A Reduced Feature Based Neural Network Approach to Classify the Category of Students

Mirza Mohtashim Alam
Daffodil International University
Sobhanbag, Mirpur Road, Dhanmondi, Dhaka-1207
turzo.mohtasim@gmail.com

Karishma Mohiuddin
BRAC University
Mohakhali, 66 Bir Uttam AK Khandakar Road, Dhaka 1212
natz.karishma@gmail.com

Amit Kishor Das
BRAC University
Mohakhali, 66 Bir Uttam AK Khandakar Road, Dhaka 1212
amitkishordas@gmail.com

Md. Kabirul Islam
Daffodil International University
Sobhanbag, Mirpur Road, Dhanmondi, Dhaka-1207
kislam@daffodilvarsity.edu.bd

Md. Shamsul Kaonain
BRAC University
Mohakhali, 66 Bir Uttam AK Khandakar Road, Dhaka 1212
mkaonain@bracu.ac.bd

Md. Haider Ali
BRAC University
Mohakhali, 66 Bir Uttam AK Khandakar Road, Dhaka 1212
haider@bracu.ac.bd

ABSTRACT

To ensure more effectiveness in the learning process in educational institutions, categorization of students is a very interesting method to enhance student's learning capabilities by identifying the factors that affect their performance and use their categories to design targeted inventions for improving their quality. Many research works have been conducted on student performances, to improve their grades and to stop them from dropping out from school by using a data driven approach [1] [2]. In this paper, we have proposed a new model to categorize students into 3 categories to determine their learning capabilities and to help them to improve their studying techniques. We have chosen the state of the art of machine learning approach to classify student's nature of study by selecting prominent features of their activity in their academic field. We have chosen a data driven approach where key factors that determines the base of student and classify them into high, medium and low ranks. This process generates a system where we can clearly identify the crucial factors for which they are categorized. Manual construction of student labels is a difficult approach. Therefore, we have come up with a student categorization model on the basis of selected features which are determined by the preprocessing of Dataset and implementation of Random Forest Importance; Chi2 algorithm; and Artificial Neural Network algorithm. For the research we have used Python's Machine Learning libraries: Scikit-Learn [3]. For Deep Learning paradigm we have used Tensor-Flow, Keras. For data processing Pandas library and Matplotlib and Pyplot has been used for graph visualization purpose.

CCS Concepts

• Computing methodologies→ Machinelearning→ Machine learning approaches • Computing methodologies→ Machine learning→ Learning paradigms

ICIAI 2018, March 9–12, 2018, Shanghai, China

ACM ISBN 978-1-4503-6345-7/18/03…$15.00
DOI: https://doi.org/10.1145/3194206.3194218

Keywords

Artificial Neural Network, Random Forest Importance, Chi2, Classification, Predictive Model, Student Category.

1. INTRODUCTION

Student categorization is a system to find out the standard of student according to their performance. It is necessary to help them to find their weakness and to guide them to the recover those weaknesses by understanding their specific performance. There are 43 variables in total from which only 3 top variables have been selected with the art of machine learning approach. Features like raising hands, topic, announcement visits, discussion, parents satisfactory, visit resources, absence of students etc. are included in the large dataset with 43 variables. In this case, we have applied the machine learning techniques to determine key features of their performance which identifies a student's category. For this purpose, we have chosen a suitable dataset [4] from Kaggle and, it has been used by Random Forest Importance (RFI) and Chi2 (Chi-Square) algorithm simultaneously to select best features from both implementations. Related authors of this dataset have also done researches on students' academic performance [5][6]. After that, data has been preprocessed to create Test and Training sets for applying the ANN algorithm. Eventually it results in a model that reflects the main reason for the categorization of students. Raising hand, announcement visit, resource visit are the top variables that determine the behavioral and enthusiastic values. Those students who are more active in class by raising hand to ask questions or to show enthusiasm, those who are more attentive to announcements and are frequently accessing study materials are marked as high ranked students. Again, those who do such activity a little less are identified as medium students and lastly, those students who hardly do such activities are acknowledged as low category students. Also, there are some other factors that include demographic variables like race and cultures etc. Students' academic attendance and parental issues related to the academic factors can also be viable for building a predictive model. If a low graded student tries to pursue better result then, that student has to go through these viable features to achieve good result in the exams.

