International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Weighted Deep Neural Network Based Clinical Decision Support System for the Determination of Fetal Health

R.Vani

Abstract*: Healthcare industry is undergoing changes at a tremendous rate due to healthcare innovations. Predictive analytics is increasingly being used to diagnose the patient's ailments and provide actionable insights into already existing healthcare data. The paper looks at a decision support system for determining the health status of the foetus from cardiotographic data using deep learning neural networks. The foetal health records are classified as normal, suspect and pathological. As the multiclass cardiotographic datset of the foetus shows a high degree of imbalance a weighted deep neural network is applied. To overcome the accuracy paradox due to the multiclass imbalance, relevant metrics such as the sensitivity, specificity, F1 Score and Gmean are used to measure the performance of the classifier rather than accuracy. The metrics are applied to the individual classes to ensure that the positive cases are identified correctly. The weighted DNN based classifier is able to classify the positive instances with Gmean score of 91% which is better than than the SVM classifier.

***Keywords* : About four key words or phrases in alphabetical order, separated by commas.**

I. INTRODUCTION

With the arrival of the fourth industrial revolution, the healthcare domain continues to embark on a transformative technology journey. There have been huge positive changes across all major segments of the industry, including hospitals, pharmaceutical, diagnostics, medical equipment, medical insurance, and telemedicine. The growth of the healthcare industry has been catalyzed by the predictive analytics marked by the overabundance of data and widespread availability of tools and techniques. Predictive analytics based on machine learning methods use statistical techniques giving the machines the ability to "learn" with incoming data and to identify patterns and make decisions with minimal human direction. Clinical health processes incorporating such targeted analytics enable doctors to assess risks, make correct diagnoses, and offer patients more effective treatments. Role of artificial intelligence in this regard is significant.

Manuscript published on November 30, 2019.
* Correspondence Author
R.Vani*, Faculty, department of computer science (PG) at PSGR Krishnammal college for Women, Coimbatore.

Cardiotography is an effective and non-invasive technique for evaluating the foetal health. The unborn foetus heart rate and the mother's uterine contractions are recorded on paper similar to ECG. This technique is cost effective and can lead to early identification of pathological states such as congenital heart defects, hypoxia or foetal distress. The risk conditions can be identified in early stages so that the obstetrician can intervene to provide remedial measures before more damage is done to the developing foetus. Due to inconsistencies in the interpretation of the signals a large number of clinical decision support systems have been developed to augment the identification of the pathological and suspect states of the foetus. In this paper, deep learning neural network has been developed to identify the CTG recordings using the cardiotographic dataset available in the UCI machine learning repository. [1] The dataset has around 2126 samples and has 21 attributes. The data is imbalanced with around 1655 samples for the normal state, 176 samples belonging to the pathological state and 295 samples belong to the suspect state.

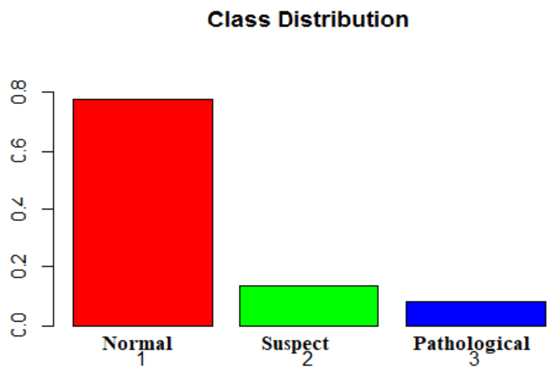

Fig 1. Imbalanced Class Distribution of the CTG Dataset

A weighted deep neural network is used to classify the various states to overcome the imbalance in the dataset. The dataset faces the accuracy paradox as number of positive instances of the pathological condition is significantly much less than the normal instances causing the dataset to be imbalanced or skewed. The **accuracy paradox** for predictive analytics states that the predictive models with the given level of accuracy may have greater predictive power than models which have higher levels of accuracy.[2]

Using the accuracy as a performance metric does not assist us in determining if the classifier was successful in determining the positive states effectively as it includes the performance of the classifier on both the positive and negative cases. Metrics such as sensitivity, precision, F1 score and G-Mean are used to determine the validity of the predictions on the three classes. [3] The rest of the paper is organized as follows. Section II discusses related work done for determination of foetal health. Section III takes a look at the deep neural network classifier used for the study. Section IV discusses the dataset. Section V provides information on the performance metrics used for evaluation of the models. Section VI reviews the results obtained on the CTG datasets. Section VII concludes the entire research done in this paper.

