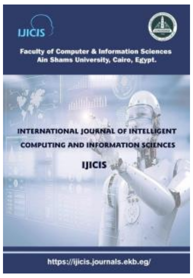
IJICIS, Vol.22, No.4, 25-34
DOI: 10.21608/ijicis.2022.152368.1207

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



Automatic Dialect identification of Spoken Arabic Speech using Deep Neural Networks

Mona A.Azim*

Information systems Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo,Egypt
monaabdelazim@cis.asu.edu.eg

Wedad Hussein

Information systems Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo,Egypt
wedad.hussein@cis.asu.edu.eg

Nagwa L.Badr

Information systems Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo,Egypt
nagwabadr@cis.asu.edu.eg

Received 2022-07-25; Revised 2022-08-31; Accepted 2022-09-01

Abstract: *Dialect identification is considered a subtask of the language identification problem and it is thought to be a more complex case due to the linguistic similarity between different dialects of the same language. In this paper, a novel approach is introduced for identifying three of the most used Arabic dialects: Egyptian, Levantine, and Gulf dialects. In this study, four experiments were conducted using different classification approaches that vary from simple classifiers such as Gaussian Naïve Bayes and Support Vector Machines to more complex classifiers using Deep Neural Networks (DNN). A features vector of 13 Mel cepstral coefficients (MFCCs) of the audio signals was used to train the classifiers using a multi-dialect parallel corpus. The experimental results showed that the proposed convolutional neural networks-based classifier has outperformed other classifiers in all three dialects. It has achieved an average improvement of 0.16, 0.19, and 0.19 in the Egyptian dialect, and of 0.07, 0.13, and 0.1 in the Gulf dialect, and of 0.52, 0.35, and 0.49 in the Levantine dialect for the Precision, recall and f1-score metrics respectively.*

Keywords: *Arabic Dialect, Automatic Speech Recognition , CNN, Dialect Identification.*

1. Introduction

In 1973, the United Nations officially recognised Arabic as a language [1]. It is used for work in many UN agencies and is the official spoken language of around 19 UN members. It is the mother tongue of more than 422 million residents in 22 different countries located in the Middle East. The Arabic dialect is composed of two different formats: The Dialectal Arabic (DA) and the Modern Standard Arabic

* Corresponding author: Mona A.Azim
Information systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo,Egypt
E-mail address: monaabdelazim@cis.asu.edu.eg

(MSA). The MSA form is the formal language that is used mainly in T.V. News shows, newspapers, most of the written books and official speeches. Although it is not the language that is being used on a daily basis, it is understood by all the Arabic Speakers. While the dialectal Arabic is the form that is being used in daily life and it differs from one region to another thus it is only spoken by the speakers living in the same community due to its variability. The same country may have more than one dialect thus it is difficult to compute the accurate number of dialects spoken all over the Arabic countries and most of the Arabic dialects cannot be understood among different speakers in different countries. The Dialectal Arabic can be divided into five main dialects according to the geographical basis. The main dialects include: Gulf, Levantine, Egyptian, Iraqi and Maghrebi [2]. In this paper, several classification approaches were investigate in the task of dialect identification. Four different classification approaches were used: Gaussian NB, SVM, RNN and CNN-RNN based approaches. The structure of the paper goes as follows: section 2 discuss the related work in the field of the Arabic dialect identification while section 3 introduces the classification approaches used during the study. The details of the proposed system and the experiments settings and results are described in section 4. Finally, the study conclusion is in section 5.

2. Related Work

In the field of dialectal Arabic speech identification, several studies took place using different approaches. In [3], the authors looked into various methods for Arabic dialect identification based on lexical and phonetic parameters that were obtained from an automatic Arabic speech recognition system. They first used a binary classifier to distinguish between dialectal Arabic and Modern Standard Arabic MSA, and then they used a multiclass classifier to distinguish between five different Arabic dialects: Egyptian, Gulf, Levantine, North African, and MSA. They achieved an accuracy of 59.2%. In [4], the authors developed a new technique for Arabic dialect identification that was based on identifying the distinctive features of many Arabic dialects by studying the repetitive sequences that make each one unique (motif). They extracted the 12 Mel Frequency Cepstral Coefficients (MFCCs) for each motif and applied their approach on different motif lengths. An enhancement of this approach was presented in [5]. The developed system aims to find the repeated sequences (motifs) of the Arabic dialect directly from the speech signal by representing it as a time series. For motif extraction task, it adopted the Scalable Time series Ordered-search Matrix(STOMP) for motif discovery. The system achieved a total accuracy of 62.75% compared to 60.2% that was achieved by a similar system in [6]. For the Algerian Arabic dialect, a deep neural network based approach was introduced in [7] to evaluate a web based corpus for the dialects of Algeria KALAM'DZ [8]. The results showed that the DNN based approach and the support vector based approach performed similarly. Recently, CNN based approaches have been used in the problem of dialect identification. In [9], a novel phonotactic based feature representation approach was presented to discriminate among various occurrences use various phone duration and probability statistics on the same phone n-grams. When compared to other systems that represent features using di-vectors with bottlenecks, the results showed that the used approach has reduced relative error rates by 24.7% and 19.0%.

