Iris Flower Classification

Rohan Das
Computer Science Engineering
SRMIST
Shillong, India
rr2704@srmist.edu.in

Md Safruddin Ansari
Computer Science Engineering
SRMIST
Giridih, India
ms7364@srmist.edu.in

*Dr. M. Kanchana, Associate Professor
Computer Science Engineering
SRMIST
Chennai, India
kanchanm@srmist.edu.in

***Abstract*—We use the classification technique in our project to identify and classify flowers from a given dataset. This technique is one of the foremost approaches to machine learning. The primary task of machine learning is data analysis. Our objective is to group the flowers using features. We employ algorithms like Support vector Machine(SVM), Random Forest Classifier, K-means clustering and K-medoids in this project. We also employ four features in addition to a new feature. They are sepal length and width and petal width and length. These features are measured in centimeters(cms). We use the scikit tool for the purpose of implementation. This project applies regression and classification algorithms on the given dataset by analyzing and discovering the patterns.**

Keywords—Classification, dataset, K-means clustering, supervised learning, K-medoids.

1. INTRODUCTION

The study of computational learning theory and pattern recognition in machine learning and artificial intelligence explores the algorithms that can learn as well as make the required data predictions, these algorithms overcome the program instructions by making decisions or predictions, by making a model from given sample inputs.

A method is presented for the identification of iris flower species. There are two phases, testing and training. During the training phase, machine learning models are loaded with data sets and after this the labels are assigned. The iris flower belongs to which group is predicted by the predictive model. Therefore, the iris species gets labeled. In this paper we focus on classification of the species of iris flower by using algorithms of machine learning along with sci-kit tools. In order to classify the iris data set we have to discover patterns by examining sepal size and petal size of the flowers. Then the prediction is made by inspecting the pattern to form the class of the iris flower. The training of the machine learning model is done by providing it with data sets and if any unrecognized data is discovered or seen then the machine learning model will predict the flower species from what it has learned by training the data. The recognition of iris species based on the flower characteristics is our task.

2. LITERATURE SURVEY

Zainab Iqbal uses Gaussian Naive Bayes algorithm to classify species of iris flower. A scatter matrix and scatter plot is created which gives us an analysis of the iris dataset. The algorithm is used along with python in the paper to classify the iris flower species. 95% accuracy is achieved which shows us that this algorithm is efficient for supervised learning classification.[2]

Joylin Priya Pinto uses KNN, SVM and Logistic Regression algorithms in order to get good accuracy results. She has applied the technique of cross validation in order to maximise the accuracy in her paper. She has focused on comparing the accuracy by including and excluding the technique of cross validation. Three algorithms are used which are K-Nearest Neighbour, Logistic Regression and SVM. Using these three algorithms we have found out that SVM is the most effective method among the others as it gives us the best accuracy.[5]

Shashidhar Halakatti made predictions in his paper on unseen data. The data which we do not use to train the machine learning model is called the unseen data. He proposed the identification of iris flowers using classification. We see models of machine learning which can predict features of the species accurately. He has trained data sets on the machine learning model and he has discovered a model for predicting iris species rather than labels.[6]

Patrick used the iris flower dataset to focus on the statistical analysis of it. In his paper they are analyzing two different methods. The dataset is plotted in order to determine the various patterns in classification. Then they are able to extract statistical information by developing an application in java.[3]

For classification Viashali used an efficient neural fuzzy approach. The technique proposed by her is connected to iris indexes and after that it groups the dataset into classes. Her system can choose the highlights which are great and for the grouping assignment remove a bit but satisfactory arrangement of standards.[11]

Bin Shi used the classification of iris flower problem as an example to show us that the photonic neural network concept permits us to get identical accuracy as compared to electronics. The final accuracy predicted gets reduced by 9.2%. A photonic DNN having three layers is used for image classification problems. A comprehensive error analysis suggests us that a chip on optical neural networks is implemented to expect to improve photonic neural performance. An analysis is worked upon to get to know the on-chip induced impairments.[14]

Poojitha in her work used neural networks to review iris flower data sets. Machine learning is a division in computer science. The iris dataset is already loaded and we classify it into three different categories. They used the k-means algorithm to group the dataset. Neural system is mainly used for grouping large sets of information. It is also used for highlight extraction, designing acknowledgment, quantization of vectors, work approximation, division of picture and mining of information. The results are grouped into three different iris species with no supervision.[8]

