IEEE Instrumentation & Measurement Magazine

Intelligent Agricultural Machinery Using Deep Learning

--Manuscript Draft--

Manuscript Number:	IMM-D-20-00078R1
Article Type:	Special Issue
Keywords:	Machine learning, signal processing
Corresponding Author:	Gabriel Thomas University of Manitoba Winnipeg, Manitoba CANADA
First Author:	Gabriel Thomas
Order of Authors:	Gabriel Thomas
	Simone Balocco, Dr.
	Danny Mann, Dr.
	Avery Simundsson
	Nioosha Esmaeilzadeh Khorasani
	Nioosha Khorasani

Intelligent Agricultural Machinery Using Deep Learning

Gabriel Thomas, Simone Balocco, Danny Mann, Avery Simundsson, and Nioosha Khorasani

Artificial intelligence, deep learning, big data, self-driving cars …, these are words that have become familiar to most people and have captured the imagination of the public and have brought hopes as well as fears. We have been told that artificial intelligence will be a major part of our lives, and almost all of us witness this when decisions made by algorithms show us commercial adds that are specifically targeting our interests while using the web. In this paper, the conversation around artificial intelligence focuses on a particular application, agricultural machinery, but offering enough content so that the reader can have a very good idea on how to consider this technology for not only other agricultural applications such as sorting and grading produce, but also other areas in which this technology can be a part of a system that includes sensors, hardware and software that can make accurate decisions. Narrowing the application and also focusing on one specific artificial intelligence approach, that of deep learning, allow us to illustrate from start to end the steps that are usually considered and elaborate on recent developments on artificial intelligence.

Agriculture importance as not only providing the world with food but also as an economic factor

Agriculture plays a significant role in Canada's economy. From a recent release from Statistics Canada, we can highlight that the sector contributed $49.0 billion to Canada's gross domestic product (GDP) in 2015, accounting for 2.6% of total GDP. Agriculture industries contributed $25.1 billion or 51% of GDP in the sector, while agri-food manufacturing industries contributed $23.9 billion or 49%. In the future, the industry is expected to grow at a considerable rate. "Canada's agriculture industry can generate $11 billion for the country's GDP annually by 2030 if the government invests in people and technology" as indicated in [1]. Canada is a global leader in agricultural production and by taking on research and development on new technologies this can be of great benefit to the country and the world, as this activity is vital and will help the country to remain competitive.

Technology impact in agriculture

The impact of technology is an old story, dating back more than 4000 years. Historians agree that the invention of the plow has been a major factor on the transformation of humanity [2]. Agricultural engineers have contributed to numerous advances in agricultural machinery during the past century, and we are now beginning to see prototype autonomous agricultural machines being designed and built by companies and universities around the world. To date, such vehicles are not commercially available due to several limitations (for instance, difficulty operating in dynamic environments with uneven terrain and the need for control systems to account for the intricacies of specific tasks). As these obstacles are overcome, humans will likely transition from machinery operators to a more supervisory role. The advent of precision agriculture technologies in the past decades has increased our ability to cope with on-farm and in-field variability and to incorporate a much larger volume of information into management decisions [3]. Different sensors have been proposed to incorporate information such as microphones, as in-cab operators can extract auditory information to quickly discern the state of mechanical operation. In a supervisory role that may be off-site, these channels of information may be harder to access without creative and innovative monitoring methods. Many operations in agriculture are time-sensitive and machinery breakdowns can be very costly in terms of both time and service. It may be difficult to monitor operations with many moving parts, such as a harvester, without the benefit of sensory information. A few seconds of delay in processing information may be the difference between preventing a failure and a serious harvest delay. Other sensors have been used such as image cameras. In [4], researchers at the University of Saskatchewan in Canada used a support vector machine to perform classification of potato diseases using images that were segmented and features extracted from the leaves were used as an input to the classifier. Features used in [4] were based on color information and statistical features obtained using a gray level co-occurrence matrix. Thus, computer vision and machine learning techniques did an excellent job as they reported classification accuracies of 95%. Now, if one would like to think of using newer techniques to possibly improve the 95 % success classification, Deep Learning (DL) comes to mind. This last discussion leads to a very good question: is using DL a good idea for such applications? The purpose of this paper is to guide readers towards making such a decision as well as the different aspects that need to be considered. For example, going back to the work presented in [4], as 300 images

were used, maybe this data set is not big enough for implementation using a Convolutional Neural Network (CNN) as it usually requires thousands of images to train successfully. Having said that, we would like to present this paper with a series of questions the reader may ask, on the understanding that not all the questions may be relevant to everybody and that some may be skipped.