3. System Architecture

Classification approaches have been used in several research fields to resolve different problems. For example, they have been used for weather forecasting [10], scene and image classification in computer vision [11,12,13] and protein classification in bioinformatics [14]. For the purpose of dialect identification, each dialect was handled as a separate class and then apply the classification approach. In this study, four different classification approaches were used: The Gaussian Naïve Bayes classifier, Support vector machines, Simple RNN based classifier and CNN-RNN Classifier. The details of each methodology is discussed below.

3.1 Gaussian Naïve Bayes:

Naive Bayes Classifiers are founded based on Bayes Theorem [15]. They assume the strong independence between data features. Gaussian Naïve Bayes classifiers assume that each class follows Gaussian distribution. The likelihood of features is assumed to be:

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)(1)$$

Where μ_y is the mean of the class y and σ^2 is the variance.

3.2 Support Vector Machine

Support Vector Machines (SVMs) are regarded as a classification strategy. that can be used in multiple fields: i.e. document classification, protein and cancer classification, face detection and Hand-written character recognition. The SVM main idea is to find a hyperplane in multidimensional space that separate several classes in the training dataset with a hyperplane that increases the difference between expected and true classes.

3.3 Simple Recurrent Neural Network Model:

Recurrent Neural Networks (RNNs) have proven significant effectiveness in modeling sequential data [16]. The RNN will take the time sequence of audio features as input then the output will be the probability that the i-th dialect is spoken in the given time sequence. RNN uses the earlier information in the input sequence to produce the current output of the given step. By processing more steps, RNNs may suffer from gradient vanishing problem rather than other networks architectures. To overcome this problem two customized versions of RNN were created: Gated Recurrent Unit(GRU) and Long Short Term Memory (LSTM). The GRU cell has two gates: reset gate and update gate. The update gate is responsible for defining the amount of previous information required to move to the next state, while the reset gate is responsible for defining which is needed to be neglected. LSTMs were also invented to solve the vanishing problem in RNNs. The LSTMs cells contains two additional gates than GRU cells which are: forget gate and output gate. The forget gate decides what is kept and what will be forgotten from previous state and the output gate decide which parts will be output to the hidden state. The diagrams of GRU and LSTM cells are shown in Figures and [17]. Although both cells are being used to resolve the gradient vanishing problem, there are slight differences between two types. The LSTM has three gates compared to the GRU's two. GRU doesn't maintain any internal memory because they lack the output gate found in LSTM.

In the LSTM, the input and target gates serve as the reset gates. In practice, LSTM works more efficiently with longer sequences but GRU cells requires less training time and less memory as it trains less number of parameters thus it works faster than LSTM. In this study, the RNN model used GRU cells was to perform the classification. The total number of parameters trained in this model was 11,745. The details of this architecture is shown in Figure. 3.

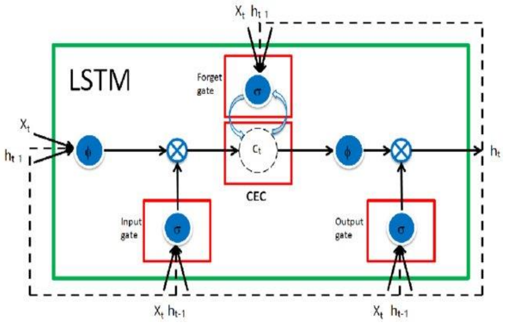


Figure. 1: The Architecture of LSTM Cell[17]