Deeptam Dutta in his work used a method to train artificial neural networks. In his paper the data set is classified with the help of neural systems. The problem statement reviews the recognition of species based on the estimations of bloom quality assessments. The task is to search for designs by processing sepal size and petal size of the iris. He used this example and this information can be used in the coming years. He processed that artificial neural systems can be effectively used for issues in work approximations, design arrangements, affiliated recollections and advancement.[1]

Swain in his work using the iris data set presented the procedure of developing the artificial neural network based on a classifier which groups the iris data sets. The multilayer perceptron neural network is used by this classifier to solve the classification problem.[17]

Ettaouil used a process for the iris dataset. The proposed model is applied to iris dataset. The iris plant is classified into three different species by using pattern classification technique. The architecture used in neural networks by using the back propagation algorithm is multilayer feed forward.[19]

Vyas used a different methodology for iris flowers. The problem assertion is to identify iris species based on the features. The multilayer feed-forward networks are used for training neural networks by using logistic function for the activation function and BP algorithm.[20]

Panwar presents us a study of using the existing data set of iris flowers using a neural network tool for clustering and classification. This gives the model the capacity to distinguish the three species of iris data sets.[9]

Borovinskiy used three different neural systems procedures by connecting them. The base model and neural system is 98%. Further, introduction of a coordinated grouping and characterisation shows us that the iris dataset gives us 98.66 % precision.[12]

L.P. Gagnani and K.H. Wandra used the WEKA data mining instruments with various artificial intelligence calculations such as Naïve Bayes, multi layer perceptron and RBF on iris dataset. Multilayer perceptron gives us better results of accuracy at 97.33 %.[7]

Mohan in his work used bolster vector machine strategies with many varieties on SVM. The iris dataset gave 96.7 % accuracy which is amazing for Q- SVM.[4]

Chang in his work used bunching calculations for order of iris flowers. Bunching is approached by the chart theory. The chart theory is utilized. He connected different scikit artificial intelligence apparatus with K-nearest neighbour. The calculated result on iris data set was 96%.[15]

Kumar in his work suggested us the adjustment of system loads using Particle Swarm Optimisation. He proposed a device to push the performance of artificial neural networks in implementation of the data sets that gives us a 97.3 % accuracy.[18]

Rathee in his work showed us a feed forward neural system model which is based on features connected to iris data set which gave us 98.3 % accuracy. When multilayer perception is used on the iris data set it gives 98.82 % accuracy.[16]

Alejandro in his work used a hybrid system which is based on hypothetical assessments that have been already presented and assessed. A group of blocks such as uniocular weighted adder and uniocular subtractor is described to use the self-educating applications in equipment. The results indicate that self-learning and classification tasks can be performed by the proposed hybrid solution using simple digital blocks. The hypothetical block works properly if a high intensity of noise is provided at the inputs because of its hypothetical nature.[13]

Nima in her work showed us how to improve machine learning algorithms in terms of required computational resources, success rate and energy. Also, the reliability of the use of machine learning techniques. She gave us a closer look at the existing models and analyzed them.[19]

Many studies have been conducted using various strategies for the identification of the iris flower species. A different strategy is used in every study. The problem is the classification and recognition of iris flower species based on its features.

We classify the iris data set by determining patterns after inspecting the features of the iris flowers and then we predict the processing of the patterns to form the iris flower class.

After using this classification and pattern the unknown data can be predicted in the coming years more precisely. The machine learning prototype for the iris flower species technique is loaded with the dataset belonging to iris flowers.