Has DL been used recently in agricultural machinery applications and what is needed to develop a proof of concept based on this technology?

DL has reached relevance in agricultural applications, for example for avoiding obstacles in autonomous agriculture machines using CNN architectures such as Alexnet [5] and ICNet [6]. Another CNN, AMTNET was used for automatic recognition of agricultural machinery images [7]. Regarding the ability to implement a proof of concept system, there are available pre-trained CNNs that can be used for these agricultural applications. Using free software such as PyTorch [8] and TensorFlow [9], these networks can be modified to adjust the number of classes needed for a specific application by modifying the last output layer. Additionally, the input layer can be modified to fit the size of the data. Table I shows some of the different pre-trained CNNs available in Matlab and Figure 1 shows a comparison of these pretrained network prediction times versus their complexity using a computer with a GPU card [10]. A trade off is expected between number of layers and prediction time and accuracy. As shown in Figure 1, Alexnet offers the least accuracy in the slowest prediction time and at the other end of the spectrum we found nasnetlarge. The database used in Figure 1[11] consists of general images such as musical instruments, furniture, animals etc. but this is not an impediment for training a pretrained network with data for a different application as during training the network starts with random values and with training algorithms such as gradient descent these values adjust to a different application.

Table I

Examples of pretrained CNN networks. Depth indicate the number of layers.

Network	**Depth**
squeezenet	18
googlenet	22
inceptionv3	48
densenet201	201
mobilenetv2	53
resnet18	18
resnet50	50
resnet101	101
xception	71
inceptionresnetv2	164
shufflenet	50
darknet19	19
darknet53	53
alexnet	8
vgg16	16
vgg19	19
nasnetlarge	1244

An important aspect to consider when selecting a pretrained network is that of the time the network requires to obtain an output. From a study of driver responses that took place on highways in personal vehicles covering a variety of driving conditions [12, 13] and assuming that reaction times to stimulus while driving farm machinery would be similar, we can estimate that a decision made in less than one second is a good benchmark. Therefore, any system that can provide a real-time reaction time (time from sensing the issue to implementing a response mechanism) of less than 1 second can be considered to be faster than a human response, and sufficient for a vehicle control system [14].

For fast output, one can design a CNN with few layers such as the one proposed by researchers at the Universities of Calgary and Saskatchewan [15] where an architecture consisting of three convolutional layers followed by a 3 × 3 pooling layer was used for leaf counting in rosette plants.

Moreover, in the last few years a novel network architecture called R-CNN has been proposed for object detection in real-time applications. The initial convolutional pipeline was proposed by Girshick et et al. [16] in 2014 and won the Pascal VOC challenge. The author successively proposed an updated version of the architecture (so called Fast R-CNN) [17] which jointly trained the CNN, classifier, and bounding box regressor using fully connected layers. The network was successively improved by Shaoqing Ren (Faster R-CNN) [18] by replacing the Region Proposal Network with a fully convolutional region extraction technique. Such network is particularly used in agricultural applications since it allowed a reliable and real-time processing of the images, as shown in [19, 20, 21].

The discussion above has been focused on images. As CNNs were very successful for tackling the problem of image description when compared to what had been available before, the pretrained networks often use inputs in the form of images consisting of matrices of unsigned integers with pixel intensity values ranging from 0 to 255 and of fixed sizes that vary from 224 by 224 to 331 by 331 as the ones listed in Table I. If images of different sizes are available, one can also rescale them to fit one of these networks, the image pattern will be distorted but the CNN network is expected to find features from these distorted images as well as if the images were kept in their original sizes. If audio is to be used, one can transform these 1D signals into 2D images via time-frequency or time-scale analysis [22, 23, 24, 25, 26]. Furthermore, a newer type of deep network that allows for the inclusion of temporal behaviour, such as the long-short term recurrent neural network, has proven to be very efficient for audio applications [27, 28] where the inputs are not the standard unsigned 8 bit 2D data that form images but real numbers that form 1D input vectors. Inputs then can be in the form of data coming from sensors that record vibration, humidity, temperature, etc.