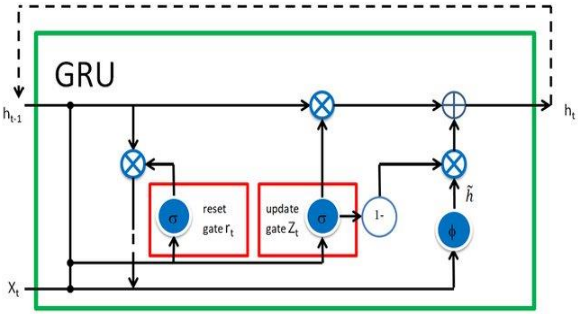
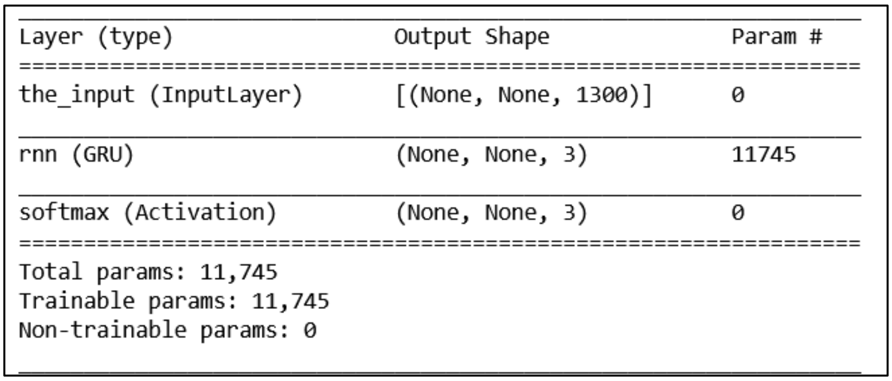


Figure. 2: The Architecture of GRU Cell[17]

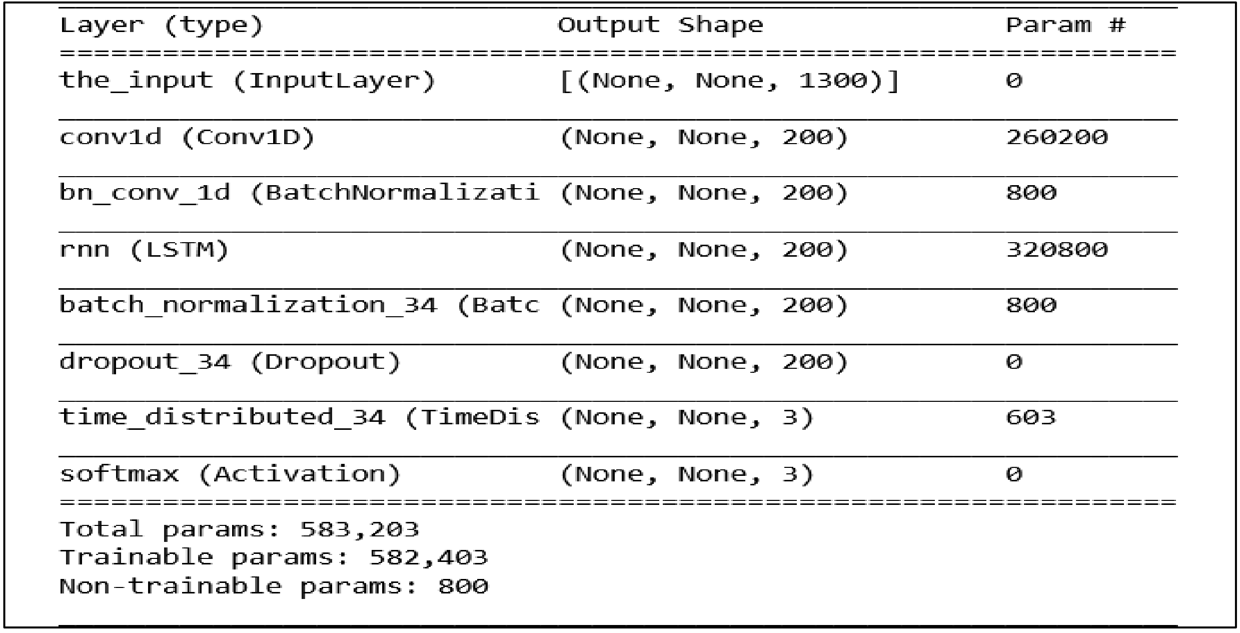
3.4 Proposed Convolutional-Recurrent Neural Network Model