2. RELATED WORKS

The authors in their paper used machine learning approach to predict the dropout students or the students who might switch their schools [1]. Also, they have ranked student on the possibility

of dropout at high school graduation level. For this reason, they adapted the ML techniques such as RFS, Logistic regression. Another paper [7], the authors have developed a system of selecting the suitable subject for the student of postgraduate degree with machine learning approach, in this case, their aim is to detect best subject for a student for post-graduation degree on the basis of previous educational record so that authors could reduce the possibility of dropout of student and aid them successfully to complete post-graduation degree by suggesting them suitable subject. In paper [8], the authors initially developed a model to predict student performance with the state of the art of machine learning like Bayesian classification and Neural Network that give the accuracy 0f 70.48%. Further, in another research conducted by the authors, where they have come with a personal module for distant learner [9]. Additionally, their goal is to create automated module for e-learner, hence, machine learning technique has been applied on the system model to inspect the performance of distant learner. However, our approach is completely different from others as we have developed a new model to determine student's performance by ranking them into high, medium and low category. Without any doubt a high ranked student would perform better than medium or low ranked students. The categorization has been created on the basis of top attributes which is conducted by machine learning approach such as RFS, Chi2, ANN and the system results in 85% accuracy. By ranking the student level, it is possible to identify student's weakness specifically so that the teacher can easily guide them to overcome their flaws.

3. SYSTEM MODEL

Our System model consists of several components. Firstly, we have encoded the categorical variables in order to deal with them. Secondly, we have applied Random Forest Feature importance to obtain the weights of the features. Additionally, Chi2 technique also has been applied to draw the weights of the features. Each of the weights and the features are then sorted along with the feature names. For plotting purpose, we only have taken top three common features according to their weights for further processing. However, we have not applied any feature scaling techniques because the chosen variables are in suitable range. Finally, the whole training set has been trained with multilayer Artificial Neural Network. Our system model has been provided in Figure 1.

3.1 Data Preprocessing

In the data preprocessing phase, we have handled the categorical data by encoding them. Both the features and labels contained several data columns which were categorical. Individual features such as *gender*, *nationality*, *parent answering survey*, *parent school satisfaction*, *student absence days* are categorical data. The outcome or label is also considered categorical data because it contained three categories: High, Low and Medium. Each of the categories is encoded. For this purpose, we have taken each of the categorical columns and made them encoded and created binary dummy columns for each. To achieve such purpose Panda's get_dummies[1] function has been used. Let's consider a categorical variable column as *cvcol*, column headers as *cvcolheaders*. We can find the dummy columns using the following equation.

$$cvcol_i = getdummies(cvcol_i.values) \quad (1)$$

[1]http://pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.get_dummies.html

3.2 Random Forest Importance

With the average of all the positive result of decision trees, the random forest can be calculated for determining the feature importance [10] [11]. For gathering all the values of feature in one, Scikit-Learn Random Forest library has been utilized [11]. We have used the Random Forest Feature importance for accessing the weights of each feature so that we can work with the highest weighted features. To get the Random Forest importance we have created a Random Forest Classifier[2] with the entropy criterion and

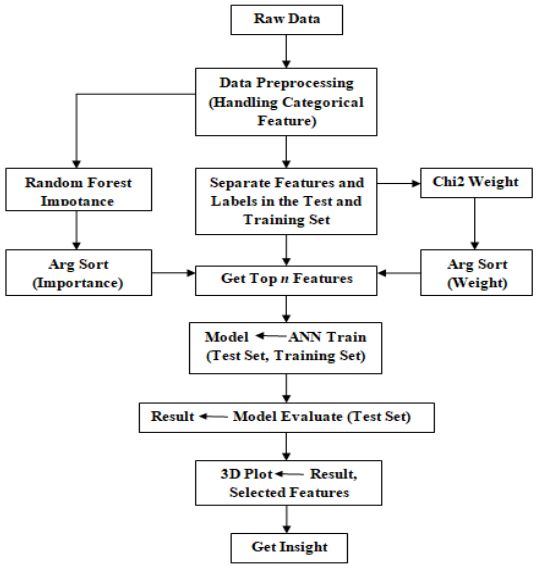

Figure 1. System Model

estimator number 1000. Here, estimator indicates the number of trees in the forest. We have fitted the feature and training set to the Random Forest Classifier. Finally, we have accessed the feature importance with the classifier's *feature_importances_* properties. We have sorted the importance values and obtained the importance along with the feature names.