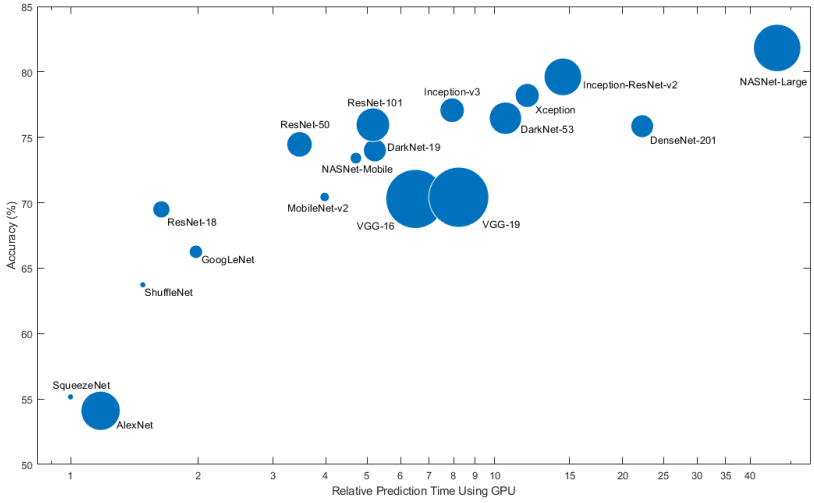

Figure 1: Comparison of different pretrained CNN networks in terms of accuracy, complexity (blue area size) and decision-making times

Not enough data, what can be done?

As long as data is available and labeled with its correct classification to train a network, a prototype can be designed. This availability of data can actually be an obstacle as many examples, in the order of thousands, are usually needed when using a deep network such as the CNNs mentioned before [29]. One way to alleviate the problem of a smaller data set is by creating more images doing operations such as rotation, scaling, shifting, etc. of the original data. This step is known as data augmentation [30]. Nevertheless, it is highly recommended to start with a big data set that can include different scenarios such as possible occlusions of objects, different types of illumination, or any other factors that would generate very different images that were not considered during the network training. In the case of audio, one must include cases that can be foreseen, such as noise coming from traffic if the system is to be recording near a

public road. As indicated in [31], it is not just the amount of data but the data significance and usefulness for the application that has to be considered.

The scenario of not having enough data is likely when an innovative solution to a particular problem is sought. In such cases, one can start with a shallow network by using a conventional neural network. Not as complex as CNNs, these networks can still use low resolution images and have been used in the past for self-driving vehicles. An example of this is ALVINN (Autonomous Land Vehicle in a Neural Network) which back in 1989, used a simple NN consisting on one hidden layer with 29 nodes, having as an input 30x32 video images as well as input from a range finder that was trained to yield 45 direction outputs [32]. ALVINN can be considered as a proof of concept and because of the accuracy importance of such application, is not up to now that DL is taking on commercial autonomous vehicles having reached much better results than the ones reported in [32]. Usually such shallow networks take on inputs in the form of features which total number is much less than the total number of pixels in an image. They execute very fast as the network requires only a matrix multiplication and the results of this multiplication are evaluated with a sigmoid function in which outputs can be logical ones and zeros by simply thresholding the output or without this thresholding, the numerical output can give an indication on how probable the network found that particular output to be. This can actually lead to a preliminary system in which a shallow network is designed first in which classifications made with high probabilities are used to create a new labeled database and the less likely classifications can be manually assessed by an expert, a process known as active learning [33]. Once a bigger data set is formed this way, a DL can potentially improve the accuracy later.

Training databases available

Computer vision applications require the training on large amount of images. A popular training strategy is to exploit training databases such as COCO, PASCAL VOC, SUN. Microsoft COCO is a large-scale object detection, segmentation, and captioning dataset for object segmentation and recognition in context including 330K images [34]. Pascal VOC [35] provides standardized image data sets for object class recognition. SUN [36] provides a benchmark for scene categorization. The Places [37] database introduced by MIT provide a novel scene-centric database with 205 scene categories and 2.5 millions of images with a category label. The importance of this dataset is not only to provide large data-set for network training, but also international challenges allowing researchers to compare novel network architectures. However, training on such large data-bases sometimes requires hardware and GPU resources not accessible to all research groups. For this reason a popular training strategy is to use transfer learning techniques for adapting networks pre-trained on public databases to custom image data-sets [38].

Hardware and preprocessing considerations

If we are starting with no data, this can be seen as an advantage as data collection can be done in such a way to facilitate the implementation of a system. Take for example improper illumination from light reflections. Figure 2 shows an example of this scenario. In this case, the objective is to grade tomatoes into different quality categories. If the shape of a tomato is to be important, that is, how round the object is, this feature can be easily calculated using circle detection and find how many pixels in the border of the tomato are within that circle. Figure 2 (a) and (b) show those normalized numbers to be 0.3434 and 0.1365 and this feature can be good enough and may not require any level of sophistication in terms of what type of classifier to use. However, if defects on the tomato are also being considered, then the reflection from the light source can present difficulties as shown in parts (c) and (d). This may be solved by using a circular polarizer in the camera acquisition system or diffusing the light source. A DL approach would require thousands of images with different illumination possibilities so that the features can be found automatically under less than optimum illumination circumstances.