In this model, an additional 1D convolutional layer was added to increase the network complexity. This layer is responsible for extracting the representable features from the acoustic features of the input signal. While the RNN part is responsible for predicting the class label for the input signal. The batch normalization is used for scaling the output of the previous layer by normalizing the activations of each variable in the mini batch. This helps in regulating the training of the network. The dropout layer helps in avoiding overfitting during the training by dropping off a partial set of the network neurons. This technique aids in building deeper and bigger network architectures that can make good predictions on un seen data. In this study, a 1D convolutional layer was added for the acoustic feature modeling, while the classification task was handled using the RNN part. The total number of parameters trained in this architecture was 583,203. The details of this architecture is shown in Figure. 4.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
the_input (InputLayer)       [(None, None, 1300)]      0
_________________________________________________________________
rnn (GRU)                    (None, None, 3)           11745
_________________________________________________________________
softmax (Activation)         (None, None, 3)           0
=================================================================
Total params: 11,745
Trainable params: 11,745
Non-trainable params: 0
_________________________________________________________________
```

Figure. 3: Simple RNN Model Architecture



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
the_input (InputLayer)       [(None, None, 1300)]      0
_________________________________________________________________
conv1d (Conv1D)              (None, None, 200)         260200
_________________________________________________________________
bn_conv_1d (BatchNormalizati (None, None, 200)         800
_________________________________________________________________
rnn (LSTM)                   (None, None, 200)         320800
_________________________________________________________________
batch_normalization_34 (Batc (None, None, 200)         800
_________________________________________________________________
dropout_34 (Dropout)         (None, None, 200)         0
_________________________________________________________________
time_distributed_34 (TimeDis (None, None, 3)           603
_________________________________________________________________
softmax (Activation)         (None, None, 3)           0
=================================================================
Total params: 583,203
Trainable params: 582,403
Non-trainable params: 800
_________________________________________________________________
```

Figure. 4: CNN-RNN Model Architecture

4. Experiments and Discussion

4.1 Data Set

In this study, a multi dialect Arabic speech parallel corpus was used [18]. The corpus includes four main dialects; Modern Standard Arabic (MSA), Gulf, Egypt and Levantine dialects. The text transcriptions were selected for a particular language area, namely travel and tourism.

There were 1291 recordings of the four main dialects used in the parallel transcriptions for the four dialects. Table. 1 presents a portion of the data set. 52 participants gave their consent for the recording of the audio files, which produced 67,132 speech files totaling roughly 32 speech hours. The corpus speakers and files count details are displayed in Table. 2. The corpus is divided into training set includes 90% of the audio files and testing set of 10% of the audio files.

Table 1 Data Set Samples

MSA	Gulf	Egyptian	Levantine	English translation
لا َأستطيع َسماع صوتك	ما َأقدر َأسمعك	أنا َمش َسامع صوتك	ما َبأدر َأسمع صوتك	I cannot hear your voice
مَتَى يُغلَقُ المَطعَم	مِتَى تسَكّرُون المَطعَم	المَطعَم بِيئفِل إمتَه	ايمَتْ بِسَكِّرْ المَطْعَم	When does the restaurant close?

Table 2 Data set Number pf Speakers and Files Count

Dialect	Speakers count	Files Count
MSA	12	15,492
LEV	8	10,328
Gulf	12	15,492
EGY	20	25,820
Total	52	67,132

4.2 Experiments Details and discussion

In order to conduct the identification task of Arabic dialects, several experiments took place to explore the best identification model for Arabic Dialects. The first experiment included using simple classifiers: Gaussian Naïve Bayes and Support vector machines. Then the second experiment conducted used the neural networks with two different architectures: RNN and CNN-RNN architectures. For each audio signal, the 13- mel frequency cepstral coefficients were calculated. Due to the variability of the audio lengths, only 100 frames were only used to feed the classifiers resulting into an input size of 1300. Precision, recall, and f1 scores were calculated for each model. to measure the system performance using Eq. (2), Eq. (3) and Eq. (4).

$$Precision = \frac{Number\ of\ correct\ predications}{Number\ of\ predications} \quad (2)$$

$$Recall = \frac{Number\ of\ correct\ predications}{Number\ of\ samples} (3)$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Table 3 Precision, Recall and F1-score of 4 Classifiers

