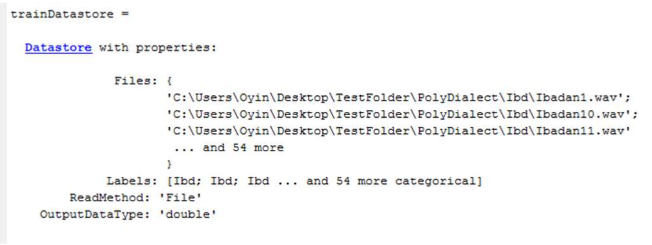
In this work, the datastore consists of 72 dialect samples (24 samples each of the three dialect classes). Each dialect class was split into two (2) parts; 19 dialect samples of each class were used for network training while 5 samples were used for network testing.



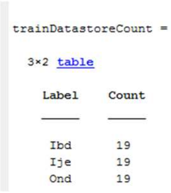
```
>> [trainDatastore, testDatastore]  = splitEachLabel(ads,0.80);
```



```
>> trainDatastore
trainDatastoreCount = countEachLabel(trainDatastore)
```



```
trainDatastore =

  Datastore with properties:

             Files: {
                    'C:\Users\Oyin\Desktop\TestFolder\PolyDialect\Ibd\Ibadan1.wav';
                    'C:\Users\Oyin\Desktop\TestFolder\PolyDialect\Ibd\Ibadan10.wav';
                    'C:\Users\Oyin\Desktop\TestFolder\PolyDialect\Ibd\Ibadan11.wav'
                     ... and 54 more
                    }
            Labels: [Ibd; Ibd; Ibd ... and 54 more categorical]
        ReadMethod: 'File'
    OutputDataType: 'double'
```



```
trainDatastoreCount =

  3×2 table

    Label    Count
    _____    _____

     Ibd      19
     Ije      19
     Ond      19
```

C. Dialects Classification

The K-NN model was developed with a K value of 5 for better performance.

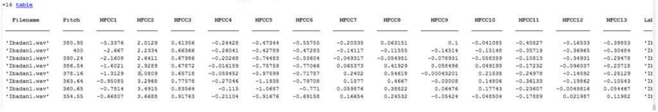
Feature extraction

Pitch and MFCC features were extracted from each frame using ComputePitchAndMFCC function which performs the following actions on the data read from each audio file: Collect the samples into frames of 30 ms with an overlap of 75%.

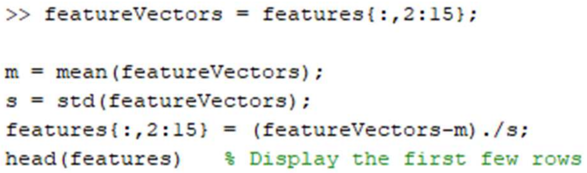
- For each frame, use audiopluginexample.SpeechPitchDetecto is Voiced Speech to decide whether the samples correspond to a voiced speech segment.
- Compute the pitch and 13 MFCCs (with the first MFCC coefficient replaced by log-energy of the audio signal) for the entire file.
- Keep the pitch and MFCC information pertaining to the voice frames only.
- Get the directory name for the file. This corresponds to the name of the dialect and will be used as a label for training the classifier.

