Indian Journal of Agricultural Sciences **91** (6): 861–6, June 2021/Article
https://doi.org/10.56093/ijas.v91i6.114287

Comparative analysis of machine learning based classification for abiotic stress proteins

BULBUL AHMED[1], ANIL RAI[1], MIR ASIF IQUEBAL[1] and SARIKA JAISWAL[1]*

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

Received: 09 October 2020; Accepted: 04 January 2021

ABSTRACT

For thousands of years, cereals which include rice, wheat, maize, sorghum and millets etc. have been playing major role in human civilization. These are the principal components of human diet and important staples for daily survival of billions of people globally. The cereal crops belong to poaceae family and rich in vitamins, minerals and fiber. They are reported to reduce the coronary heart disease and other serious diseases. These crops are adversely affected by biotic and abiotic stresses like cold, drought, heat and salinity. With the advent of modern NGS technologies, the plethora of molecular data leads to infer many unexplored facts of the cereal crops using *in-silico* approach. In the present work, computational techniques were applied to study thoroughly the classification of abiotic stresses (cold, drought, heat and salinity) responsive genes in cereals. The datasets of four stress responsive genes in poaceae family was retrieved from public domain. The machine learning based methodologies namely, Random forest, Support Vector Machines and Deep Learning-Long Short-Term Memory (DL-LSTM) were applied. A comparative analysis was carried out for classification of the retrieved data with *k*-fold cross validation applying the machine learning techniques at different parameters. It was observed that for all the four sets of data, accuracy was maximum, i.e. 95.11%, 76.88%, 94.31% and 82.04% for cold, drought, heat and salinity, respectively using DL-LSTM. Comparison of the methodologies obviates the outperformance of deep leaning. Such approach of computational studies will help researchers to study the complex biological problems of gene classification more efficiently.

Keywords: Classification, Deep learning, LSTM, Poaceae, Random forest, SVM

Cereal crops belong to the monocot grass family (Poaceae/Gramineae), having wheat, rice, millets, maize, sorghum, barley etc. and are the staple food crops. They are rich source of minerals, vitamins, protein, carbohydrates, fats and oil (Sarwar *et al.* 2013).The global cereal production and use is estimated to expand by 13% and 14%, respectively frombase year 2017 to 2027. Cereals are staple food crops grown globally, containing essential nutrients such as vitamins, minerals and fiber. These are reported to reduce risk factor of developing coronary heart disease, diabetes and colon cancer (Sarwar *et al.* 2013). India ranks third (~313 MT in 2017) in cereal production after China and United States (https://www.indexmundi.com/). Globally, cereal crops cover the lion's share in production among all the crops. Since 1960, the food production in India has been satisfactory to handle food scarcity. The application of agri-biotechnology enhanced by genetics has been playing a major role in reforming agriculture.

The cereal crops are adversely affected by biotic (diseases, insect and pests etc.) and abiotic (heat, cold, water and salinity etc.) stresses leading to major economic losses. The cereals occupied 51.4 Mha land area in 2010 but due to drought stress, *kharif* crops production reduced by 10% and 15 Mha land area was affected for all the agricultural produce. Harsh agro-ecologies and soil conditions threaten the agricultural production because a set of genes play important role in stress tolerant mechanisms (Bal *et al.* 2014). It is tough to understand the response of crops under such stresses due to multigenicity. The discovery of abiotic stress resistance genes can assist breeders in endeavor of higher production. But this discovery is very expensive, time consuming and limited due to complexities of genotype interaction between genes and their associated environment (G×E) interaction. The *in-silico* way to decipher the stress resistance genes by computational/ bioinformatics approaches are widely used these days. In the present study, we applied deep learning and other machine learning techniques like Support Vector Machine (SVM) and Artificial Neural Network (ANN) over the heat, cold, water and salinity stresses in cereals.

MATERIALS AND METHODS

Data collection and pre-processing: The data was retrieved from the public domain from Uniprot database

Present address: [1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi. *Corresponding author e-mail: sarika@icar.gov.in.

(https://www.uniprot.org/). The search was performed using the keywords "salt stress", "drought stress", "heat stress" and "cold stress" in poaceae family using appropriate Boolean operators like AND, OR and NOT till January 2020. The poaceae family included data from crops like rice, wheat, sorghum, barley, *Lolium arundinaceum*, maize etc. Consecutively, the negative data was retrieved from Uniprot and NCBI in the form of genes which were not associated with cold stress, drought stress, heat stress and salinity stress from poaceae family (Roh *et al*. 2019).