Classifier Approach	Precision			Recall			F1-Score			Accuracy
	EGY	Gulf	Lev	EGY	Gulf	LEV	EGY	Gulf	Lev	
simple RNN	0.67	0.67	0.39	0.66	0.8	0.68	0.77	0.79	0.53	0.67
SVM	0.88	0.65	0.49	0.8	0.96	0.33	0.79	0.73	0.03	0.73
Gaussian NB	0.93	0.78	0.44	0.94	0.79	0.01	0.84	0.77	0.4	0.70
CNN-RNN	0.99	0.77	0.96	0.99	0.98	0.69	0.99	0.86	0.81	0.92
Average improvement	**0.16**	**0.07**	**0.52**	**0.19**	**0.13**	**0.35**	**0.19**	**0.097**	**0.49**	**0.22**

From the results shown in Table 3, the following conclusion can be addressed: CNN-RNN has outperformed the other classifiers with a precision of 0.99, 0.77 and 0.96, recall of 0.99, 0.98 and 0.69 and f1-score of 0.99, 0.86 and 0.81 for the EGY, Gulf and LEV dialects respectively. While the Simple RNN has achieved the least performance with precision of 0.67, 0.67 and 0.39, recall of 0.66, 0.8 and 0.68 and f1-score of 0.77, 0.79 and 0.53 for EGY, GULF and LEV dialects respectively. The Gaussian naïve Bayes classifier and Support vector machine classifier have achieved a moderate performance. From the graphical representation of the classifiers precision, recall and f1-score is shown in Figure. 6, Figure. 7 and Figure. 8 the LEV dialect achieved the lowest performance while the Egyptian dialect has achieved the maximum one and this is due to the number of speakers found in the dataset where the number of Egyptian speakers is 20 while the number of LEV speakers is only 8 speakers. The confusion matrix of the best model (CNNN-RNN model) is shown in Figure. 9 representing the number of the correctly classified samples in the matrix diagonal.According to table 4, the best performance conducted by the CNN-RNN architecture out performed the similar studies.

Table 4 Proposed Approach in comparison with similar systems

Related Work	Dialects	Dataset size	Acurracy
[3]	Egyptian, Gulf, Levantine, North African, and MSA	74 hours	59%
[5]	Egyptian, Gulf, Levantine and North African.	30 hours	62.75%
[6]	Egyptian, Gulf, Levantine and North African.	30 hours	60.2%
Proposed approach	Egyptian, Gulf, and Levantine.	32 hours	**92%**