3. PROPOSED WORK

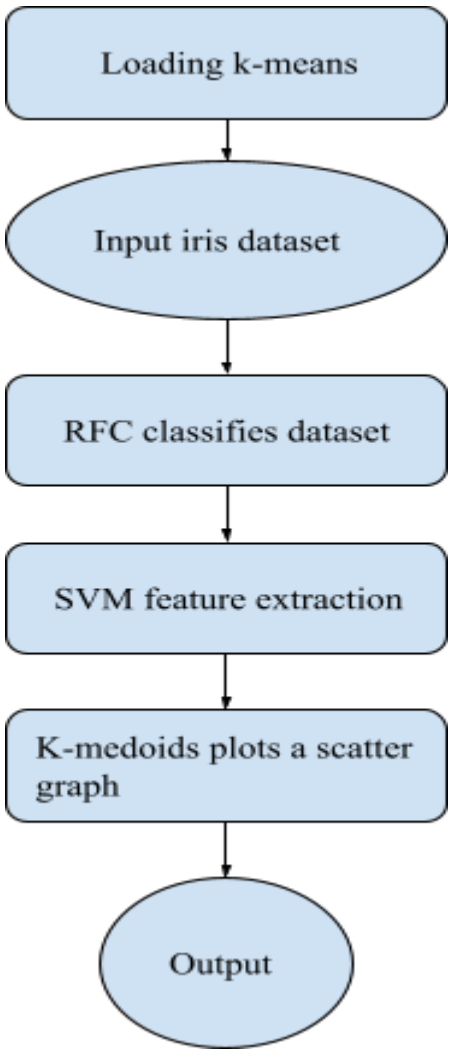

Fig. 3.1. Architecture Diagram

The architecture diagram of our project is shown in Fig. 3.1. We first load the k-means algorithm. We then input the iris data sets having 150 samples. We use the Random Forest Classifier for classification of the data set. After that we use the Support Vector Machine for feature extraction. Then we use k-medoids for plotting a graph. The last step is the final output which is displayed as a graph.

The execution using python and its various libraries makes the system accurate, fast and helps in future advancements. We upskill our prototype with our data. If any data is found or located which cannot be recognised then the iris species is predicted by the model from whatever it has learned using trained data.

3.1. Methodology

Data sets: The dataset contains 150 samples and these samples belong to the mentioned species: iris versicolor, iris setosa and iris virginica. This data is based on the famous Fisher's model and has become an important dataset for many uses in classification under machine learning. This dataset is included in the scikit-learn package. The rows are the samples and the columns are the iris flower features. We add an extra column in our dataset. This new column will contain the species colour of all flowers. The data set is loaded in the predictive model. The four features used to measure each of the sample are:

- petal length
- sepal width
- petal width
- sepal length

These four properties are measured in centimeters(cms). Using the four properties the iris species can be anticipated. The iris dataset is already loaded.

Fig. 3.1.1. Iris Flower Species

- Training: We use the dataset to train our model to predict output accurately. We focus on classifying the iris flower class by extraction of data from this dataset. The data provided is processed in a way of analyzing every parameter. In the machine learning process, data preprocessing is required so that the data gets converted to such a form that the machine can now easily understand it. We can say that the data features can now be easily interpreted by the algorithm. In our project the data gets converted to binary format. Now, the algorithm does the necessary computation. The output is converted from binary to hexadecimal format so that we can understand it. Each color has a unique hexadecimal code. The task is to classify the flowers based on the flower features.

- Testing: We use the Random Forest Classifier in our code to classify the testing data. After using the K-means algorithm we have found out that its accuracy is very high. We use the Random Forest Classifier to get the color codes of the flowers.

3.2. Algorithms used

- K-means clustering: It is an unsupervised learning algorithm and is the most common clustering method. It is faster than other clustering algorithms and we can use it for huge data sets. It allows us to group the data together or form clusters based on similar characteristics.

  We load the k-means algorithm first. The program inputs data by using the load iris function that is available from sklearn. The data gets split. We import the required libraries in our code. We calculate the Euclidean distance. The distance between the two points is called the Euclidean distance. Like this we calculate distances between many points. After comparing all these distances with one another we find the least distance. This method will form clusters of our dataset based on colour feature. We predict the accuracy now which is a number between zero and one. The accuracy is better if it is close to one.

  This algorithm finds meaningful structure among our data and discovers underlying patterns. It minimizes the total squared error.

- Random Forest Classifier: Random forest classifier works well for both regression and classification tasks. It can handle large datasets. It gives good accuracy and is fast in predicting. It reduces redundancy and so it removes duplicate values .

  Random Forest Classifier runs simultaneously with k-means. We input the iris data here. The task is to check the accuracy of k-means. If the accuracy is close to one then it allows the program to proceed to further steps. We use the iris decision tree classifier. The training of the data set is done after loading it. We classify the data set based on the decision tree.

  Random Forest Classifier uses the decision tree algorithm to go to the next step. Decision Tree is unsupervised and it is used for non linear data sets. It is used to determine action or course. Each branch of the decision tree depicts a probable end result. The prediction of the labels belonging to the data set is done based on the trained model.

- Support Vector Machine(SVM): It is a process for the ordering of linear and nonlinear knowledge. It uses the nonlinear mapping to modify the developing knowledge for advanced computation. It is used for both regression and classification problems.

  The features of a random dataset are plotted in a graph in Fig. 3.2.1. We use the scatter function to draw a scatter plot. The data is displayed as a collection of points. This algorithm is a part of supervised machine learning. We will use this algorithm to classify images. It requires full labelling of input data. It reduces the time complexity.