Machine learning and deep learning are the widely used techniques for classification problems. Deep learning can capture more complicated patterns of inputs (Deng 2011) rather than manual features selection (Arel *et al*. 2010). Deep learning approaches can establish great impact and has vital role in providing predictive analytics solutions for large-scale data sets in future work (Qiu *et al*. 2016). In order to carry out the leaning analysis, the retrieved data was processed as follows:

Imbalance data: One class of the data may have smaller number of samples compared to the others. In such case, learning may ignore the classes with smaller number of samples or gives only one or two or more number of samples. To overcome this, scaling and standardization of data is required.

Scaling: The features with larger range will dominate during training. So, scaling can achieve consistency in a varying range of data. Generally scaling is use to convert the data ranging between 0 to 1 which can improve and speedup the model learning.

Standardization: All the variables are taken as zero mean with unit variance so that no variable should dominate over the variance with larger variance $(Z = \frac{x-\mu}{\sigma}; \bar{\mu} = 0; \bar{\sigma}^2 =)$ where, Z is standard normalization, x the variables, μ is mean and σ^2 is variance) (Tauber and Sánchez 2002).

All the stress tolerance data have been taken separately with almost equal number of negative data. Numerical representation was taken in matrix form in CSV format with tab delimited. Only the significant features with correlation <80% were considered for further analyses.

The classification data observed are based on categorization of features, i.e. model construction to train the model and test the model based on new input data. Machine learning techniques such as Support Vector Machines (SVM), Random Forest (RF) and deep learning algorithms were employed on the described data.

Support Vector Machines (SVM): It is a supervised machine learning algorithm developed by Vapnik where data is divided into optical hyper-plane (Cortes and Vapnik 1995). Kernel functions are used to measure the similarity between single as well as multi class input data. The different kernel functions of SVM are linear, polynomial, Radial Basis function and Sigmoid (Vapnik 1995). Accuracy of model depends on kernel function and tuning of parameters of the model.

Random Forest (RF): RF is an ensemble method which uses divide and conquer to improve the weak learners into strong learner which can use in regression as well as classification problems (Biau 2012). Random forest construct group of decision trees and each tree is used to train the data. It works better with large number of data having large number of features for finding the variance of each trees. RF basically reduce the over-fitting the data. Large data used can reduce the variance with higher accuracy and flexibility even with large amount of missing data. However, complexity of forest makes learning and testing tough and time consuming for (Breiman 2001).

Deep learning (DL): Deep learning network is a technique that mimics the working of human brain by processing and creating patterns for the data in decision building. It can work with unstructured data contacting large number of features that can be extract automatically without any human intervention (Le Cun *et al*. 2015). A DL model works on the following three steps:

Input data: Input data variables are fed to the model using some weights where these weights are learnt during training.

Model learning: Some function is performed using the input data so that the model can learn the features.

Class prediction: The features learnt during training are used to classify any unknown given input dataset during testing. y is output of the given function f.

It requires large number of information to train the data, necessitating the voluminous data for accurate finding.Non-linear function and sigmoid functions are used to connect to all the layers including input and output layers (Wen *et al.* 2018). The most widely used deep learning in sequence data is recurrent neural network (RNN). RNN is useful dynamic problems but problem may arise during back propagation as it may show growth or shrink at each step. All the layers are given same weight in feed forward network which are good for long term dependencies but difficult to learn in theoretical and empirical learning (Vieira *et al*. 2017). Long Short-Term Memory (LSTM) is a special recurrent neural network (RNN) architecture used for sequences more accurately. Effective use of hyper-parameters (Bergstra *et al*. 2011) outperforms and can converge quickly as compared to rest of the deep learning algorithms (Sak *et al*. 2014).

Architecture of proposed Recurrent Neural Network of Deep Learning model: To convert the dataset into numeric vector, 48 features associated were amino acid compositions of 20 amino acids, Coiled-coil domain, Polo-like kinase, Prediction of Pupylation, A-*Nitrosylation*, B-*Nitrosylation,* C-*Nitrosylation,* Total N*itrosylation*, A-*Nitrotyrosine*, B-*Nitrotyrosine*, C-*Nitrotyrosine*, Total-*Nitrotyrosine*, SUMOylation I, SUMOylation II, SUMOylation III, Total SUMOylation, amino acid number, number of negative amino acids, number of positive amino acids, molecular weight, theoretical pI, number of carbon atoms, number of hydrogen atoms, number of nitrogen atoms, number of oxygen atoms, number of sulphur atoms, instability index, aliphatic index and grand average hydropathy (GRAVY). The encoded numeric representation is passed into an input

layer with 150 nodes and *He* normal as weight initializer in the Keras framework. The syntax followed dense layer with 500 units and 1D recurrent layer for the binary output to check if the input is stress tolerant or not. It helps to create further compact representation of input symbols yielding similar symbols close to each other in the vector space. This embedding layer can be taken for training with other layers of RNN updating the weights during training.This was done for all the four classes of stress, i.e. Cold (C), Drought (D), Heat (H) and Salinity (S).

