ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Utilizing Multi-Layer Perceptron for Language Identification

Ankita Tiwari

Department of Engineering Mathematics, College of Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram,

Andhra Pradesh, India.tdrankita@gmail.com

***Abstract*: -**

Speech-based language identification is one of the most promising areas. Language identification is the strategy for perceiving a particular language from short speech data. The wording of Telugu, Hindi, and English falls within this work's purview of language identification. Accuracy is the primary language identification problem in the literature. This study proposes a multi-layer perceptron with a sequential model where every epoch uses the Adam Optimizer to lower the error rate and boost accuracy. Mel Frequency Cepstral Coefficient is also utilized to separate highlights from sounds. With 85% accuracy, the proposed language identification model surpassed competing models in the literature.

Keywords: Mel Frequency Cepstral Coefficient, Sequential Model, Multi-Layer Perceptron, Adam Optimizer

INTRODUCTION

The task of language identification has become a focal point for researchers in the field, aiming to discern the language spoken by an opponent during communication. The widely accepted and most effective approaches to language identification involve two distinct phases. Initially, a deep learning-based model is developed to predict at least one dialect within a given dataset. Given the real-world scenario where not everyone possesses the ability to comprehend the language spoken by opponents during conversations, addressing the language identification problem becomes imperative. The deep learning-based model incorporates a bottleneck, typically a low-layered element serving as a feature extractor. I-vectors for each case are subsequently derived using these bottleneck features as information highlights [1]. While this work employs Mel-frequency cepstral coefficients (MFCC) for feature extraction, existing literature explores alternative methods such as Linear Predictive Coding (LPC) and Gammatone Filterbank Analysis [2-3]. The choice of feature extraction methods varies, with

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

some utilizing Multi-Layer Perceptron (MLP) due to its capability to handle numerous input features and other advantages, while others opt for Random Forests, Decision Trees, among others [2-3]. In instances where the test information only encompasses a subset of primary languages, MLP proves useful for quick adaptation. However, challenges arise when test languages are unidentified during the framework's development. In such cases, the system must possess the ability to classify a test as an unknown language if it does not align with any recognized dialects in the system [4-5]. The proposed language identification model adopts a Multi-Layer Perceptron through Sequential Modeling and Adam optimizer to enhance accuracy. I-vectors or embeddings may be integrated as a potential contribution to MFCC. Additionally, the Gaussian Mixture Model serves as a commonly employed backend for language discovery [6-7]. The model operates on calculating probabilities rather than likelihood ratios (LRs), making it applicable in closed-set scenarios. The two-covariance MFCC model anticipates a Gaussian distribution around a language-specific mean for the vector representing a signal. It further assumes that the language-specific means are similarly distributed by Gaussians [8-9]. Researchers like D. Zhu and M. Adda-Decker have explored the diversity of dialects within datasets, noting that some language dialects share commonalities with specific groups, while others exhibit no such connections. The extension of the MFCC technique is investigated, applying Gaussian assumptions to groups of dialects rather than individual ones [10-12]. The proposed work incorporates the MFCC model, drawing from various online datasets for language identification, including NIST LRE data, BABEL, KALAKA, and others [13-14]. This inclusion highlights that discriminative training of the standard MFCC model significantly enhances its generative capabilities. Section II delves into the existing literature in the domain, highlighting key challenges. Section III outlines the dataset and proposed methodology, while Section IV presents the results and discussions arising from the conducted study. Finally, Section V provides a summary of the conclusion and outlines future avenues for exploration.

LITERATURE REVIEW

In this section, a literature study in the domain of language identification is described.

C. Fan et al. outline a robust end-to-end speech recognition process using gated recurrent fusion with common training frameworks. One of the challenges encountered and addressed in this work was speech distortion problems affecting the speech recognition component [8].

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

M. Yousefi et al. propose block-based high-performance CNN

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

architectures to detect speech in audio streams with frames as short as 25ms. The frame-based model architecture effectively detects speech and produces highly accurate, precision and recall [9]. S. Herry et al. present a method for detecting language using a discriminatory approach and a temporal decision using neural network models, with the MFCC parameter as the evaluation parameter. This Model has an advantage in language pair discrimination becausethe units are defined so that they are common to all languages [10].

Leeet al. describe spoken language recognition as a process that automatically determines the language spoken in a speech sample. The main challenge faced was to extract prosodic features. Additionally, it was summarized that spoken language recognition provides excellent generalization abilities [11].