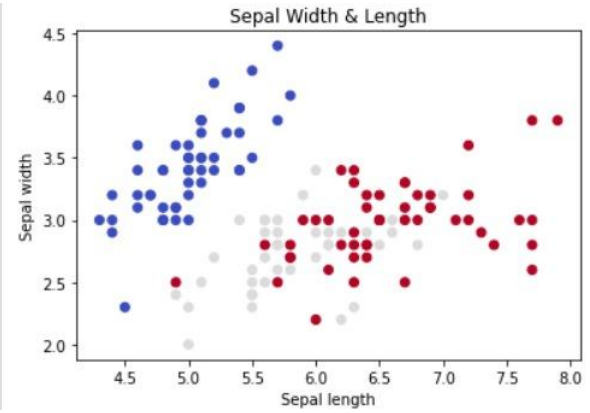

Fig. 3.2.1. SVM graph

- K-medoids: This is a clustering algorithm which is similar to k-means. The sum of dissimilarities between two points is minimised. The point labelled in a cluster and a point designated as the center of the cluster is minimised.

  We create a mesh plot using the mesh function that gives us a three-dimensional surface in a graph in Fig.3.2.2. We get a more precise graph compared to Fig. 3.2.1. by using the same feature extraction as in SVM. This method gives better prediction and reduces overfitting of models.

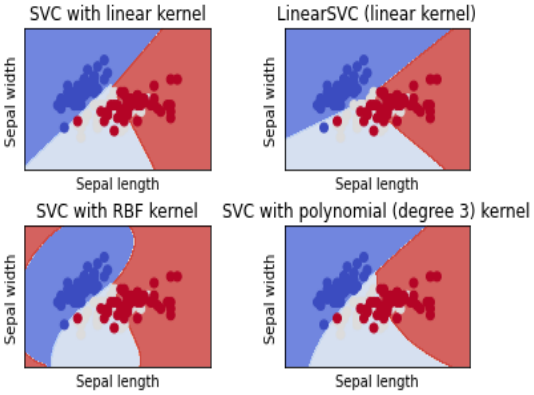

Fig. 3.2.2. K-medoids graph

4. IMPLEMENTATION

- Mathematical formula: We use the Euclidean distance formula in the K-means algorithm.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

The two coordinates are x and y. The distance is the variable d. The value of n is chosen randomly.

- Data collection: The data we have collected has 150 samples in total. Each of the samples belong to a particular iris species. Our data set contains the three iris species. We add color features in our data set.

- Data preprocessing: The data set is loaded into the model and it gets converted into binary form. Data preprocessing is that step in which data gets converted into that form which can be understood by the machine. A data set is a collection of data objects which are also known as records, entities, samples or vectors.

- Feature extraction: A feature is an individual characteristic of an observable phenomenon which can be measured. Color, power and mileage of a car are considered as features. It is a name for methods which combine variables into features so that it can reduce the amount of data that is to be processed. The process of feature extraction is beneficial if we want to reduce the assets which we need for processing without losing any germane information. Feature extraction can also reduce data redundancy.

Table 4.1. Dataset sample for iris setosa

sepal_length	sepal_width	petal_length	petal_width	species	species_color
5.1	3.5	1.4	0.2	Iris-setosa	#FF0000
4.9	3	1.4	0.2	Iris-setosa	#FF0000
4.7	3.2	1.3	0.2	Iris-setosa	#FF0000
4.6	3.1	1.5	0.2	Iris-setosa	#FF0000

This table 4.1 depicts iris setosa species from our dataset.

Table 4.2. Dataset sample for iris versicolor

sepal_length	sepal_width	petal_length	petal_width	species	species_color
6.5	2.8	4.6	1.5	Iris-versicolor	#0000FF
5.7	2.8	4.5	1.3	Iris-versicolor	#0000FF
6.3	3.3	4.7	1.6	Iris-versicolor	#0000FF
4.9	2.4	3.3	1	Iris-versicolor	#0000FF

Our table 4.2 depicts iris versicolor species from our dataset.

Table 4.3. Dataset sample for iris virginica

sepal_length	sepal_width	petal_length	petal_width	species	species_color
6.3	3.3	6	2.5	Iris-virginica	#FFFF00
5.8	2.7	5.1	1.9	Iris-virginica	#FFFF00
7.1	3	5.9	2.1	Iris-virginica	#FFFF00
6.3	2.9	5.6	1.8	Iris-virginica	#FFFF00

This table 4.3 depicts iris virginica species from our dataset.