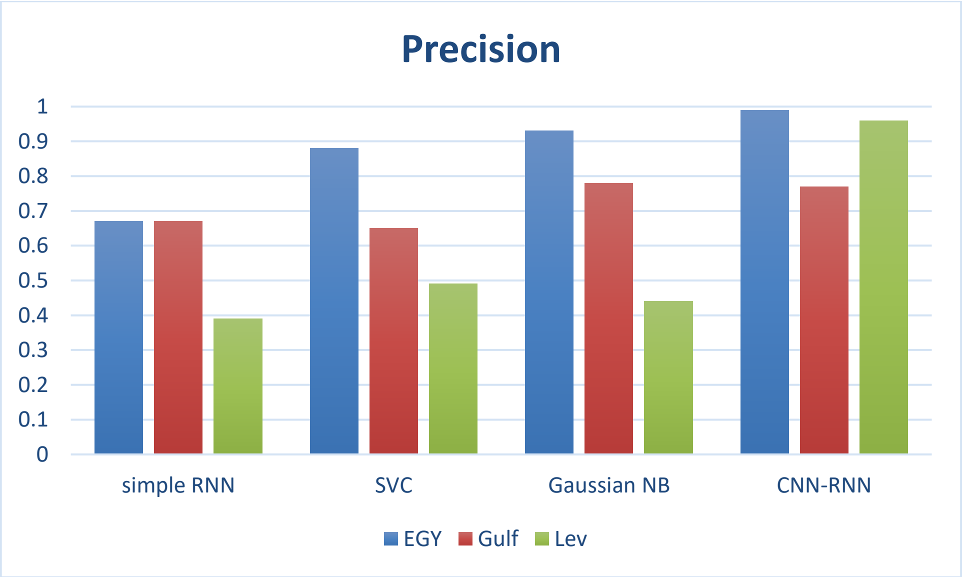


Figure. 5: Precision Results

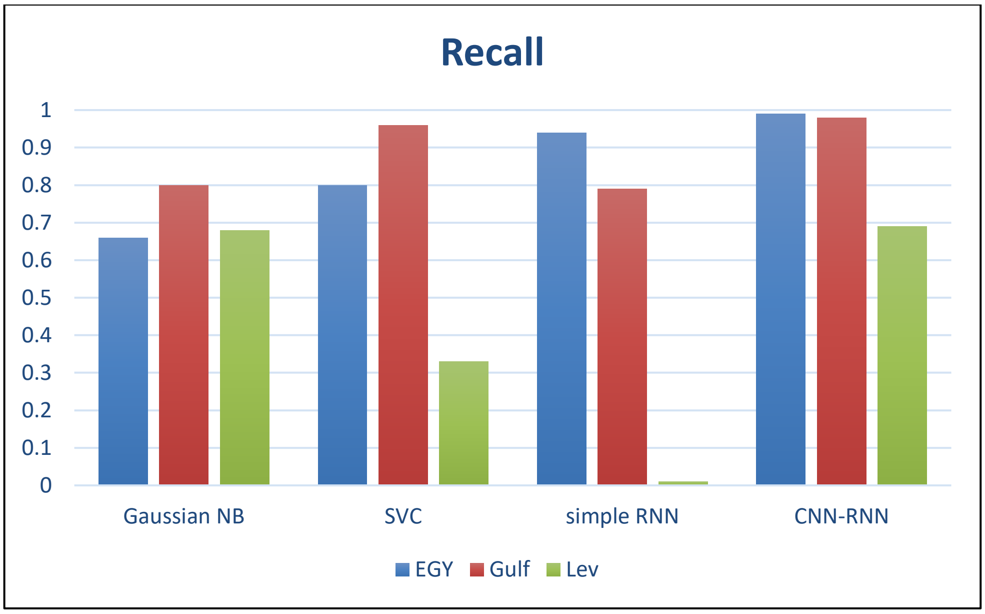


Figure. 6: Recall Results

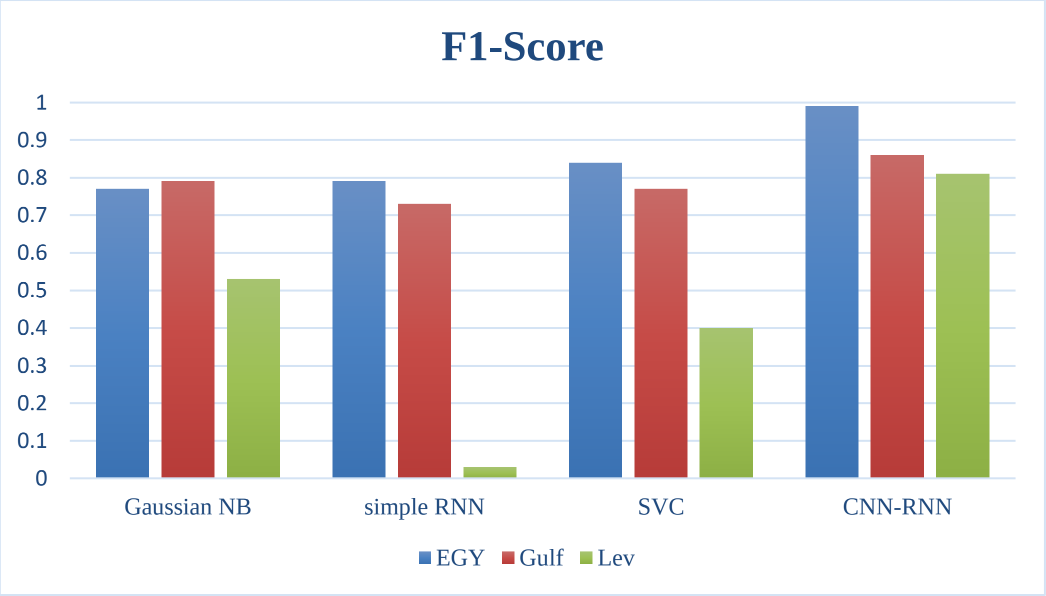


Figure. 7: F1-Score Results

5. Conclusion and Future Work

In this study, several experiments were conducted for the dialect identification problem in the Arabic Language. A proposed CNN-RNN dialect classifier was built to classify the input signal into one of the three main Arabic dialects used within the study: Egyptian, Levantine and Gulf Dialects. The proposed classifier results were compared with other three classifiers that included: Support vector machine, Gaussian Naïve Bayes, Simple and RNN model. A multi dialect corpus was used to train each classifier. The data set was divided into 90% that was used for training and 10% that was used for testing then the precision, recall and f1-score were calculated. From the experiments conducted, the maximum performance was achieved by the CNN-RNN model with a precision of 0.99, 0.77 and 0.96, recall of 0.99,0.98 and 0.69 and f1-score of 0.99, 0.86 and 0.81for the EGY, Gulf and LEV dialects respectively. While the Simple RNN has achieved least performance with precision of 0.67,0.67 and 0.39, recall of 0.66, 0.8 and 0.68 and f1-score of 0.77, 0.79 and 0.53 for EGY, GULF and LEV dialects respectively. The LEV dialect has achieved the lowest performance in all experiments due to the lack of number of speakers in the dataset while the Egyptian Dialect has achieved the maximum recognition accuracy as it has the maximum number of speakers in the dataset.

In the future, larger data with more number of speakers for the LEV and GULF Dialects will be used to enhance the recognition accuracy using the best architecture.

6. References

[1] Inclusion of Arabic among the official and the working languages of the General Assembly and its Main Committees. Adopted at the 2206th plenary meeting, 18 Dec. 1973.

[2] Ali, Ahmed, Hamdy Mubarak, and Stephan Vogel. "Advances in dialectal arabic speech recognition: A study using twitter to improve egyptianasr." *Proceedings of the 11thInternational Workshop on Spoken Language Translation: Papers*. 2014.

[3]Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., ... &Renals, S. (2015). Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

[4]Moftah, M., Fakhr, M. W., & El Ramly, S. (2018, April). Arabic dialect identification based on motif discovery using GMM-UBM with different motif lengths. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)* (pp. 1-6). IEEE.