II. REVIEW OF LITERATURE

Various research has been undertaken in the healthcare domain for detection of foetal distress. Sunder et al. in their work have used a supervised ANN and support vector machine to classify the CTG dataset. The model performs well in classifying normal cases. However for the abnormal states the performance measures are lesser than the normal cases. [4] A novel clinical decision support system was proposed by Sindhu et al. [5] in their work for evaluating the fetalwell-being from the cardiotocogram dataset(CTG). The paper uses a search algorithm with three different fitness functions (two single objective fitness functions and multi-objective fitness function) to assess its performance. The resulting classification accuracy of 94% is obtained with an optimal feature subset using IAGA. They also compare the benchmark datasets to further strengthen their findings. Divya Bhatnagar et al [6] in their paper provide an analysis of CTG data set and generated classification rules to identify normal, suspicious and pathological cases using WEKA classifiers. In their comparative study on machine learning techniques, Z.Comert et al. [7] compare models based on artificial neural network, support vector machine, extreme learning machine, radial basis function network, and random forest. They found that ANN yielded the best results with sensitivity of 99.73% and specificity of 97.94%. They use a restricted two class dataset instead of the three class and hence achieve a higher accuracy measure in their result. In their paper Belal Amin et al. [8] use Rough Neural Network in classifying the cardiotocography dataset. Accuracy rate and consumed time during the classification process are used as performance metrics using the WEKA tool on the CTG dataset. The paper also compares the different algorithms such as the neural network, decision table, bagging, the nearest neighbor, decision stump and least square support vector machine algorithm with the RNN model.

III. DATASET DESCRIPTION

The CTG dataset consists of measurements of FHR and UC for the fetus, the important features of cardiotocograms classified by an obstetricians' expert, and the data set is available publicly at the data mining repository of University of California Irvine (UCI). Data set was split into training data and testing data with percentages 70% and 30% respectively. The data set has 21 attributes and classified according to the FHR pattern or fetal state class code [3, 4]. In this study, fetal state class code is used as the target attribute instead of FHR pattern class code and classification into one of three groups Normal, Suspicious or Pathological (NSP) classes. The dataset includes a total of 2126 samples. Attributes description of the dataset is given in Table I.

Table 1. CTG Dataset Attributes

Attribute	Description
LB	Fetal Heart Rate baseline (beats per minute)
AC	number of accelerations per second
FM	number of fetal movements per second
UC	number of uterine contractions per second
DL	number of light decelerations per second
DS	number of severe decelerations per second
DP	number of prolonged decelerations per second
ASTV	percentage of time with abnormal short term variability
MSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
MLTV	mean value of long term variability
Width	width of FHR histogram
Min	minimum of FHR histogram
Max	maximum of FHR histogram
Nmax	number of histogram peaks
Nzeros	number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency
NSP	fetal state class code (N = normal; S = suspicious ; P = pathologic)

IV. RESEARCH METHODOLOGY USED

For the CTG data set we perform tests using support vector machines and the Deep Learning Neural Network based on Keras with Tensorflow as the backend. The objective is to ensure that the positive cases are identified and isolate metrics for successful identification of the positive metrics. Performance metrics used as discussed in section V.

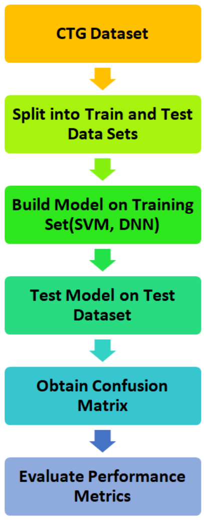

Fig 2. Research Methodology adopted.

A. Support Vector Machines

Support Vector Machine (SVM) is a discriminative classifier defined in formal terms by a separating hyper plane.