ComputePitchAndMFCC returns a table containing the filename, pitch, MFCCs, and label (Dialect name) as columns for each 30 ms frame.

```
lenDataTrain = length(trainDatastore.Files);S
features = cell(lenDataTrain,1);
fori = 1:lenDataTrain
[dataTrain, infoTrain] = read(trainDatastore);
features{i} = ComputePitchAndMFCC(dataTrain,infoTrain);
end
features = vertcat(features{:});
features = rmmissing(features);
head(features)   % Display the first few rows
```



```
×16 table

     Filename       Pitch     MFCC1      MFCC2     MFCC3      MFCC4       MFCC5       MFCC6       MFCC7       MFCC8         MFCC9         MFCC10      MFCC11      MFCC12       MFCC13      Lal
  _____________    ______   ________   ______   _______   _________   ________   ________   _________   _________   ___________   _________   ________   __________   ________   ___

  'Ibadan1.wav'    380.95    -3.3376   2.0129   0.41356    -0.24428   -0.47344   -0.55755    -0.20335    0.063151           0.1   -0.041085   -0.40827     -0.16533   -0.39853   'Il
  'Ibadan1.wav'       400     -2.667   2.2334   0.66366    -0.26041   -0.42789   -0.47283    -0.14117    -0.11555      -0.14514    -0.15148   -0.35719     -0.36965   -0.30484   'Il
  'Ibadan1.wav'    390.24    -2.1608   2.6411   0.67986    -0.20268   -0.74483   -0.53604   -0.049317   -0.054951     -0.078931   -0.058359   -0.10813     -0.34931   -0.29478   'Il
  'Ibadan1.wav'    366.54    -1.6021   2.9289   0.67872   -0.016159   -0.75759   -0.77066    0.065373     0.41929      0.058496    0.049193   -0.17232    -0.096037   -0.23718   'Il
  'Ibadan1.wav'    378.16    -1.3129   3.0809   0.65718   -0.053452   -0.97599   -0.71757      0.2402     0.54618   -0.00043201     0.21538   -0.24978     -0.14592   -0.25129   'Il
  'Ibadan1.wav'    363.64   -0.95085   3.2968   0.77578    -0.27046    -1.1938   -0.79708      0.1577      0.4667      -0.03008     0.14806   -0.36133     -0.19862   -0.10543   'Il
  'Ibadan1.wav'    360.65    -0.7814   3.4915   0.83564      -0.115    -1.0687     -0.771    0.059876     0.38822       0.06476     0.17743   -0.23607   -0.0049816   0.054467   'Il
  'Ibadan1.wav'    354.55   -0.66807   3.6688   0.91763    -0.21104   -0.91676   -0.69158     0.16654     0.24532      -0.05424   -0.048504   -0.17889     0.021987    0.11982   'Il
```



```
>> featureVectors = features{:,2:15};

m = mean(featureVectors);
s = std(featureVectors);
features{:,2:15} = (featureVectors-m)./s;
head(features)   % Display the first few rows
```

D. Determination of Performance Evaluation of the Developed Model.

The performance evaluation of the Model was determined using equations 1 and 2.

$$Accuracy = \frac{\text{TP+TN}}{\text{Total Samples}} \quad (1)$$

$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{FN+TP}} \text{x100} \quad (2)$$

Where, TP, TN, FP and TP are True Positive, True Negative, False Positive and True Positive respectively.

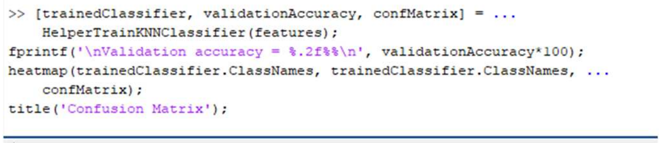
The total number of samples of each class - TP +FN FN for each class = sum of the corresponding rows excluding TP. FP = sum of corresponding column excluding TP TN = sum of all columns and rows excluding that class column and row.

III. Result and Discussion

A. Results of the classifier

The entire document should be Times New Roman at 10 points in size. Other font type and size may be used if needed for special purposes. Recommended font type and sizes are shown in Table 1.

After the features extraction for all dialects were performed, the network was trained using the K-Nearest Neighbor (KNN) Classifier. Figure 4 shows the waveforms of the dialect classes. Confusion Matrix was computed as shown in Figure 5. The model was tested using a new set of data of 15 dialect samples (i.e 5 dialect samples each of the three classes (see Table 2).



```
>> [trainedClassifier, validationAccuracy, confMatrix] = ...
    HelperTrainKNNClassifier(features);
fprintf('\nValidation accuracy = %.2f%%\n', validationAccuracy*100);
heatmap(trainedClassifier.ClassNames, trainedClassifier.ClassNames, ...
    confMatrix);
title('Confusion Matrix');
```