3.3 Chi2 Algorithm

Chi2 algorithm is mainly used to determine the highest valued features from the test Chi2 statistic which is consisting of training and test set [12]. To achieve such purpose, we have separated the features into training set and test set. The training set of our, is consisted of non-negative features that are in Boolean form or in frequencies because Chi2 algorithm's performs on these features [12].

Chi2 algorithm mainly develops on the statistic of X^2 which executes in two stages. In first stage, for each pair of epochs it evaluates the value of X^2 from the significant level and then by combining all the values of intervals with lowest value of X^2 the procedure runs till the values of all X^2 have exceeded the sigLevel which is determined parameters. The continuation of stage one goes on until the rate of inconsistency has passed beyond the detached data. However, stage two is more efficient and far better version of stage one because it starts at sigLevel0 where each level is connected sigLevel[s]. Further, checking consistency and adding up the features are also part of the procedure. Until all the

[2]http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

attributes are merged up, the process continues and checks if the inconsistency label has passed or not. If it has passed the level, then the sigLevel[s] has been decreased for every feature's (s) subsequent stage of merging. The mathematical equation of Chi2 algorithm has given below [13].

$$X^2 = \sum_{S=1}^{2} \sum_{t=1}^{n} \frac{(Ust - Vst)}{Vst} \quad (2)$$

Here, n= number of classes,

U_{st} = number of patterns in s^{th} epoch

V_{st}= expected frequency of U_{st}

3.4 Feature Selection and splitting the dataset

We have compared the values of the result from the Chi2 and Random Forest importance and taken the top three variables: *visited resources*, *raised hands* and *announcements view* as these contained the highest weights accordingly. Taking these three variables, we have created the final features of the dataset. The labels contained the encoded dummy variables of the student's categories as before. Then, the training and test set has been separated and the test set contained 20% of the whole dataset. No feature scaling has been done as the training set was in suitable range. Finally, the training and test set has been ready to be fitted into an Artificial Neural Network (ANN).

3.5 Artificial Neural Network

Artificial Neural Network has been implemented to provide a prediction model where each of the student labels (*High*, *Mediumand Low*) can be identified based on the top features (*VisITedResources*, *raisedhands* and *AnnouncementsView*). Artificial Neural Network has been inspired from biological background to that can be utilized to perform certain tasks such as classification, pattern recognition, clustering etc. [14]. ANN is consisted of three different layers namely Input layer, Hidden layer and Output layer. Since we have three reduced feature sets, the input layer's value is three. We had several hidden layers because we wanted to have consistency in the accuracy of our model. Additionally, the multi hidden layer approach has been used to gain a better accuracy while training the model, though the time complexity increases after adding an additional hidden layer. We have a total of five Hidden layers each of which contains a value of the multiplication factor of the input layer, where the value has been designated as, *input layer*2k* where k is equal to a set of values: 1, 2, 4, 2, 1 periodically. Let us consider the input layer be defined as *inp*, variable containing set of sequential values as *k* and a particular hidden layer as *h*. The equation can be deduced as,

$$h \leftarrow inp \times 2k \; where \; k = \{1,2,4,2,1\} \quad (3)$$

There are several activation functions[3] to use for activating each of the layer components: *Rectifier Linear Units*, *Sigmoid Units*, *Tanh Units* etc. since, our labels or outcomes are in categorical form we have used *softmax* function in the output layer. The optimizer is *adam* [15], The hidden layer activation function is *Rectifier Linear Units (ReLU)* [16] and loss function is given as *categorical crossentropy*. Since, training of a supervised deep neural network fastens in case of using ReLU as hidden layer activation function [16]. While training we have used 50000 epochs so that our training accuracy goes to a consistent level as we are working with reduced number of features. We have taken the batch size as 48. Our whole approach has been described in Algorithm 1 and Algorithm 2 accordingly.