When labeled training data (supervised learning) is used, the SVM algorithm outputs an optimal hyper plane which is capable of categorizing new examples. SVM give the largest minimum distance to the training examples. Twice, this distance is termed as the Margin. Hence, the optimal separating hyper plane maximizes the margin of the training data [10]. Various SVM algorithms use different types of kernel functions including linear, nonlinear, polynomial, radial basis function and sigmoid. We use radial basis functions for implementation of the SVM model. Radial basis kernel is also known as the Gaussian Kernel, is non linear and can be used to classify data that is non-linearly separable. C type classification is a regularization parameter which helps impose a penalty on the misclassifications that are done during the classification process.

B. Deep Neural Networks

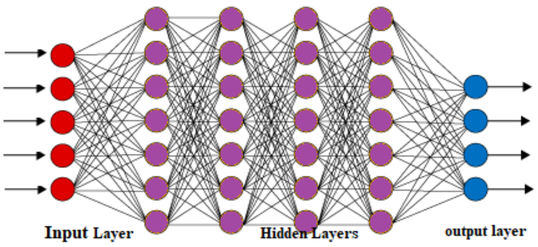

Fig 3. Deep Neural network with multiple hidden layers.

Deep-learning networks have significantly more hidden layers than shallow models which mostly have single hidden layer. The earlier versions of the neural networks such as the first perceptrons are mostly composed of a single input layer and single output layer with at the most a hidden layer inbetween. The deep neural networks learn to identify the features based on their depth. Also a significantly large dataset performs much better than smaller datasets as it is capable of providing a larger amount of training data to the DNN. When training on unlabeled data, each node layer in a deep network learns features automatically by repeatedly trying to reconstruct the input which it draws its samples, attempting to minimize the difference between the network's guesses and the probability distribution of the input data itself. This enables them to create so-called reconstructions in this manner. In the process, these neural networks learn to recognize correlations between certain relevant features and optimal results. They draw connections between feature signals and what those features represent, whether it is a full reconstruction, or with labeled data. We compare outputs from both the models – SVM as well as DNN and evaluate them based on the performance metrics discussed in section V.

V. PERFORMANCE METRICS

For imbalanced data sets accuracy metric may not be suitable. Most common metrics used as Geometric Mean and F1 score which can help us evaluate the strength of the classifier for skewed data. In our work, we use True positive Rate (Sensitivity) and True Negative (Specificity) rate together with Geometric Mean as metrics to evaluate the performance of the algorithms. **True positive rate** also known as, **sensitivity** given by formula (1) **measures** the proportion of positives that are correctly identified as the percentage of cases which are correctly identified as having the condition. **True Negative Rate** (TNR) also known as **specificity** given in (2)**,** measures the proportion of negatives that are correctly identified as such as the percentage of healthy people who are correctly identified as not having the condition. Precision (3) and F1 score (4) are also tracked. However, these metrics cannot be key indicators of the fetal health as the help predict the normal states rather than the abnormal states. The geometric mean G-mean is the product of the prediction accuracies for both classes, i.e. sensitivity and Specificity as given in (5). As accuracy is a poor indicator of the performance of the algorithm on the positive samples, G-mean overcomes the problem since the value is dependent on both the positive and the negative samples in the dataset. Loss of accuracy in identifying negative cases may be tolerated for the above area.

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad (1)$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad (2)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (3)$$

$$\text{F1Score} = 2 * \frac{TPR * TNR}{(TPR + TNR)} \quad (4)$$

$$\text{Gmean} = \sqrt{TPR * TNR} \quad (5)$$

TP is the true positives, TN is the True Negatives, FP is the False Positives and FN is the False Negatives. TPR and TNR represent the True positive Rate and the True Negative Rate.

VI. RESULTS AND DISCUSSION

Models were simulated in R using the R studio package. For training phase, 70% of the data was allocated randomly and the remaining 30% used for testing.

SVM model was implemented using Radial Basis based Kernel and C type classification. We obtain an accuracy of 91% with a misclassification error of 8%. Resulting accuracy and metrics are provided in the Table 2. The SVM model identifies the negative cases with high accuracy but the accuracy is lesser for the suspect and the pathological cases.