B. Duvenhage et al. describe how a naive Bayes classifier and character n-gram frequency have become the standard for language identification, highlighting its ability to accurately predict the language [12]. Z. Tang et al. identify language using phonetic temporal neural models based on anSVM (Support Vector Machine) and a naive Bayes classifier.The study mainly emphasized the advantages of phonetically aware systems and demonstrated that phonetic knowledge could generate phonetic features [13]. A study by F. Adeeba et al. identifies native languages from very short utterances using bidirectional LSTM (Long Short-Term MemoryNetwork). It proposes using spectrograms and cochleagramsto infer an Urdu speaker's native language from concisespeech utterances. Accurate language identification is helpfulfor a wide range of human-machine interfaces [14].Muthusamy et al., in their work, recognized language for long utterances and described methods such as softmax and performance evaluation carried out using the F1-score. The main challenge was incorrect input shapes. [15].

The summary of an accomplished literature review in tabularform is in Table I. In literature, the accuracy of the proposedworks could be more desirable due to incorrect input shapes and wrong selection optimizer. The other challenge is that the proposed models could be more precise in languageidentification because of the different dialects of the same language

DATASET AND PROPOSED METHODOLOGY

This section describes the used dataset and the proposed methodology.

Dataset Used: The database contains a wide range of languages. Thirty-two speakers, aged

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

22 to 76, were used to produce more than 16,000 utterances. Speech files in Kaggle are available in three different languages. The database includes terms from a 35-word range. The speech was sampled using a 16-bit A/D converter [16].

This work mainly focuses on overcoming the problem of inaccuracy due to incorrect input shapes and the wrong selection optimizer. The input shapes are handled using the Ravel function to make them more accurate than the original [17-19]. Most literary works use stochastic gradient descent optimization, which fails to get the desired and precise output. The time complexity of the stochastic gradient descent optimizer is also high [20-22].

TABLE I. LITERATURE REVIEW

Problem Statement	Solution	Dataset	Evaluation Parameters	Challenges	Advantage
Framework for joint training with gated recurrent fusion [8]	These gated recurrent fusion representations of noisy and improved characteristics.	AISHELL – 1	Speech Enhancement	The speech distortion problem affects the speech recognition component.	It would effectively address the issue of speech distortion.
Speech recognition using a block-based, high-performance CNN architecture [9]	The modelling of speech detection is addressed via a block-based CNN architecture.	GRID corpus	Precision, Recall, f-score	It is still difficult to identify speech segments and extract useful information from them.	This Model, with such high accuracy, could provide an effective solution.
Language Detection combining a discriminating approach [10]	The answer is based on neural network discrimination between language pairs.	Call Friend corpus	MFCC Parameter	It places a significant restriction because APDs require corpuses that are	One benefit of discriminating across language pairs is that they are specified using

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

				phonetically labelled.	the same set of units.
When a language is spoken, it can be identified automatically using a technique known as said language recognition.[11]	The GMM training procedure provides an answer to this issue.	Ethnologue Dataset	The primary evaluation measure is the average detection cost.	The main challenge is to extract prosodic features reliably.	It offers the benefit of solid generalization ability.
Identification of native language in brief utterances[12]	The use of features based on spectrograms and cochleagrams that are taken from short voice utterances	Single word utterances speech corpus	Accuracy, MFCC	The main challenge is data paucity.	It can be helpful in numerous human-machine voice interface
Using a phonetic temporal neural model to identify languages [13]	An SVM and a naïve bayes classifier for language identification	The AISHELL -1	Support vector machine	The PRLM method and the GMM/i-vector method require long test utterances.	We can obtain the benefit of using phonetic knowledge.
Recognizing the language present in the speech signal using DNN and BN [14]	Naïve Bayes is used to reducing character n-frequency speech features.	MSLT corpus	Accuracy is between 70-85%	Effects in pruning	By this, we can get better accuracy.
Recognizing speech for long-form utterances	Methods proposed 1. VAD	PMC articles, Kaggle	F1-score	The main challenge is input shapes	We can find language for long

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

[15]	2. Soft-Max			need to be corrected.	sentences.

Therefore, the Adam optimizer is utilized in the proposed work, producing better results and having lower time complexity. The brief explanation of the Adam optimizer is as follows:

Adam Optimizer: -Adaptive moment estimation or Adam is simply a combine of both momentum and RMSprop (root mean squared propagation). It acts upon:

i) Where m used as the gradient component, the exponential moving average of gradients (like in momentum), and

ii) The learning rate component divides the learning rate by the exponential moving average of squared gradients (like in RMSprop), which is the square root of v.

iii) Mathematical Formulas

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t + g}} \cdot \hat{m}_t$$

where, $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$ and $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

A. Proposed Methodology

This section discusses various methods used in the proposed work with their step-by-step working.