In our paper we have gone through iris data set. We apply algorithms and then classify the entire data.

We are able to get the desired results based on the colour feature. The output is in hexadecimal format which is converted from binary form. We use python and its libraries which gives us accurate and fast results. Each color has a unique hash code.

Finally, we display the outputs and then we obtain the graph.

4.1. Results and Discussion

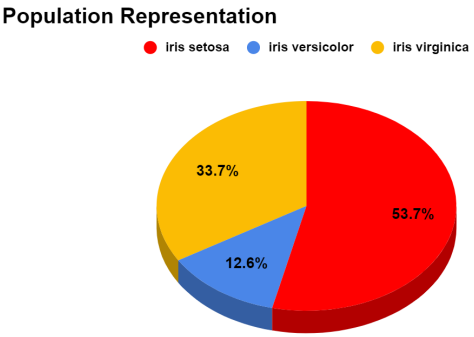

Fig. 4.1.1. Pie chart

The chart in Fig. 4.1.1 represents the population of our dataset. We show the species present in our dataset as a percentage out of the total datasets on the basis of K-means accuracy. Red color represents the iris setosa, iris versicolor for blue and yellow for iris virginica. Majority of the population is iris setosa, followed by iris virginica and iris versicolor respectively.

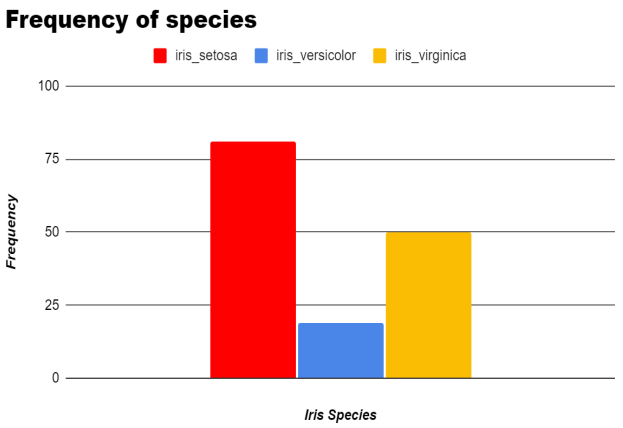

Fig. 4.1.2. Bar graph

This graph in Fig. 4.1.2. represents frequency of the species. The x-axis represents iris species and y-axis represents the frequency of species based on the colors. We see that iris setosa which is in red is the most frequent followed by iris virginica in yellow and then by iris versicolor in blue.

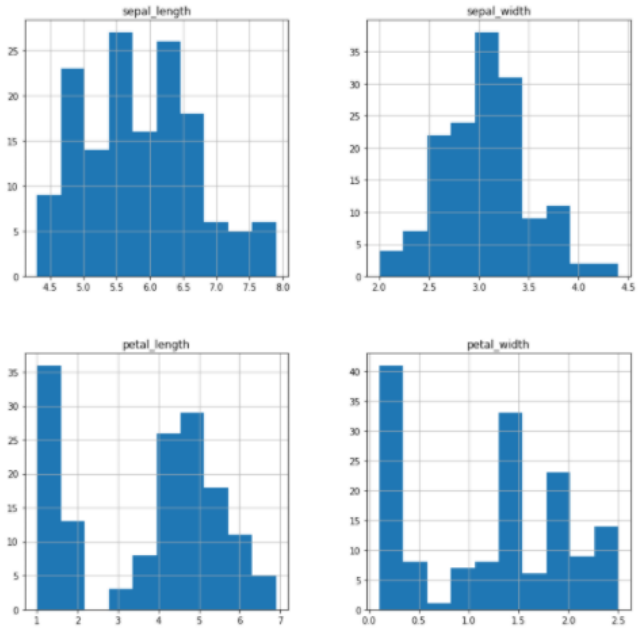

Fig. 4.1.3. Individual features

This graph in Fig. 4.1.3. represents the individual features or characteristics of the flowers in a graph. For each feature we have a separate graph.

5. CONCLUSION

Technology is developing rapidly. Many fields have used artificial intelligence. In order to achieve artificial intelligence, machine learning is the most basic approach. In our project our work includes loading the iris dataset, installation of Anaconda and using the algorithms to get our results. Our paper mentions the various algorithms used for the analysis of data sets. Random Forest Classifier, SVM and K-means algorithms are used to give great accuracy results. Using the three algorithms, we can conclude that SVM gives the best accuracy. K-means is a simple unsupervised algorithm which is used for clustering purposes. Our machine learning model classifies the flowers based on their colours by clustering them. Our paper indicates the importance of scikit-learn tools which are applied in machine learning.