The DNN model is trained using the 70 – 30 split for the training and validation samples. Keras with the Tensor flow backend allows the sequential models to be created which allows the different layers to be stacked up in a linear fashion. [9] Optimiser used is the adam optimizer which is an optimization algorithm that can be used instead of the usual stochastic gradient descent procedure to update network weights iteratively for the training data.

```
summary(model)
Layer (type)                 Output Shape              Param #
================================================================
dense_6 (Dense)              (None, 8)                 176
________________________________________________________________
dense_7 (Dense)              (None, 3)                 27
================================================================
Total params: 203
Trainable params: 203
Non-trainable params: 0
________________________________________________________________
```

Fig 4. DNN model with no weights applied.

Categorical crossentropy is used as this is a classification problem where only one output is correct out of the three values of the response variables. Softmax activation is used for the output layer while Relu(rectified linear units) is used for the hidden layers. The softmax function outputs a vector that represents the probability distributions of a list of potential outcomes.

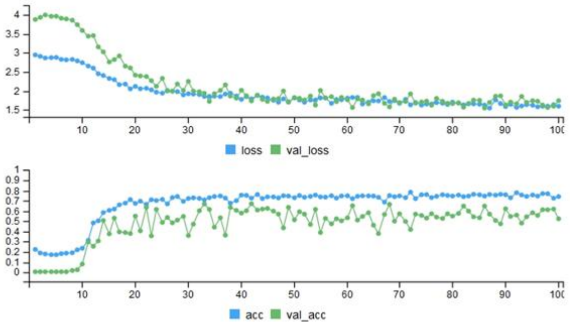

Fig 5. DNN without weights – accuracy and loss

We perform the tests on the DNN model using a simple one hidden layer network with no weights and no dropouts incrementally we add weights and drops outs to the network. The number of hidden layers are increased for the various test cases with weights.

```
Layer (type)                 Output Shape              Param #
================================================================
dense_24 (Dense)             (None, 8)                 176
________________________________________________________________
dense_25 (Dense)             (None, 3)                 27
================================================================
Total params: 203
Trainable params: 203
Non-trainable params: 0
________________________________________________________________
```

Fig 6. DNN model with weights single hidden layer

Also the accuracy and the loss can be seen in figure 5for the DNN with no weights. The weights for the various test cases and the summary of the model with the number of units in the DNN are provided in Figure 4 and Figure 4 respectively. The DNN using single hidden layer and with weights yields 82 % accuracy with a 54% loss. However the pathological case identification is much lesser than that of the SVM model.

We create DNN with dropout and 3 hidden layers with a 3 unit output layer for each of the foetal health states. The DNN provides an accuracy of 78%.

Fig 7. DNN – single layer with weights.

We use performance metrics and tabulate the finding. Although SVM has significantly higher accuracy than the deep learning model, the higher accuracy is obtained due to the identification of the negative test cases. In healthcare domain, identification of negative test cases is important but may not critical. Hence use the weighted classifier and increasing the weights for the positive cases we achieve a better performance results for the pathological cases which are critical and needs immediate medical intervention.

```
Layer (type)                 Output Shape              Param #
================================================================
dense_12 (Dense)             (None, 40)                880
________________________________________________________________
dropout_3 (Dropout)          (None, 40)                0
________________________________________________________________
dense_13 (Dense)             (None, 30)                1230
________________________________________________________________
dropout_4 (Dropout)          (None, 30)                0
________________________________________________________________
dense_14 (Dense)             (None, 20)                620
________________________________________________________________
dropout_5 (Dropout)          (None, 20)                0
________________________________________________________________
dense_15 (Dense)             (None, 3)                 63
================================================================
Total params: 2,793
Trainable params: 2,793
Non-trainable params: 0
```

Fig 8. DNN model with weights and three hidden layers

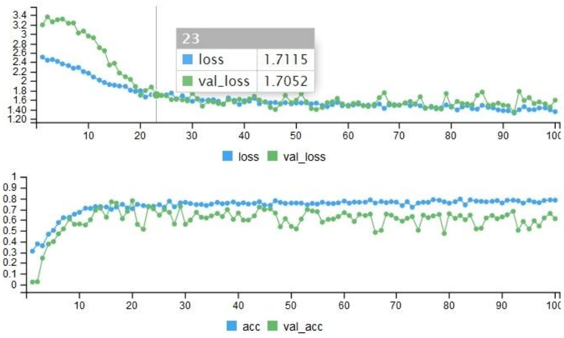

Figure 9. DNN with weights and three hidden layers

We obtain the confusion matrix for each of the test cases and derive the performance metrics as given in the section v. The tables summarizes the precision, sensitivity, specificity metrics, f1 score and gmean for the different test cases for the DNN and the SVM models. We analyze the different metrics for the individual classes instead of the macro values which are averages of the individual classes. Measurement of the class wise performance leads to better understanding of how the model performs under the different test case scenarios.