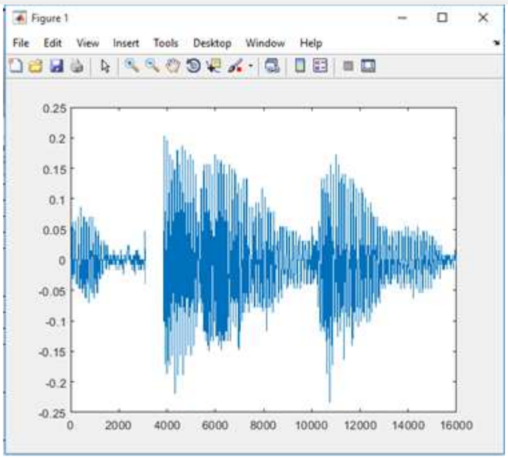
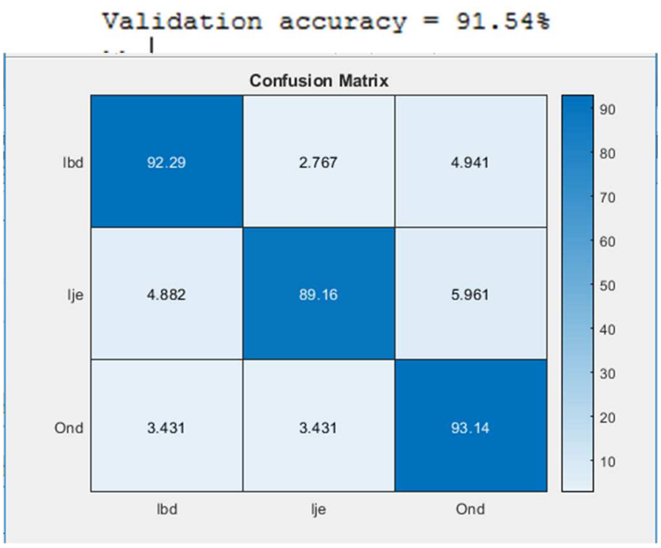


Fig. 4 Dialects Waveforms



```
Validation accuracy = 91.54%
```

Confusion Matrix

	Ibd	Ije	Ond
Ibd	92.29	2.767	4.941
Ije	4.882	89.16	5.961
Ond	3.431	3.431	93.14

Fig. 5 Confusion Matrix for Validation Data

TABLE II
CONFUSION MATRIX OF THE SPEECH SIGNALS PREDICTED

	Predicted		
Classes	Ibadan	Ijebu	Ondo
Ibadan	4	0	1
Ijebu	0	4	1
Ondo	0	1	4

B. *Results of Evaluation*

Considering Table 2, total samples = 15
Samples of each class = 5

IBADAN:

FP = 1
FN = 1
TN = 10
Tp = 4

$Accuracy = \frac{\text{TP+TN}}{\text{Total Samples}} = \frac{TP+TN}{TP+TN+FP+FN}$x100 $= \frac{4+10}{15}$ x 100 = 93.33%

Recall (Sensitivity) = $\frac{\text{TP}}{\text{FN+TP}}$ x100 = $\frac{4}{1+4}$ x 100 = 80%

IJEBU:

FP = 1
FN = 1
TN = 9
TP = 4

$Accuracy = \frac{\text{TP+TN}}{\text{Total Sample}} = \frac{TP+TN}{TP+TN+FP+FN}$x100 $= \frac{4+9}{15}$ x 100 = 86.67%

Recall (Sensitivity) = $\frac{\text{TP}}{\text{FN+TP}}$ x100 = $\frac{4}{1+4}$ x 100 = 80%

ONDO:

FP = 2
FN = 1
TN = 8
TP = 4

$Accuracy = \frac{\text{TP+TN}}{\text{Total Sample}} = \frac{TP+TN}{TP+TN+FP+FN}$x100 $= \frac{4+8}{15}$ x 100 = 93.33%

Recall (Sensitivity) = $\frac{\text{TP}}{\text{FN+TP}}$ x100 = $\frac{5}{0+5}$ x 100 = 100%

C. *Discussions*

This research presents the development of Yoruba dialects classification Model for automatic speech recognition systems using the KNN. To achieve the goal of this research, the work was divided into four (4) major stages namely: audio signals acquisition, data pre-processing, audio data classification and Model training, testing and evaluation (see Figure 1).

EZ CD Audio Converter Software was used to convert the input waveforms of audio samples recorded (Table 1) from ".opus file" format to ".wav" format (see Figure 2). The Program4Pc Video Converter Pro was used to trim the converted audio waveforms to the same size and converted them to image signals (see Figure 3). The datasets were divided into two (2), 19 samples each of the classes were used for training the network and 5 samples each for predictions.