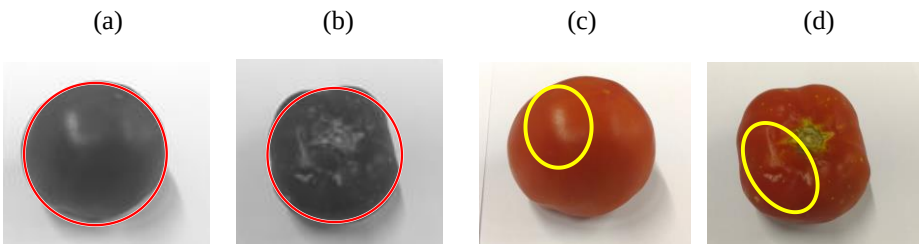

Figure 2. Reliable features such as roundness depicted in (a) and (b) and possible illumination problems that can be caused by light reflections.

Thus, no data can actually be a good start as better data can be obtained considering the patterns a network is about to learn from the data set. Even preprocessing the data before using a neural network can help. Take for example grading dates. Figure 3 shows colour transformations of the gray scale value images of dates by easily converting them into color using a colormap and linearly assigning a particular gray level value to a specific value in the colormap table. The three targeted date categories are soft, semi soft and hard dates. Using 900 images, Figure 3 show how training accuracy is improved when using a shallow CNN.

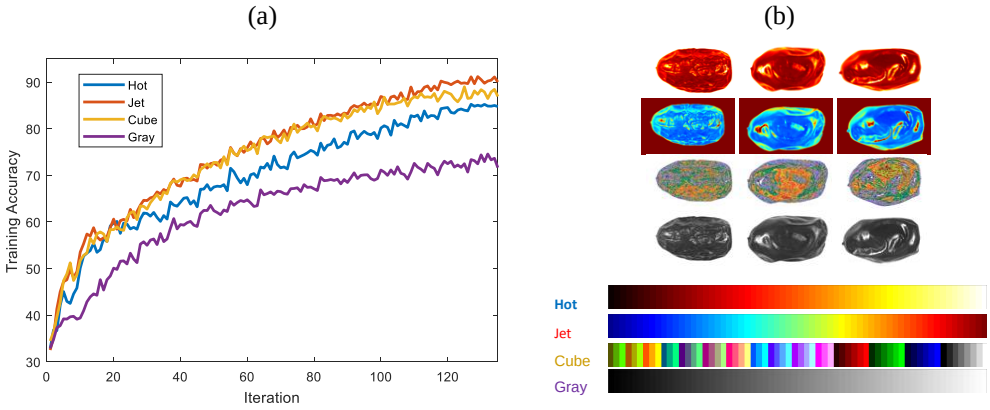

Figure 3. (a) Training accuracies. (b) Date images and colormaps used.

For real time applications, not only the complexity of the classifier is in question, but also the hardware that will make those decisions. If a stand-alone instrument is to be designed as a second step of a successful proof of concept, NN can be very fast for decision making. As CNNs require the calculation of features from the input images via convolutions, the calculations take more time. Here there is a trade off again, as a shallow conventional NN can compute an output extremely fast, but features need to be calculated first and that can take some time too. These features though can be calculated really fast using FPGAs. For example, circle detection used in Figure 2 can be implemented with a video capturing board that includes a specialized FPGA [39]. Furthermore, in computer vision, features can be extracted from images using morphological operations, which based on set theory and working with binary images, can be computationally fast [40]. Thus, not only the complexity of the CNN is in question, but also the hardware that will make those decisions. Low-cost digital signal processing boards, such as the Texas Instruments C5535/C5545 eZdsp USB Stick Development Kit, can be used to analyze the auditory data in real-time and implement the matrix multiplication needed using a NN. For DL solutions, Google Coral / Edge TPU boards can be used. The costs are quite similar and all the computations can be carried out on such devices.

If 1D input data is used and features are to be calculated from the frequency signatures of the data, the FFT is such a fast algorithm that it has been even successfully implemented in real time on mobile phone platforms [41]. Even though features are to be calculated, the computational cost can be minimal. This time-consuming feature extraction is one good reason why only a few features are used and also highlights the importance of selecting the ones that are the most useful [31]. For that selection of features, there are very well-known statistical tools that can reduce the

number of features, for example, selection of features corresponding to the maximal statistical dependency criterion based on mutual information [42]. Sequential forward selection is widely used due to its simplicity and efficiency [43, 44]. Another popular feature reduction method, principal component analysis [45], is an alternative feature selection method that transforms the original variables into another feature space based on principal components.