Table 2. Precision values class wise for test conditions.

Model Used	Normal	Suspect	Pathological
SVM	0.95	0.72	**0.93**
DNN – no weights	0.89	0.64	**0.91**
DNN – Single Hidden Layer weights	0.93	0.45	0.54
DNN – Two Hidden layers with weights	0.95	0.42	0.65
DNN –Three Hidden Layers with weights	0.98	0.41	0.55

It is observed that the SVM model outperforms the DNN model in terms of the precision metric. However as per the accuracy paradox the metric cannot be a conclusive predictor of the classifier performance. Table 2 provides precision values for the different test cases. Also DNN with no weights has better precision than the models with weights.

Sensitivity is by far the most important metric as it is a significant predictor of the model performance. We find that the DNN with three hidden layers has a significantly better sensitivity value for the pathological and suspect classes.

Table 3. Class wise Sensitivity for test conditions

Model Used	Normal	Suspect	Pathological
SVM	**0.97**	0.73	0.76
DNN – no weights	0.95	0.53	0.61
DNN – Single Hidden Layer weights	0.83	0.68	0.73
DNN – Two Hidden layers with weights	0.81	0.73	0.78
DNN –Three Hidden Layers with weights	0.73	**0.77**	**0.89**

Hence this will prove extremely useful in detecting the foetal health state compared to the other metrics. Table 3 gives the values of the sensitivity metric.

Table 4. Class wise Specificity for test conditions

Model Used	Normal	Suspect	Pathological
SVM	0.82	0.96	**0.99**
DNN – no weights	0.62	0.94	0.91
DNN – Single Hidden Layer weights	0.76	0.87	0.95
DNN – Two Hidden layers with weights	0.84	0.84	0.97
DNN –Three Hidden Layers with weights	0.92	0.77	0.89

SVM model performs extremely well in terms of the specificity values. Specificity relates to the test's ability to correctly reject healthy fetus which do not exhibit any abnormalities.

Table 5. Class wise F1score for test conditions

Model Used	Normal	Suspect	Pathological
SVM	0.96	0.73	**0.84**
DNN – no weights	0.92	0.58	0.73
DNN – Single Hidden Layer weights	0.88	0.54	0.62
DNN – Two Hidden layers with weights	0.88	0.53	0.71
DNN –Three Hidden Layers with weights	0.83	0.54	0.68

Hence the SVM model can be used for detecting the negative cases rather than the positive cases. DNN with two layers has significant performance over the other DNN based models. The values are provided in table 4.

F1 score is applied to seek a balance between the precision and sensitivity values. The F1 score for SVM model is significant for all the three classes as given in table 5. DNN models have lesser values which are mostly due to poor precision values as F1 score is dependent of the precision of the models.

Table 6. Class wise Gmean for test conditions

Model Used	Normal	Suspect	Pathological
SVM	0.88	0.83	0.86
DNN – no weights	0.76	0.70	0.78
DNN – Single Hidden Layer weights	0.79	0.77	0.83
DNN – Two Hidden layers with weights	0.82	0.78	0.86
DNN – Three Hidden Layers with weights	0.84	0.79	**0.91**

The Gmean measure for the DNN with three layers has significantly better measure over the other models with 91% score for the pathological state as seen in table 6. We also see that the SVM model performs evenly in all the three states.

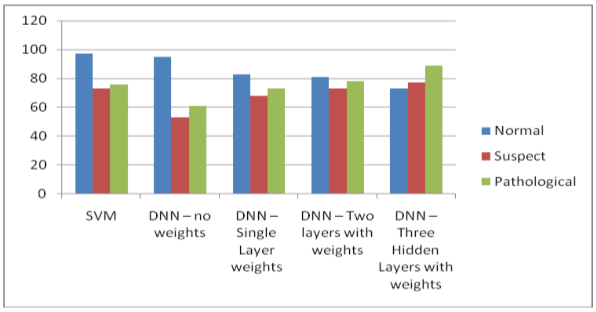

Fig 10. Sensitivity score classwise

Sensitivity score indicating the tests response for detecting the pathological states and suspect states which are abnormal are depicted in figure 10. Sensitivity is a primary indicator of the model performance.