Figure 4 shows the input dialect waveforms while Figure 5 shows the Confusion Matrix for training and validation data and audio Signals predicted. The validation accuracy is 91.54%. Table 1 shows the Confusion Matrix for model testing and prediction. Additional 15 audio samples (i.e 5 samples each of Ibadan, Ijebu and Ondo) were used for this purpose. From Table 2, 4 tested audio signals in IBADAN were correctly predicted while 1 was wrongly predicted as ONDO. For IJEBU, 4 were correctly predicted while 1 was wrongly predicted as ONDO. For EKITI, 4 were correctly predicted while 1 was wrongly predicted as IJEBU.

Accuracy and Recall (Sensitivity) were determined to evaluate the accuracy of the developed classification Model. Table 3 shows the summary of the calculated accuracy of the dialect Model developed. Accuracy obtained for IBADAN, IJEBU and ONDO was 93.33%,, 86.67% and 93.33% respectively while their Recalls (Sensitivities) are 80.00%, 80.00% and 100% respectively. Tables 4 and 5 showed the comparison of experimental and calculated predicted results and the average evaluation of the developed classification Model.

TABLE III
SUMMARY OF EVALUATION OF THE DEVELOPED MODEL

Classes	Accuracy (%)	Sensitivity (%)
Ibadan	93.33	80.00
Ijebu	86.67	80.00
Ondo	93.33	100.00

TABLE IV
COMPARISON OF EXPERIMENTAL AND CALCULATED PREDICTED RESULTS

Classes	Experimental Results (%)	Evaluated Results (%)
Ibadan	92.29	93.33
Ijebu	89.16	86.67
Ondo	93.14	93.33

TABLE V
AVERAGE PERFORMANCE OF THE CLASSIFICATION MODEL.

Method	Average Accuracy (%)	Average Sensitivity (%)
KNN	91.11	86.67

IV. Conclusion

In this research, a Yoruba dialects classification Model for an automatic speech recognition systems using KNN was developed. The Model classified three (3) south-western states' dialects namely; Ibadan, Ijebu, and Ondo. The KNN classification Model was implemented on MATLAB 2018 platform. The system was evaluated using accuracy and recall (specificity). An average performance of 91.11% accuracy and 86.67% sensitivity were achieved for the classification Model developed. The results showed that the KNN developed Model worked successfully. However, a more power classification Model such as convolutional Neural network (CNN) is recommended since K-NN is slow in learning and non-parametric. It also works on small datasets while facing problems when dealing with8 large datasets.

References

[1] Oladipo, FO., Habeeb, R.A. Musa, A.E. mezuruike, C. and Adeiza, O.A. (2021): "Automatic SpeechRecognition and Accent Identification of Ethnically Diverse Nigerian English Speakers". International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 12– No.36, May 2021 – www.ijais.org. Pp 41-48.

[2] Wendy Baker, David Eddington, Lyndsey Nay, (2009): "Dialect identification: The effects of region of origin and amount of experience". American Speech, American Dialect Society. Vol. 84, No. 1, Pp 48-71.

[3] ZongzeRen, Guofu Yang, ShugongXu, "Two-stage Training for Chinese Dialect Recognition", 2019 In Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China.

[4] Nagaratna B. Chittaragi, AsavariLimaye, N. T Chandana, B Annappa, Shashidhar G. Koolagudi, "Automatic Text-Independent Kannada Dialect Identification System", 2019 In Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, Mangalore, India.

[5] Rashmi Kethireddy, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty, "Exploration of temporal dynamics of frequency domain linear prediction cepstral coefficients for dialect classification",2021.

[6] Sunija A. P, Rajisha T.M, Riyas K.S, (2016): "Comparative Study of Different Classifiers for Malayalam Dialect Recognition System". Elsevier Ltd .Procedia Technology 24 (2016)Pp 1080-1088.

[7] Bo Li, Tara N. Sainath, Khe Chai Sim, MichielBacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yonghui Wu, Kanishka Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model",2017 in Google Inc., USA.

[8] Cynthia G. Clopper, et. al., "Variation in the strength of lexical encoding across dialects",2016.

[9] Mohamed O., and Aly S. A. "Arabic Speech Emotion Recognition From Saudi Dialect Corpus ", 2021 in Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia.

[10] Yusof, S.A.,,Atanda, A.F. and Hariharan, M. (2013): "A Review of Yorùbá Automatic Speech Recognition". 2013 IEEE 3rd International Conference on System Engineering and Technology, 19 - 20 Aug. 2013, Shah Alam, Malaysia. Pp 242-247.