After convolution layer, we applied max pooling layer with 500 nodes to down samples the filter values by sliding windows of length 50 nodes which were non-overlapping with 0.2 drop out. The other parameters were kept as default to overcome the issue of over fitting. This also helps to improve the RNN's predictive performance. The output from this LSTM layer passed through the final dense layer, which made use of sigmoid activation function giving the final prediction in the range of [0, 1]. The Keras model was compiled with 500 epochs, adam optimizer, loss: mean squared error and metrics: accuracy. The 5-fold cross validation was applied for validating the developed model.

Model tuning and construction: We considered the two separate models, namely, a 'training' model for evaluating the testing partition and a 'validation' model which is trained on all available data. Model parameters were tuned through different combination of hyper parameter and based on the most suitable outcome, hyper-parameter have been selected for all the four models. The prediction probability threshold of >0.5 was used to denote the gene as "specific stress"while those <0.5 were labeled as "non-specific stress".

Model evaluation: The classification performance of the stress associated genes was evaluated in terms of sensitivity, specificity, accuracy and Matthews Correlation Coefficient, which are the functions based on the number of true positive (TP), true negative (TN), false-positive (FP) and false-negative. These are defined as follows:

$$\text{Specificity} = \left(\text{TNR} = \frac{\text{TN}}{\text{TN}+\text{FP}}\right),$$

$$\text{Accuracy} = \left(\text{SE} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}\right),$$

$$\text{Precision} = \left(\text{PE} = \frac{\text{TP}}{\text{TP}+\text{FP}}\right),$$

$$\text{Recall} = \left(\text{RE} = \frac{\text{TP}}{\text{TP}+\text{FN}}\right),$$

$$F1 = \frac{2 \times \text{Precesion} \times \text{Recall}}{\text{Precesion}+\text{Recall}},$$

Matthews correlation coefficient

$$\cdot\,(\text{MCC}) = \left(\frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}\right),$$

These are not totally independent of each other. However, they have different performance on ML and deep learning algorithms and treat independently (Zhang *et al*. 2020).

RESULTS AND DISCUSSION

A total of 739 cold stress, 642 drought stress, 977 heat stress and 473 salt stressproteins were retrieved from public domain data search. Consecutively, 1305, 1284, 1305 and 946 negative dataset that are not associated with any stress condition were considered for cold, drought, heat and salt stresses, respectively. After removing variables with 80% or above similarity, we reached to 36 variables, namely CCD, POLO, PUP, Nitro A, Nitro B, Nitro C, YNO A, YNO B, YNO C, Sumo I, SUMO II, Sumo III, Amino acid number, Theoretical pI, Instability Index, Aliphatic index and 20 Amino acid compositions have been selected and standardized the data so that they follow normal distribution and scale between 0 to 1.The total data was split into 80% for training and 20% for testing the model. In all four datasets, we applied machine learning techniques like Support Vector Machine (SVM), Random Forest (RF) and deep learning technique (Long Short-Term Memory (LSTM)) for classification and comparison of the result. The hyper-parameters were fine tuned to get maximum accuracy using SVM, RF and RNN-LSTM methodologies.

For the four datasets under study, it was found that SVM model with polynomial kernel with default parameters were the best fit. The confusion matrix for calculation of further evaluation measures as reported in Table 1. For the same set of four datasets, random forest (RF) technique was also applied. This methodology constructed 101 trees with leaf weight of 0.05 at default parameter. The confusion matrix constructed for all the four datasets are represented in Table 1. Finally, we made an attempt to apply deep learning technique to solve the classification problem. We applied DL-LSTM in our four datasets with input layer of 150 nodes, hidden layer with 500 nodes and padding layer with 50 nodes. The Gaussian error linear units (*gelu*) activation function (Hendrycks and Gimpel 2016), *he* normal weight initialization (Young-Man *et al*. 2019), *adam* optimizer