Figure 4 summarizes what has been discussed and shows how the instrumentation part can add many steps to an effective solution before data is to be used for training and decision making using a NN or DL network. Thus, we note that it is not just a matter of inputs – classifier – output, but several other considerations can yield a more efficient device.

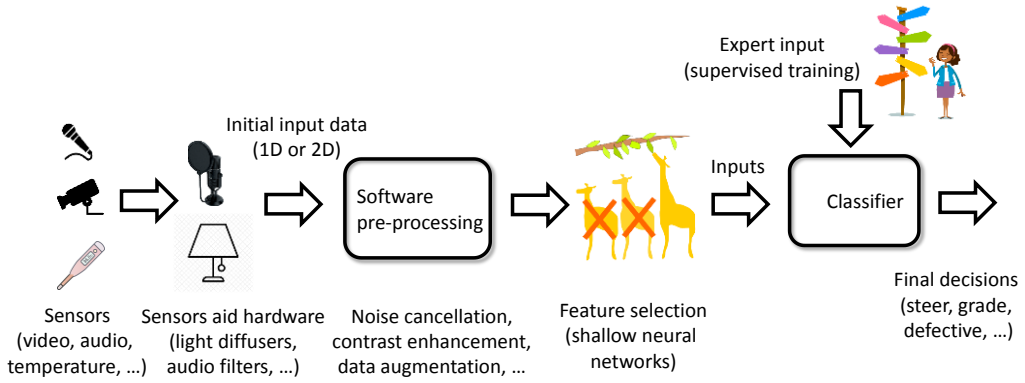

Figure 4. Possible hardware and software needed for an automatic decision-making instrument.

Proof of concept example

The work presented here formulates a proof of concept approach for the classification of harvester sounds that will facilitate the creation of an important part of the automatization of this farm vehicle. Sound recordings were taken from harvest video feed during the canola harvest in East Selkirk, Manitoba. The canola was harvested with an S680 John Deere combine and audio was captured with a GoPro Hero Session. The recordings were taken from the rear of the combine near the straw chopper. All recordings were taken in the same field on the same day from the same machine. Sound sampled at a rate of 48 kHz with AAC compression and automatic gain control was converted to .wav files (Waveform Audio Format) for analysis. Sound recordings were isolated into three different operating modes (classes) of the combine:

> Mode 1: The combine's engine is running, and mechanized threshing is not engaged ("Empty"); total recorded time was 9 seconds.
>
> Mode 2: The combine engine is running, and mechanized threshing is engaged with no actual threshing being performed ("Engaged"); total recorded time was 5 seconds.
>
> Mode 3: The combine's engine is running, and mechanized threshing is engaged and utilized at approximately 80% capacity ("Full"); total recorded time was 10 seconds.

CLASSIFICATION BASED ON A CONVENTIONAL NN

Extracting features from the frequency information of a signal via the absolute value of the Fourier transform, Periodogram (PG), can be an effective way to extract information to be used in a classifier, as for example it was used in [46] for roller bearing fault diagnosis. For the harvester sounds, the position of the highest peaks differ and the energy of different frequency bands also differs between the different 3 modes listed before. With this in mind, a series of features were extracted from the location of the three strongest peaks, the ratio of the amplitudes of the first and second strongest peaks, the distance between the two strongest peaks and the PG center of gravity. These features can be easily normalized and be robust to the size of the FFT used as well as the amount of noise, as noise tends to spread over a large frequency band and does not considerably affect these feature values.

Defining the PG as $X(f)$, the calculated features are:

$\widehat{f_1} = X(f_1)$ where $X(f_1) \geq X(f_i)$
$\widehat{f_2} = X(f_2)$ where $X(f_2) \geq X(f_i) \ \vee_i \notin i = 1$
$\widehat{f_3} = X(f_3)$ where $X(f_3) \geq X(f_i) \ \vee_i \notin i = 1{,}2$
$\widehat{f_4} = |f_1 - f_2|$
$\widehat{f_5} = \frac{\sum_{i=1}^{N} iX(f_i)}{\sum_{i=1}^{N} X(f_i)}$
$\widehat{f_6} = \frac{X(f_1)}{X(f_2)}$

These features become the inputs to a classifier and as it is well known that the normalization process for the inputs can have great effect on preparing the data to be suitable for training [47]. Normalization was done by dividing the above features by the total number of frequency samples *N* used.