REFERENCES

[1] Kaustav Choudary, Arga Roay, Dipptam Duta, “By Using Particle Swarm Optimization Algorithm Training the Artificial Neural Network”, Proc. IEEE , vol. 110, July 2020.
[2] Shilpi Jain, V Poojitha, “By Using Neural Network Clustering tool in MATLAB Collecting the IRIS Flower”, Proc. IEEE , vol. 109, 2020.
[3] K R Rathy, Arya Vaishali, “Classification of Dataset using Efficient Neural Fuzzy Approach”, vol. 099, August 2019.
[4] Shneiderman, Caard, S., Mackkinlay, J B. Readings in Information Visualization: Using Vision to Think, vol. 612, pp.1-34; 1999, Morgan Kaufmann Publishers, Inc., USA.
[5] Lichman and K, Bache M. 2013. Machine Learning Repository. Irvine, CA: School of information, Computer Science and University of California, vol. 516, 2015.
[6] A. L. Plana, Galluppi F, Furber S. B, Temple S, and L. A. Plana, “The Spinnaker project,” Proc. IEEE, vol. 102, no. 5, pp. 652–665, May 2014.
[7] R.A, Fisher. 1936. “Iris Data Set by using Machine Learning Repository”, vol. 116, May 2020.
[8] D. Decoste, E. Mjolsness. 2001. “State of the art and future prospects by using Machine Learning”, vol. 320, 2013.
[9] F and Varoquaux, Pedregosa, G. 2.11., Scikit-learn: machine learning in Python— Scipy lecture notes, vol. 220, Jan 2010.
[10] Duiin R. P. W, Taax D. M. J, Breukelen Van M and Kittler J. Combining multiple classifiers by averaging or by multiplying. Pattern Recognition, vol. 315, Feb 2019.
[11] Canny J. “Edge detection using Computational Approach. Machine Intelligence and Pattern Analysis”, vol. 440, June 2011.
[12] Shi P, He. X, An S. “Fake iris detection using support vector machines advances in biometrics by using Statistical texture analysis”, vol. 501, 2009.
[13] Gray J, Szalay A. “Science in an exponential world,” Nature, vol. 440, no. 7083, pp. 413–414, Mar. 2006.
[14] Lynch C, “How do your data grow? By Big Data” Nature, vol. 455, no. 7209, pp. 28–29, 2008.
[15] Wong P. H. S, Theis N. T, “A new beginning for information technology by Moore’s law,” Comput. Sci. Eng., vol. 19, no. 2, pp. 41–50, 2016.
[16] Chhugani J et al., “CPU myth vs Debunking the 100X GPU,” ACM SIGARCH Comput. Archit. News, vol. 38, no. 3, pp. 451–460, 2012.
[17] Shen Y. et al., “extreme scale systems are measured by Silicon Phonotics,” Lightw. J Technol., vol. 37, no. 2, pp. 245–259, Jan. 2019.
[18] Hines J, “Stepping up to summit,” Comput. Sci. Eng., vol. 20, no. 2, pp. 78–82, 2018.
[19] Martin C. K. A, Douglas J. R, “Recurrent neuronal circuits in the neocortex,” Current Biol., vol. 17, no. 13, pp. 496–500, 2004.
[20] Marr B, Hasler J, “To achieve large neuromorphic hardware systems using Finding a roadmap”, vol. 212, 2005.
[21] Merolla P. A. et al., “A scalable communication network and interface with a million spiking neuron integrated circuit” Science, vol. 345, no. 6197, pp. 668–673, Aug. 2014.
[22] Benjamin B. V. et al., “Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations”, vol. 150, 2007.
[23] C. Bishop, 2006. Machine Learning and Pattern Recognition. New York: Springer, pp.424-428, vol. 120, 2017.
[24] Neckar A. et al., “Braindrop: A dynamical systems-based programming model with a mixed signal neuromorphic architecture”, vol. 312, 2018.
[25] Zhang C. et al., “Optimizing FPGA-based accelerator design for deep convolutional neural networks,” in Proc. ACM/SIGDA Int. Symp. FieldProgrammable Gate Arrays, 2015, pp. 161–170, vol. 171, 2017.