Algorithm 1: Student's category classification(dataset preperation AND Defined Functions)

```
j:=0;
i:=0;
do
    dataset.featureColumn[i]:=dataset.featureColumn[j].getDummies(columnName);
    i:=i+1;
    j:=j+1;
while (j <= (dataset.featureColumn.length − 1));
Def featureImportanceWRforest(X,Y):
    featureLabels:=Dataset.ColumnNames[all];
    classifier:=RandomForestClassifier(criterion='entropy',estimators=1000,rstate=TRUE)
    classifier.fit(X,Y);
    importances:=classifier.featureImportances;
    indices:=sort(importances);
    f:=0;
    do
        display featureLabels[indices[f]] AND importances[indices[f]];
        f:=f+1;
    while (f <= featureColumn.length − 1);
    return all features(sorted) according to importance in X,Y;
Def chi2Calculation(Xtrain,Ytrain,Xtest,Ytest,kNumber):
    selector:=SelectorKBest(chi2,k=kNumber);
    Xtrain:=selector.transform(Xtrain);
    Xtest:=selector.transform(Xtest);
    scores=selector.scores;
    return Xtrain,Xtest,scores;
Def createANNModel(inputLayerDimension,hiddenLayerAct,outputLayerAct):
    model:=Sequential();
    model.add(inputLayerDimension);
    model.add(inputLayerDimension*2*1, activation:=hiddenLayerAct);
    model.add(inputLayerDimension*2*2, activation:=hiddenLayerAct);
    model.add(inputLayerDimension*2*4, activation:=hiddenLayerAct);
    model.add(inputLayerDimension*2*2, activation:=hiddenLayerAct);
    model.add(inputLayerDimension*2*1, activation:=hiddenLayerAct);
    model.add(outputLayerDimension, activation:=outputLayerAct);
    model.compile(optimizer:=adam,loss:=categoricalCrossentropy);
    return model;
```

Algorithm 2: Student's category classification((Algorithm Analysis)

```
RFImportanceFeautreVector:=Get feature importance and feature
  column names by calling featureImportanceWRforest(*paramlist);
Xtrain,Xtest,Ytrain,Ytest:=TrainTestSplit(dataset.features,dataset.labels,tSize=0.20)
Chi2WeightVector:=Get Chi2 Weights and feature
  column names by calling chi2Calculation(*paramlist);
RFImportanceFeautreVector(Sorted):=Argsort(RFImportanceFeautreVector);
Chi2WeightVectorVector(Sorted):=Argsort(Chi2WeightVectorVector);
CommonBestK(RFImportanceFeautreVector(Sorted),Chi2WeightVectorVector(Sorted

Subsequently Xtrain,Xtest will contain no more than three features;
inputLayerDimension:=3;
hiddenLayerAct:=relu;
outputLayerAct:=softmax;
createdModel=createANNModel(inputLayerDimension,hiddenLayerAct,outputLayerA
createdModel.fit(Xtrain,Ytrain,epochs:=50000,batchSize:=48);
Ypred:=classifier.predict(Xtest);
Ypred:=(Ypred>=0.5);
confusionMatrix:=confusionMatrix(Ytest,Ypred);
df:=create new data frame columns using Xtest AND Ypred column wise;
define X Y and Z axis column set from df;
create colormap as col for each of the labels or outputs;
plot population with colormap for each of the population;
visualize3D(X,Y,Z) using different colors to identify the categorical regions;
```

3.6 Visualization using 3D Graph

After getting the predicted result of each of the categories of the students (*High*, *Medium* and *Low*), we have merged the result with the test set features we have just evaluated. Finally, we have created a new dataset with all the test set features and predicted result and visualized them in different color-maps.