Table 1 Confusion matrix for classification of different stress proteins using SVM, RF and DL-LSTM classifier

	SVM				RF				DL-LSTM			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
Cold	134	250	5	20	107	251	4	47	141	248	7	13
Drought	58	231	12	84	21	239	4	121	95	201	42	47
Heat	176	258	2	21	145	257	2	52	167	257	3	30
Salt	44	183	6	51	44	185	4	51	63	170	19	32

and mean square error loss function and sigmoid output performed better. Besides, in certain cases Rectified Linear Unit (*ReLU*) activation (Eckle, and Schmidt-Hieber 2019) with *he* normal weight initialization, *adam* optimizer and mean square error loss function (Reddy *et al*. 2018) and sigmoid output also performed good. In case of cold stress, as the data are less deviated, the training and validation, accuracy curves keeps increasing without intersecting with each other. It gives a constant and stable model with better learning capacity and maximum accuracy. However, in case of drought and heat, there are variation in datasets hence, training and validation curve either keep changing or overlap with each other. This provides a less stable model and results in less accuracy in all three models. The confusion matrix of DL-LSTM model underall the four datasets are represented in Table 1.

It is observed from the results that the accuracy of deep learning based methodology is maximum as compared to SVM and RF methodologies. Also, the F1 score, sensitivity and MCC highly fluctuate with little variation of accuracy in SVM and RF, whereas DL-LSTM shows least effect of all these. It was also found that DL-LSTM takes more time to train as compared to SVM and RF which may be due to the multiple layers. The results have been compared on the basis of accuracy and MCC concluding that as the data complexity increases, LSTM shows better performance. The evaluation measures (Specificity, Sensitivity, Accuracy, Precision, Recall, F1 Score and MCC) are calculated for these four datasets. Table 2 delineates the comparison of these methodologies for cold, drought, heat and salt, respectively. Graphical representation of training and validation curve of DL-LSTM model at 500 epochs for the four stress datasets is represented in Fig 1.

Also, these developed models have been cross validated with standard data of heart disease which is available in UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/heart+Disease). It consists of 303 samples with 139 are positive and rest 164 are negative, i.e. they do not have the disease. This dataset had 76 features. BayesNet, SVM and Function Tress (FT) were used, out of which SVM outperformed as compared to remaining with an accuracy of 83.8% (Otoom *et al.* 2015). We used this dataset for cross-validation of our developed models using SVM, RF and LSTM. Surprisingly our results show an accuracy of 84.21%, 86.84% and 92.98% for SVM, RF and LSTM respectively. LSTM was found to have maximum accuracy as compared to SVM and RF.

With the advancement of recent NGS technologies, the voluminous data generation can help to make inferences of many biological events. It also plays an important role for