$f_i = \frac{\widehat{f_i}}{N}$ for $i = 1{,}2, \ldots{,}6.$

As a classifier, a neural network with one hidden layer with 5 nodes using backpropagation for training yielded accuracies of 78.26 when using 5000 samples of audio segments. The total time required for feature extraction and final classification was 2.5 ms using an Asus laptop with a 64 bit Intel i7 CPU @ 2.6 GHz and 16 GB of ram. We used the Matlab platform version 2017b using parallel processing via an NVIDIA GeForce GTX 960 M GPU card with 10 Gb of memory. Figure 5 (a) show the steps taken.

CLASSIFICATION BASED ON A CNN

The absolute value of the Short-Time Fourier Transform (STFT), known as the Spectogram (SG) was used to obtain images. A seven-layer CNN was used. The layers are:

1. Image input layer with 'zerocenter' normalization
2. Convolution layer, eight 8x8x3 convolutions with stride [1 1] and no padding
3. ReLU layer
4. Average Pooling layer with 2x2 average pooling with stride [2 2] and no padding
5. Fully Connected layer
6. Softmax
7. Classification Output layer, crossentropyex

Using the same number of samples as before, the CNN achieved accuracies of 97.97 in an execution time below the one second mark using the same equipment and software. Figure 5 (b) shows the steps.

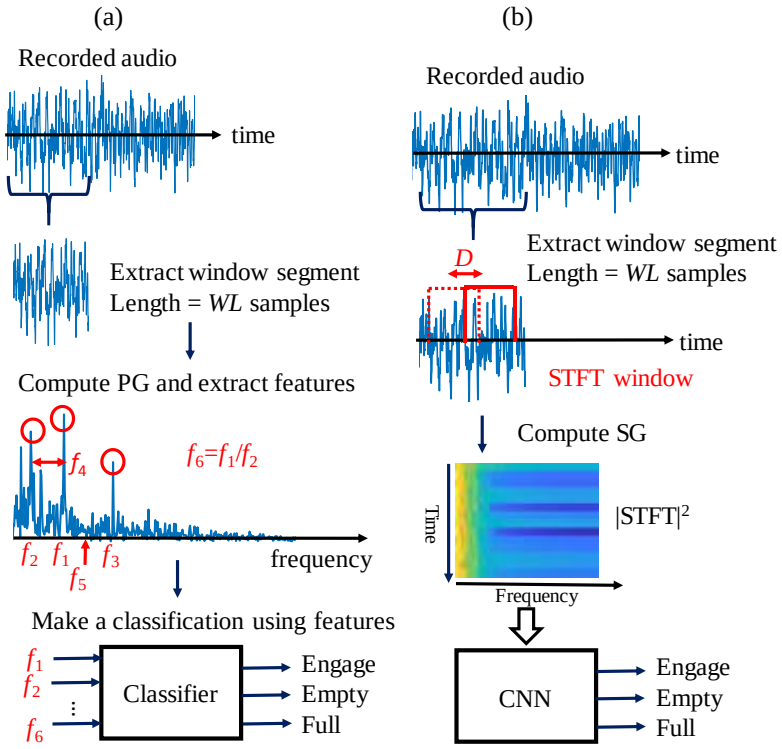

Figure 5. Steps needed for the classification of a harvester based on sound recording. (a) Using a conventional neural network. (b) Using a CNN.

For comparison purposes, using the same six features and same number of samples, five different classifiers yielded the accuracies given in Table II.

TABLE II
CLASSIFICATION TESTING ACCURACY

Method	Accuracy
Linear Discriminant	54.88
Linear Support Vector Machine	55.33
Ensemble Bagged Trees	53.97
k-Nearest Neighbor	56.01
Binary classification decision tree	55.1
NN	78.26
CNN	97.97

We can see how promising the use of DL can be and we also highlighted the impact that data acquisition and data preprocessing can have for a system. The topics discussed here not only apply to agricultural machinery but these ideas can also play an important role in other DL applications.

Conclusions

In this paper, the recent techniques for image and signal classification are reviewed with particular emphasis to agricultural applications. The experiments presented in this paper highlight the limitation of classical machine learning techniques vs neural network approaches. As demonstrated, classical machine learning approaches reaches lower performances compared to NN and CNN architectures. However, the need of a larger amount of training data labeled by an expert may not always be available in real world technologies, and the hardware requirements (GPUs) for computation may limit the use of such system in real-time applications.

References

[1] A. Sagan, "Agriculture could add $11B a year to Canada's GDP by 2030," Canadian Business, August 2019.

[2] F. L. Pryor, "The Invention of the Plow." Comparative Studies in Society and History, vol. 27, no. 4, pp. 727–743, 1985.