Multi-Layer Perceptron (MLP): MLP is an ANN (Artificial Neural Network) which consists of multiple layers which are interconnected nodes, each performing a nonlinear transformation on the input dataset. MLPs are supervised learning algorithms that can be used for classification and regression tasks. They are trained using backpropagation that adjusts the weights of the connections between the nodes to minimize the difference between the predicted output and the actual output. MLPs are widely used in various applications for their ability to learn complex nonlinear relationships in the input dataset [23].

Sequential Model: Sequential models are deep learning models commonly used for sequence prediction tasks. These models are based on a sequential structure, meaning the input data is processed sequentially, one element at a time [24].

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Step 1 gathers the audio dataset files from the Kaggle onlinerepository. In step 2, the audio files are passed for pre-processing to find the missing audio files, which convert the dataset from categorical to numerical format and then normalization of the dataset.

1) **Data Collection**

i. Input Dataset: [16]

2) **Pre-processing**

i. is_na() //to find missing audio files

ii. LableEncoder().fit_transform() //it converts fromcategorical to numerical data

iii. MinMaxScaler() //it will normalize the data

3) **Model Building**

model=Sequential()

//A sequential model is used to judge the entire Model instead ofthe complex kind

i. model.add(Dense(,input_shape=(40,)))

ii. model.add(Activation(''))

iii. model.add(Dropout())#Repeat the steps[1],[2],[3] untilthe completion of hidden layers

###final layer

iv. model.add(Dense(num_labels))

v. model.add(Activation('softmax'))

vi. model.compile(loss='categorical_crossentropy',metrics=['accuracy'],optimizer='adam')

4) **Training and Validation Phase**

The input dataset is passedfor training the Model, and thenvalidation data is provided to evaluate the model further. model.fit(X_train, y_train, batch_size=num_batch_size, epochs=num_epochs,
validation_data= (X_test, y_test), callbacks=[checkpointer],verbose=1)

5) **Evaluation Phase**

Accuracy

i. test_accuracy=model.evaluate(X_test,y_test,verbose=0)

ii. print(test_accuracy[1])

Confusion matrix

iii. tf.math.confusion_matrix(predictions,y_pred)

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

iv. classification report(this includes precision ,recall,f1-score)

print(classification_report(predictions, y_pred))

In step 3, a model is built using a multi-layer perceptron andsequential Model. In step 4, the data is trained, tested and validated to predict the output. In the final step, performance evaluation is accomplished based on accuracy, confusion matrix, precision, and recall as parameters.

The pseudo-code for the proposed Model is given below:

RESULTS AND DISCUSSIONS

Evaluation Parameters

1) ***Confusion Matrix:*** A confusion matrix is a table that is used to define the performance of a classification, as shownin Table II.

TABLE II. *CONFUSION MATRIX REPRESENTATION*

	Positive (1)	**Negative (0)**
Positive (1)	TP	FP
Negative (0)	FN	TN

2) ***Precision:*** The ability to not label an instance positive that is negative is called precision.

Precision – Accuracy of optimistic predictions.

Precision = TP/(TP + FP)

3) ***Recall:*** Recall is the ability to find all positive instances.

Recall = TP/(TP+FN)

4) ***F1-Score:*** F1 scores are usually lower than accuracy measures. Comparing models should be done using theweighted average of F1, not overall accuracy.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

5) ***Accuracy:*** The following definition is used for accuracy. Total number of predictions by the number of correct predictions.

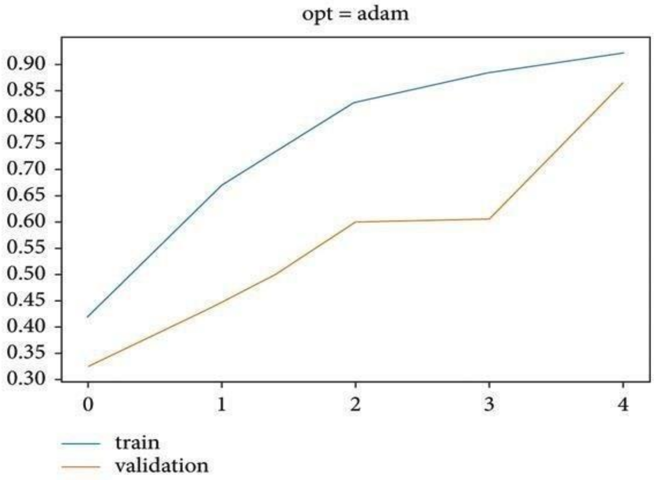

Fig. 1. *Training vs Testing*

Fig. 1. shows the use of the Adam optimizer, and the blue curve represents training accuracy while the yellow curve represents testing accuracy. Hence, according to the graph above, testing accuracy is 85% and training accuracy is also 85%. As a result, there is no overfit state because the accuracy of training and testing are both equal. So, this may be summarized that the model is reliable for predicting the specified respective languages.