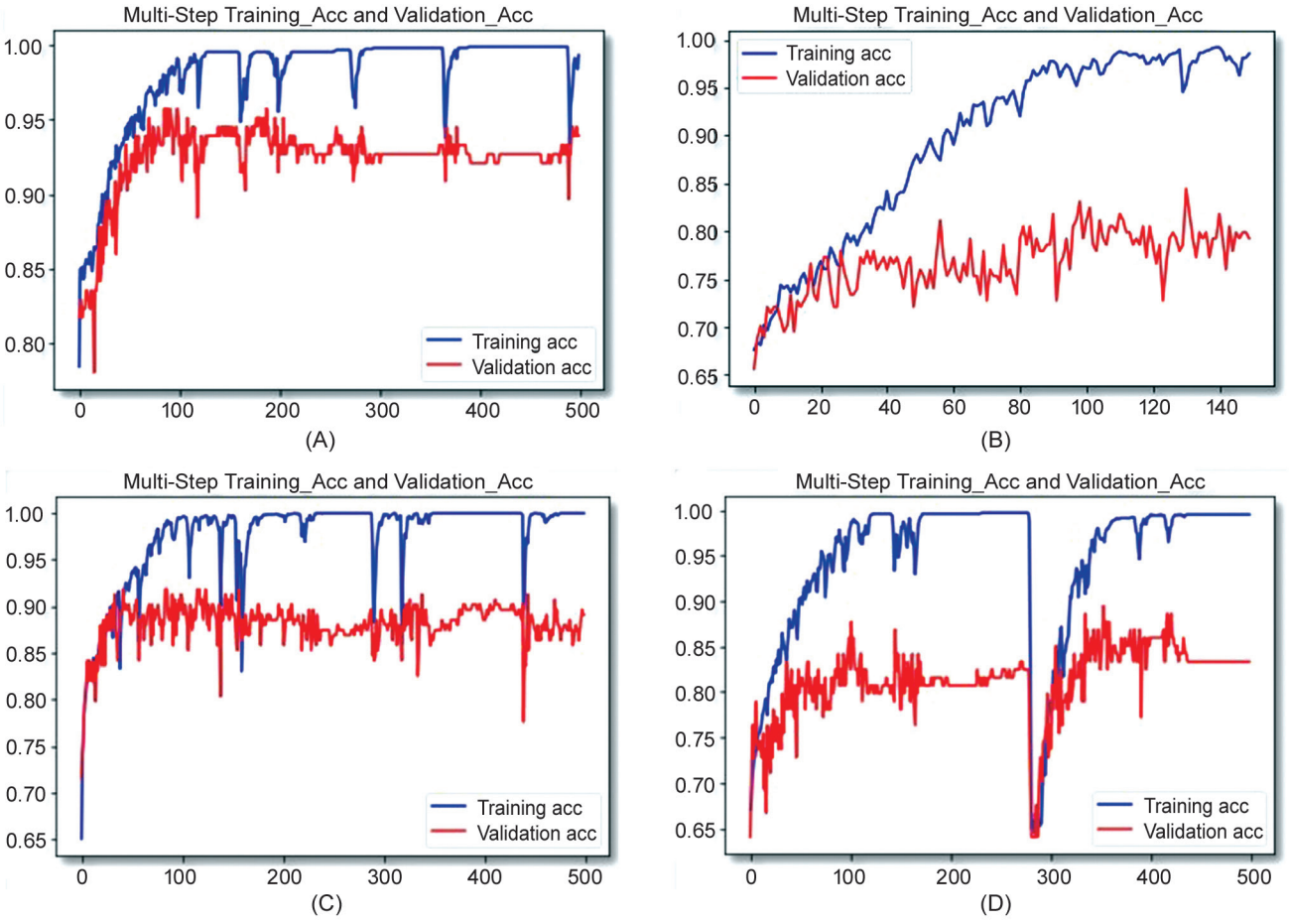

Fig 1 Graphical representation of training and validation accuracy curves of DL-LSTM model at 500 epochs for (A) Cold, (B) Drought, (C) Heat and (D) Salt.

Table 2 Comparative performance of classifiers (SVM, RF and DL-LSTM) for different stress proteins classification

	Cold			Drought			Heat			Salt		
	DL-LSTM	SVM	RF	DL-LSTM	SVM	RF	DL-LSTM	SVM	RF	DL-LSTM	SVM	RF
Accuracy	95.11	93.89	87.53	76.88	75.07	67.53	94.31	92.78	87.96	82.04	80.63	79.93
Precision	95.00	96.00	96.00	69.00	83.00	84.00	93.00	98.00	98.00	77.00	92.00	88.00
Sensitivity	92.00	87.00	69.00	67.00	41.00	15.00	94.00	85.00	74.00	66.00	46.00	46.00
F1 Score	93.00	91.00	81.00	68.00	55.00	25.00	93.00	91.00	84.00	71.00	62.00	61.00
Specificity	97.00	98.00	98.00	83.00	95.00	98.00	94.00	99.00	99.00	90.00	97.00	97.00
MCC	89.60	87.00	74.00	50.00	44.90	25.70	88.40	85.00	77.20	58.60	53.40	55.70

classification in agricultural crops, different factors affecting (such as biotic and abiotic factors) crop production etc. We conducted a study on classification of abiotic stress proteins from the crops belonging to family poaceae. Here, different abiotic stress condition such as cold, heat, drought and salinity were considered and we collected the protein sequences of the same, available in public domain. Features were extracted and different methodologies such as Random forest (RF), Support Vector Machines (SVM) and Deep learning were applied to classify them. We applied Long Short Term Memory (LSTM) deep learning method on the data. Results of these methodologies have been compared after using 5-fold cross validation. It was observed that deep learning outperformed in all conditions. We can conclude that LSTM based methodology has improved the accuracy. However, it is also concluded from the results that as accuracy deceases, recall, F1 and F2 fluctuate more in case SVM and RF, but it remains almost consistence and nearer to accuracy in LSTM. Also, it is observed that the accuracy of LSTM increase with the increase of dataalong with the tuning of hyper parameters. There is/are an opportunist(s) to explore different hyper-parameters, customize hyper-parameter(s) can improve the accuracy of deep leaning to reduce true negative and false positive rate.