[3] S. Fountas, C. G. Sorensen, Z. Tsiropoulos, C. Cavalaris, V. Liakos, and T. Gemtos, "Farm machinery management information system," Computers and Electronics in Agriculture, pp. 131-138, 2015.

[4] M. Islam, Anh Dinh, K. Wahid and P. Bhowmik, "Detection of potato diseases using image segmentation and multiclass support vector machine," IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, ON, pp. 1-4, 2017.

[5] K. A. Steen, P. Christiansen, H. Karstoft and R. N. Jørgensen, "Using Deep Learning to Challenge Safety Standard for Highly Autonomous Machines in Agriculture," Journal of Imaging, 2016.

[6] Y. Li, M. Iida, T. Suyama, M. Suguri and R. Masuda, "Implementation of deep-learning algorithm for obstacle detection and collision avoidance for robotic harvester," Computers and Electronics in Agriculture, July 2020.

[7] Z. Zhang, H. Liu, Z. Meng and J. Chen, "Deep learning-based automatic recognition network of agricultural machinery images, Computers and Electronics in Agriculture, Vol. 166, 2019.

[8] A. Paszke, S. Gross, S. Chintala, et. Al, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Advances in Neural Information Processing Systems, 2019.

[9] M. Abadi, P. Barham, J. Chen, et. Al, "TensorFlow: A System for Large-Scale Machine Learning," Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, 2016.

[10] MATLAB 9.8 and Statistics and Machine Learning Toolbox 12.0, The MathWorks, Inc., Natick, Massachusetts, United States.

[11] ImageNet. http://www.image-net.org

[12] P. L. Olson and M. Sivak, "Perception-Response Time to Unexpected Roadway Hazards," Human Factors: The Journal of the Human Factors and Ergonomics Society, 1986.

[13] T. J. Triggs, and W. G. Harris, "Reaction Time of Drivers to Road Stimuli. Victoria : Monash University, 1982.

[14] A. Simundsson , A neural network to classify auditory signals for use in autonomous harvester control systems, M.Sc. thesis, The University of Manitoba, 2019.

[15] J. Ubbens, M. Cieslak, P. Prusinkiewicz, C. Mikolaj and I. Stavness, "The use of plant models in deep learning: an application to leaf counting in rosette plants," Plant Methods, Issue 6, 2018.

[16] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Computer Vision–ECCV, pp. 345-360, 2014.

[17] R. Girschick, “Fast R-CNN”, Proceedings of the IEEE International Conference on Computer Vision, pp. 1440-1448, 2015.

[18] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” Computer Vision and Pattern Recognition, 2016.

[19] D. I. Patricio and R. Rieder, “Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review,” Computers and Electronics in Agriculture, Vol. 153, pp. 69-81, October 2018.

[20] F. Gao, L. Fu, X. Zhang, Y. Majeed, R. Li, M. Karkee and Q. Zhang, “Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN,” Computers and Electronics in Agriculture, Vol: 176, 2020.

[21] P. Christiansen, L. N. Nielsen, K. A. Steen, R. N. Jørgensen and H. Karstoft, “Deep Anomaly: Combining Background Subtraction and Deep Learning for Detecting Obstacles and Anomalies in an Agricultural Field,” Sensors, Vol. 16, 2016.

[22] J. Burriel-Valencia, R. Puche-Panadero, J. Martinez-Roman, A. Sapena-Bano and M. Pineda-Sanchez, "Short-Frequency Fourier Transform for Fault Diagnosis of Induction Machines Working in Transient Regime," in IEEE Transactions on Instrumentation and Measurement, vol. 66, no. 3, pp. 432-440, March 2017.

[23] H. Liu, L. Li, and J. Ma, “Rolling Bearing Fault Diagnosis Based on STFT-Deep Learning and Sound Signals,” Shock and Vibration, vol. 2016, 12 pages, 2016.

[24] D. Verstraete, A. Ferrada, E. López-Droguett, V. Meruane, and M. Modarres, “Deep Learning Enabled Fault Diagnosis Using Time-Frequency Image Analysis of Rolling Element Bearings,” Shock and Vibration, 2017.

[25] B. Schwerin, K. Paliwal, “Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement,” Speech Communication, pp. 49-68, Volume 58, 2014

[26] M. Huzaifah, “Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks,” Computer Vision and Pattern Recognition, 2017.

[27] H. Sak, A. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 338-342, 2014.

[28] W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1-4, 2016.

[29] T. Wiesner-Hanks, E. L. Stewart, N. Kaczmar, et al. “Image set for deep learning: field images of maize annotated with disease symptoms,” BMC research notes, Vol.11, 2018.