4. RESULT AND ANALYSIS

[3]https://isaacchanghau.github.io/2017/05/22/Activation-Functions-in-Artificial-Neural-Networks/

Firstly, we have applied Random Forest importance with our preprocessed variables. Afterwards, Chi2 have been applied. The main reason was to find out the top *n* important variables to be selected for the training set. The result of Random Forest Importance and Chi2 on 43 variables has been provided in Table 1. With the common 3 variables we have gone through a multilayered Neural Network based approach. The results obtained for the predicted values are visualized in a 3D graph for better identification of each of the output categories which is given in Figure 2.

Table 1: Result of Random Forest Importance and Chi2 weights

Variables	Random Forest Importance	Chi2 Weights
VisITedResources	0.147091	3606.48
raisedhands	0.137987	3423.82
AnnouncementsView	0.119978	2000.07
Discussion	0.088378	558.07
StudentAbsenceDays_Above-7	0.080157	106.85
StudentAbsenceDays_Under-7	0.075749	71.54
ParentAnsweringSurvey_Yes	0.027948	33.05
ParentAnsweringSurvey_No	0.026729	38.25
NationalITy_KW	0.020811	14.76
ParentschoolSatisfaction_Bad	0.017978	30.00
gender_F	0.017867	19.70
NationalITy_Jordan	0.017307	2.15
gender_M	0.017303	12.77
ParentschoolSatisfaction_Good	0.017058	21.43
StageID_lowerlevel	0.015898	3.60
StageID_MiddleSchool	0.015352	1.94
Topic_IT	0.013876	10.07
Semester_F	0.012243	2.71
Semester_S	0.011646	2.97
Topic_Arabic	0.010389	0.81
Topic_French	0.009333	0.66
Topic_English	0.009075	2.10
Topic_Science	0.007604	0.45
NationalITy_Iraq	0.006758	10.94
Topic_Chemistry	0.006713	4.12
Topic_History	0.006698	1.16
Topic_Math	0.006667	0.57
StageID_HighSchool	0.006621	1.05
NationalITy_SaudiArabia	0.006272	2.49
Topic_Geology	0.005504	7.55
Topic_Spanish	0.005327	0.55
NationalITy_Palestine	0.005319	7.08
Topic_Biology	0.004897	3.34
NationalITy_lebanon	0.004840	6.64
Topic_Quran	0.004130	1.60
NationalITy_Tunis	0.003102	0.06
NationalITy_Egypt	0.002090	1.11
NationalITy_USA	0.001615	2.91
NationalITy_Iran	0.001599	2.12
NationalITy_Syria	0.001450	0.58
NationalITy_Morocco	0.001202	0.12
NationalITy_Lybia	0.001125	13.46
NationalITy_venzuela	0.000313	2.37

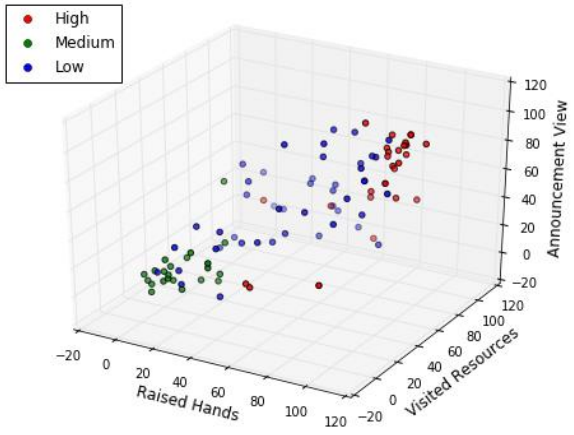

Figure 2. 3D Visualization of predicted student categories

From Figure 2 we can clearly view that the low category students (indicated by green color) are numerically low in each of the three features. The medium category students are the largely scattered in the graph space. Even small portion of them are overlapping with the high categories (indicated by red color) and low categories (indicated by green color). Medium students are stretched in the graph space. The high categories are good at each of the three variables. By visualizing this dataset, our observation has been given in Table 2.