[5]Moftah, M., Fakhre, M., & El-Ramly, S. (2018). Spoken Arabic Dialect Identification Using Motif Discovery. *The Egyptian Journal of Language Engineering, 5*(1), 25-36.

[6]Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., ... &Renals, S. (2015). Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

[7]Bougrine, H. C. S., &Abdelali, A. (2018, April). Spoken Arabic algerian dialect identification. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)* (pp. 1-6). IEEE.

[8]Bougrine, Soumia, et al. "Toward a Web-based speech corpus for Algerian Dialectal Arabic varieties." *Proceedings of the Third Arabic Natural Language Processing Workshop*. 2017.

[9]Najafian, Maryam, et al. "Exploiting convolutional neural networks for phonotactic based dialect identification." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[10]Ghaleb, M. S., Moushier, H., Shedeed, H., &Tolba, M. (2022). Weather Classification using Fusion Of Convolutional Neural Networks and Traditional Classification Methods. *International Journal of Intelligent Computing and Information Sciences*, 1-13.

[11]Soudy, M., Afify, Y., &Badr, N. (2022). RepConv: A novel architecture for image scene classification on Intel scenes dataset. *International Journal of Intelligent Computing and Information Sciences*, *22*(2), 63-73.

[12]Amin, A. E. (2021). Integrating Hexagonal Image Processing with Evidential Probabilistic Supervised Classification Technique to Improve Image Retrieval Systems. *International Journal of Intelligent Computing and Information Sciences*, *21*(3), 53-70.

[13]Ahmed, S. (2021). Comparison of Satellite Images Classification Techniques using Landsat-8 Data for Land Cover Extraction. International Journal of Intelligent Computing and Information Sciences, 21(3), 29-43.

[14]Mostafa, F., Afify, Y., Ismail, R., &Badr, N. (2022). UNCOVERING THE EFFECTS OF DATA VARIATION ON PROTEIN SEQUENCE CLASSIFICATION USING DEEP LEARNING. *International Journal of Intelligent Computing and Information Sciences*, 1-14.

[15]Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *403*.

[16] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.

[17]Su, Y., &Kuo, C. C. J. (2019). On extended long short-term memory and dependent bidirectional recurrent neural network. *Neurocomputing*, *356*, 151-161.

[18] Almeman, K., Lee, M., & Almiman, A. A. (2013, February). Multi dialect Arabic speech parallel corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)* (pp. 1-6). IEEE.