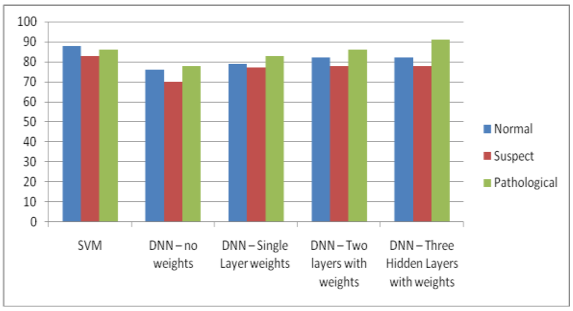

Figure 11. Gmean Score classwise

The Gmean scores are represented in the graph below in Figure 11. Deep neural network with three hidden layers outperforms the other models with Gmean of 91 % in detecting the pathological state.

VII. CONCLUSION

In healthcare detection of positive cases in life threatening situations are critical in terms of patient health and medical staff efficiency in attending to their needs. Hence we focus more on a class wise performance for the various metrics to ensure that vital classes are given their much needed attention.

We find that **Sensitivity** and **Gmean** are **key indicators** of predicting the positive test cases. SVM model performs better with respect to the precision and F1 score performance metrics with scores of 93% and 81% respectively. DNN model with three layers using weights has a significantly improved performance over the SVM in detecting the positive pathological states with 91% measure of Gmean and 89% measure of sensitivity metrics. Also the model has a significantly better performance over the suspect states for sensitivity with 77%. Hence we conclude that the weighted deep neural networks can be used to identify cases of pathological and suspect states while monitoring the foetal health for imbalanced datasets. Deep Learning neural network based tools can be developed in the future as an enhancement for detecting such critical conditions in the healthcare.

ACKNOWLEDGMENT

The author would like to express her gratitude to the Head of the department of Computer Science (PG) as well as the faculty of the department for their support in enabling this work.

REFERENCES

1. https://archive.ics.uci.edu/ml/datasets/cardiotocography.
2. Valverde-Albacete FJ, Pelaez-Moreno C (2014) 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. PLoS One 9(1):e84217
3. Vivek Bhambri, B Chopra, B Krishnan, Data Mining and Business Intelligence,2014, ch 8, pp 175-177
4. C. Sunder, Chitradevi, and G. Geetharamani, "Incapable of identifying suspicious records in CTG data using ANN based machine learning techniques", Journal of Scientific& Industrial Research, vol 73, August 2014, pp. 510-516.
5. Sindhu Ravindran, Asral Bahari Jambek, Hariharan Muthusamy, Siew-Chin Neoh, A Novel Clinical Decision Support System Using Improved Adaptive Genetic Algorithm for the Assessment of Fetal Well-Being, Computational and Mathematical Methods in Medicine, Volume 2015, Article ID 283532, 11 pages,
6. D. Bhatnagar., P. Maheshwari., "Classification of Cardiotocography Data with WEKA", International Journal of Computer Science and Network, vol 5, April 2016.
7. Z. Cömert, and A.F. Kocamaz, "Comparison of Machine Learning Techniques for Fetal Heart Rate Classification", 3rd International Conference on Computational and Experimental Science and Engineering (ICCESEN 2016), vol 132, 2017.
8. Belal Amin, A. A. Salama, Khaled Mahfouz I.M. El-Henawy, Classifying Cardiotocography Data based on Rough Neural Network, International Journal of Advanced Computer Science and Applications, Vol. 10, No. 8, 2019, pp352-356
9. Dr.S.Lovelyn Rose, Dr. L Kumar, Dr. D. Karthika Renuka, Deep Learning using Python, Wiley publication, 2019, pp130

AUTHORS PROFILE

R.Vani has around 14 years of experience including 8 years of industrial experience. Her domain areas include embedded systems, data mining and business analytics. She currently works as a faculty in the department of computer science (PG) at PSGR Krishnammal college for Women, Coimbatore.