System requirement: These analyses have been carried out using a free and open-sourceprogramming known as Python, version 3.7.8. A graphical user interface (GUI) conda package with a local Anaconda Repository was used for implementing the python and executing the codes in a Jupyter Notebook environment. Different libraries were installed in Python for executing these machine learning and deep learning models. Some basic and important libraries are:

NumPy: Itsupports multi-dimensional high-level mathematical functions for matrices and arrays and latest version used (v. 1.19.1).

Pandas (v.1.1.1): It provides Data Frame manipulation of data along with indexing to reduce data memory in different file formats, data split, missing values handling, insertion, deletion etc.

Matplotlib (v): It embeds plots into applications for any kind of purposes. It also helps SciPy for plotting the data.

Seaborn: It is a data visualization library for drawing attractive and informative graphics from the data.

Sequential: Allows to define the order of execution irrespective of content of the function etc.

All analyses were performed in HP-Z400-Workstation dual booting system where Linux - Ubuntu version with 16.04 LTS is used with memory of 99.3 GB. The RAM of the system was 16 BGB with a processor of Intel® Xeon(R) CPU W3565 at 3.20GHz × 4 having NVC1 graphics.

ACKNOWLEDGEMENTS

The authors are grateful to Indian Council of Agricultural Research (ICAR), Ministry of Agriculture and Farmers' Welfare, Government of India for providing financial and infrastructural support to carry out this research and for creation of Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility where the work was carried out. The authors further acknowledge the supportive role of Director of ICAR-IASRI, New Delhi. The grant of IARI Merit scholarship to the first author is duly acknowledged.

REFERENCES

Arel I, Rose D C and Karnowski T P. 2010. Deep machine learning-a new frontier in artificial intelligence research. *IEEE Computer Intelligent Magazine* **5**(4): 13–18.

Bal S, Saha S, Fand B, Singh N, Rane J and Minhas P. 2014. Hailstorms: Causes, damage and post-hail management in agriculture. *Technical Bulletin* **5**: 44.

Biau G. 2012. Analysis of a random forests model. *Journal of Machine Learning Research* **13**: 1063–95.

Bergstra J S, Bardenet R, Bengio Y and Kegl B. 2011. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems* **1**: 9.

Breiman L. 2001. Random Forests. *Machine Learning* **45**: 5–32.

Cortes C and Vapnik V. 1995. Support-vector networks. *Machine learning* **20**(3): 273–97.

Deng Y and Li D. 2011. Deep learning and its applications to signal and information processing. *IEEE Signal Proc Mag* **28**(1): 145–54.

Eckle K and Schmidt-Hieber J. 2019. A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Networks* **110**: 232–42.

Hendrycks D and Gimpel K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:*1606.08415.

LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. *Nature* **521**(7553): 436–44.

Otoom A F, Abdallah E E, Kilani Y, Kefaye A and Ashour M. 2015. Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and its Applications*: **9**(1): 143–56.

Qiu J, Wu Q, Ding G, Xu Y and Feng S. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing* **1**(67): 1–16.

Reddy S, Reddy K T and Kumari V V. 2018. Optimization of deep learning using various optimizers, loss functions and dropout. *International Journal of Recent Technology and Engineering (IJRTE)* **7**(4S2): 448–55.

Roh Y, Heo G and Whang S E. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering* 1–20.

Sarwar H. 2013. The importance of cereals (Poaceae: Gramineae) nutrition in human health: A review. *Journal of Cereals and Oilseeds* **4**(3): 32–35.

Sak H, Senior A W and Beaufays F. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling, pp 1–5.

Tauber L and Sánchez V. 2002. Introducing the normal distribution in a data analysis course: specific meaning contributed by the use of computers. *Proceedings of Seventh International Congress for Teaching Statistics, Citeseer*, pp 1–6.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*, 1-334. Springer, New York.

Vieira S, Pinaya W H and Mechelli A. 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews* **74**: 58–75.

Wen M, Cong P, Zhang Z, Lu H and Li T. 2018. DeepMirTar: a deep learning approach for predicting human miRNA targets. *Bioinformatics* **34**(22): 3781–87.

Young-Man K, Yong-woo K, Dong-Keun C and Myung-Jae Lim. 2019. The comparison of performance according to initialization methods of deep neural network for malware dataset. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* **8**(4S2): 57–62.

Zhang J M, Harman M, Ma L and Liu Y. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 1–37.