Table 2: Dense area distribution based on the categories

		Raised Hands	Visited Resources	Announcement View
High Class	High density area	80-100	80-100	60-100
	Mid density area	null	null	35-65
	Low density area	20-60	0-20	10-20
Mid Class	High density area	10-38	60-100	0-40
	Mid density area	40-80	20-60	40-95
	Low density area	80-96	0-10	null
Low	High density	0-15	0-12	0-36

Class	area			
	Mid Density area	20-40	20-40	null
	Low density area	null	60-62	56-58

Since, for high category students there is no mid density areas for raised hand or visited resources and low frequencies in mid low density areas, it means high category students are more likely to be attentive in the class and concerned about the resources provided by the teachers. Middle class students lack in raising hands in the class yet there are some exceptions but very little in number. They are not very serious about announcement views. The low category students fail in every aspect because of having very low frequencies in high density areas. We also have compared several algorithm results with our model's Neural Network which has been built on Keras [17]. This comparison has been provided in Table 3.

Table 3: Comparison of different algorithms

Algorithm	Accuracy with all variable	Reduced variable(3)
ANN	95%	85%
Decision tree	68%	60%
Random forest	73%	60%
MLP classifier	59%	20%
SVM	67%	69%
Gaussian	61%	72%
KNN	63%	62%

5. CONCLUSION

Our model has been defined based on numerical intervals of the grades of the students to classify them into suitable numbers of attributes. Our procedure has applied multi-layered neural network based approach which out performs above mentioned state of the art machine learning algorithms. Without the reduction of the variables, the model has shown up to 95% accuracy. For plotting purpose and getting a better insight from the prediction model, we have reduced the features in such a form which has been easily applied on the 3D plotted graph. After reducing the features, we have achieved accuracy up to 85%. Here, in our case, the preprocessed attributes have been decremented from 43 to 3 and these three have high amount of positive co-relation. Further, in future, we will try to figure out in which extent the variables shall be responsible for the betterment of the student's academic performance.

6. REFERENCES

[1] Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., & Addison, K. L. (2015, March). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 93-102).ACM.

[2] Xu, J., Moon, K. H., & van der Schaar, M. (2017). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal of Selected Topics in Signal Processing*.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.

[4] I. (2016, November 26). Retrieved November 14, 2017, from https://www.kaggle.com/aljarah/xAPI-Edu-Data

[5] Amrieh, E. A., Hamtini, T., &Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

[6] Amrieh, E. A., Hamtini, T., &Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

[7] A., USIOBAIFO, & 2. O., OSASERI. (2016). A MACHINE LEARNING APPROACH FOR PREDICTING POSTGRADUATE STUDENTS' PERFORMANCE. *Proceedings of INCEDI 2016 Conference*, 779-784. Retrieved November 15, 2017, from http://www.incedi.org/wp-content/uploads/2016/11/A-MACHINE-LEARNING-APPROACH-FOR-PREDICTING-POSTGRADUATE-STUDENTS%E2%80%99-PERFORMANCE-USIOB.pdf.

[8] Agrawal, H., &Mavani, H. (2015). In Student Performance Prediction using Machine Learning. *International Journal of Engineering Research and Technology*.

[9] Li, N., Cohen, W., Koedinger, K. R., & Matsuda, N. (2010, June). A machine learning approach for automatic student model discovery. In *Educational Data Mining 2011*.

[10] RASCHKA, S. M. (2017). PYTHON MACHINE LEARNING -.S.l.: PACKT PUBLISHING LIMITED.

[11] Random forest feature importance. (n.d.). Retrieved October 01, 2017, from http://blog.datadive.net/selecting-good-features-part-iii- random-forests/.

[12] Sklearn.feature_selection.chi2.(n.d.).Retrieved October 01, 2017,fromhttp://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.chi2.html.

[13] Liu, H., &Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Tools with artificial intelligence, 1995. proceedings. seventh international conference on(pp. 388-391). IEEE.

[14] (2017, July 17). Overview of Artificial Neural Networks and its Applications. Retrieved August 28, 2017, from https://hackernoon.com/overview-of-artificial-neural-networks-and-its-applications-2525c1addff7.

[15] Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[16] Glorot, X., Bordes, A., &Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on ArtificialIntelligence and Statistics* (pp. 315-323).

[17] Chollet, F. (2015). Keras.