The two types of gradient descent optimizers usually used inliterature are

1) Stochastic gradient descent
2) RMSprop

These optimizers are commonly used in deep learning models to update the neural network's weights through backpropagation. The results of both these optimizers are also compared to the Adam optimizer, as shown in Table III.

TABLE III. ***RESULTS OF DIFFERENT OPTIMIZERS***

Type of optimizer	**Accuracy**
Stochastic gradient-descent [14]	69%
RMSprop [11]	76%
Adam (*used in Proposed Model*)	85%

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

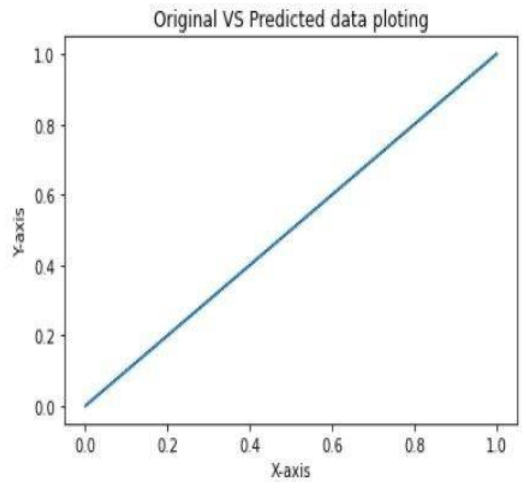

Fig. 2. *Predicted v/s Original*

In Fig. 2., the X-axis denotes the predictions, and the Y-axis represents the truth values or original (ranging from 1,0). This graph was generated by using y_pred and y_test. By observing the above graph y_test and y_pred are plotted in the same line. Thus, our Model is well-trained.

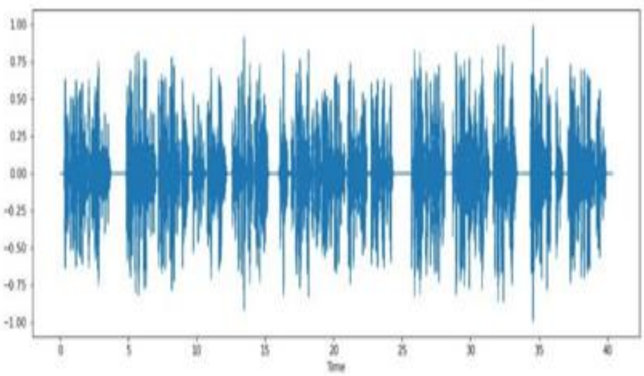

Fig. 3. *Frequency of audio sample*

In Fig. 3., on X- axis denotes the time (in Seconds), and Y- axis represents the audio (in decibels (dB)). This graph generates the audio sample frequency using the librosa.waveshow(). Librosa library is also used to extract data and the sample rate from the audio sample. Fig. 3. can be used to check whether the base and audio are shrinking. The overall results of the proposed Model are shown inTable IV.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

TABLE IV. **RESULTS OF EVALUATION PARAMETERS**

Model Name	**Accuracy**	**Confusion Matrix**	**Precision**	**Recall**	**F1-score**
Proposed model	85%	$\begin{bmatrix} 13 & 2 \\ 3 & 3 \end{bmatrix}$	81%	85%	80%

The accuracy of the proposed Model achieved is 85%, which means at training and testing, the Model went well, and a significant impact on the model performance can be determined.

CONCLUSION AND FUTURE SCOPE

The data is extracted using the librosa library and the MFCC. With MLP, we trained our Model on the Telugu, English and Hindi audio files dataset and achieved an accuracy of 85%. The proposed mode can be further improved and enhanced so that it can identify the language as well as the age and gender of the speaker. Using the GAN algorithm (Generative adversarial network), the cartoon design of the speaker canbe generated based on the frequency of an audio file.

REFERENCES

[1] P. Dalsgaard, O. Andersen, H. Hesselager and B. Petek, "Language- identification using language-dependent phonemes and language- independent speech units," Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, 1996, pp. 1808-1811 vol.3, doi: 10.1109/ICSLP.1996.607981.

[2] D. Farris, C. White and S. Khudanpur, "Sample selection for automatic language identification," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4225-4228, doi: 10.1109/ICASSP.2008.4518587.

[3] K. Markov and S. Nakamura, "Language identification with dynamic hidden Markov network," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4233-4236, doi: 10.1109/ICASSP.2008.4518589.

[4] M. C. Padma, P. A. Vijaya and P. Nagabhushan, "Language Identification from an Indian Multilingual Document Using Profile Features," 2009 International Conference

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

on Computer and Automation Engineering, 2009, pp. 332-335, doi: 10.1109/ICCAE.2009.35.

[5] Bo Yin, E. Ambikairajah and Fang Chen, "Improvements on hierarchical language identification based on automatic language clustering," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4241-4244, doi: 10.1109/ICASSP.2008.4518591.

[6] L. Sun, "Spoken Language Identification with Deep Temporal Neural Network and Multi-levels Discriminative Cues," 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), 2000, pp. 153-157, doi: 10.1109/ICICSP50920.2020.9232093.

[7] D. Zhu and M. Adda-Decker, "Language identification using lattice- based phonotactic and syllabotactic approaches," 2006 IEEEOdyssey- The Speaker and Language Recognition Workshop, 2006,pp. 14, doi: 10.1109/ODYSSEY.2006.248102.

[8] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu and Z. Wen, "Gated Recurrent Fusion with Joint Training Framework for Robust End-to-End Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 198-209, 2001, doi: 10.1109/TASLP.2020.3039600.

[9] M. Yousefi and J. H. L. Hansen, "Block-Based High Performance CNN Architectures for Frame-Level Overlapping Speech Detection," In IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 28-40, 2001, doi: 10.1109/TASLP.2020.3036237.

[10] S. Herry, C. Sedogbo, B. Gas and J. L. Zarader, "Language Detection combining discriminating approach and temporal decision with neural network modelling," 2006 IEEE Odyssey - The Speaker and Language Recognition Workshop, 2006, pp. 1-4, doi: 10.1109/ODYSSEY.2006.248107.

[11] Li, H., Ma, B., & Lee, K. A. (2013). Spoken Language Recognition:From Fundamentals to Practice. Proceedings of the IEEE, 101(5), 1136–1159. doi:10.1109/jproc.2012.2237151.

[12] B. Duvenhage, M. Ntini and P. Ramonyai, "Improved text language identification for the South African languages," 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), 2017, pp. 214-218, doi:

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019**

10.1109/RoboMech.2017.8261150.

[13] Z. Tang, D. Wang, Y. Chen, L. Li and A. Abel, "Phonetic Temporal Neural Model for Language Identification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 134-144, Jan. 2018, doi: 10.1109/TASLP.2017.2764271.

[14] F. Adeeba and S. Hussain, "Native Language Identification in Very Short Utterances Using Bidirectional Long Short-Term Memory Network," in IEEE Access, vol. 7, pp. 17098-17110, 2019, doi: 10.1109/ACCESS.2019.2896453.

[15] Muthusamy, Y.K., Barnard, E. and Cole, R.A., 1994. Reviewing automatic language identification. IEEE Signal Processing Magazine, 11(4), pp.33-41.

[16] https://www.kaggle.com/general/185386

[17] Kumar, S., Arpit Jain, Ambuj Kumar Agarwal, and Anshu Ghimire, "Object-Based Image Retrieval Using the U-Net-Based Neural Network," Computational Intelligence and Neuroscience, 2001.

[18] Sandeep, Shilpa, Arpit Jain, Chaman Verma, Maria Simona Raboaca, Zoltán Illés and Bogdan Constantin Neagu, "Face Spoofing, Age, Gender and Facial Expression Recognition Using Advance Neural Network Architecture-Based Biometric System, " Sensor Journal, vol. 22, no. 14, pp. 5160-5184, 2002.

[19] Kumar, Sandeep, Arpit Jain, Rani, Hammam Alshazly, Sahar Ahmed Idris, and Sami Bourouis, "Deep Neural Network Based Vehicle Detection and Classification of Aerial Images," Intelligent automation and soft computing , Vol. 34, no. 1, pp. 119-131, 2002.

[20] Kumar, Sandeep, Arpit Jain, Shilpa, Deepika Ghai, Swathi Achampeta, and P. Raja, "Enhanced SBIR based Re-Ranking and Relevance Feedback," in 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 7-12. IEEE, 2001.

[21] Harshitha, Gnyana, Kumar, Shilpa Choudhary, and Arpit Jain, "Cotton disease detection based on deep learning techniques," in 4th Smart Cities Symposium (SCS 2021), vol. 2021, pp. 496-501, 2001.