[30] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” Journal of Big Data, vol. 6, no. 1, 2019.

[31] D. Petri, “Big Data, Dataism and Measurement,” IEEE Instrumentation & Measurement Magazine, Vol. 23 , Issue: 3 , May 2020.

[32] J. Zurada, Introduction to artificial neural systems, West Publishing Co., January 1992.

[33] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, M. Sedlmair, “Comparing visual-interactive labeling with active learning: An experimental study,” IEEE transactions on visualization and computer graphics 24 (1), 298-308, 2017.

[34] T. Lin, M. Maire, S. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll´and C. L. Zitnick, “Microsoft COCO: common objects in context,” arXiv:1405.0312, 2014.

[35] http://host.robots.ox.ac.uk/pascal/VOC/

[36] https://vision.princeton.edu/projects/2010/SUN/

[37] http://places2.csail.mit.edu/

[38] S. Pan and Q. Yang, “A Survey on Transfer Learning,” IEEE Transactions on Knowledge and Data Engineering,” 2010.

[39] A. Elhossini and M. Moussa, "Memory efficient FPGA implementation of hough transform for line and circle detection," 2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Montreal, QC, pp. 1-5, 2012.

[40] E. R. Dougherty, An Introduction to Morphological Image Processing, SPIE Optical Engineering Press, 1992.

[41] N. Alamdari, and N. Kehtarnavaz, “A real-time smartphone app for unsupervised noise classification in realistic audio environments” Proceedings of IEEE Signal Processing in Medicine and Biology Symposium, Philadelphia, PA, Dec 2018.

[42] H. Peng, F. Long, C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Trans. On Pattern Analysis and Machine Intelligence, vol.27, no.8, pp. 1226-1238, Aug. 2005.

[43] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina-Januchs, and D. Andina, “Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network,” IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society, pp. 2845-2850, Nov. 2010.

[44] J. Schenk, M. Kaiser, and G. Rigoll, “Selecting features in on-line handwritten whiteboard note recognition: SFS or SFFS?,” IEEE 10th International Conference on Document Analysis and Recognition, pp. 1251-1254, Jul. 2009.

[45] I. Jolliffe, Principal component analysis, John Wiley & Sons, Ltd, 2002.

[46] L. Song, H. Wang and P. Chen, "Vibration-Based Intelligent Fault Diagnosis for Roller Bearings in Low-Speed Rotating Machinery," in IEEE Transactions on Instrumentation and Measurement, vol. 67, no. 8, pp. 1887-1899, Aug. 2018.

[47] T. Jayalakshmi and A. Santhakumaran, “Statistical Normalization and Back Propagation for Classification,” International Journal of Computer Theory and Engineering, Vol.3, No.1, pp. 1793-8201, February, 2011.

Gabriel Thomas (M’95) received the B.Sc. degree in electrical engineering from the Monterrey Institute of Technology, Mexico, in 1991, and the M.Sc. and Ph.D. degrees in computer engineering from the University of Texas, El Paso, in 1994 and 1999, respectively. Since 1999, he has been a faculty member in the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada, where he is currently an Associate Professor. He is a coauthor of the book Range Doppler Radar Imaging and Motion Compensation. His current research interests include digital image and signal processing, machine learning, nondestructive testing, and computer vision.

Simone Balocco is Associate Professor at the University of Barcelona (Spain) at the Department of Mathematics and Computer Science and a senior researcher at the Center for Computer Vision (CVC), Bellaterra. He received the title of doctor in Acoustics in the CREATIS laboratory, Lyon1 University, Lyon (France) and Electronics and Telecommunications obtained in the MSD Lab of the University of Florence (Italy). His main lines of research are the study of deep machine learning, artificial intelligence, computer vision and automated pattern recognition applied to medical imaging.

Danny Mann received a B.Sc. degree in Agricultural Engineering (University of Manitoba, 1992), an M.Sc. degree in Biosystems Engineering (University of Manitoba, 1995), and a Ph.D. in Biosystems Engineering (University of Manitoba, 1998). Danny joined the Department of Biosystems Engineering in 1998 as an Assistant Professor, was promoted to the rank of Associate Professor in 2004, and was promoted to the rank of Professor in 2008. He has served

as Head of the Department of Biosystems Engineering since 2009. He has research expertise in agricultural ergonomics, agricultural safety, and assistive technologies. His research bridges the disciplines of agricultural engineering and ergonomics.

Avery Simundsson obtained her B.Sc. in Mechanical Engineering from the University of Manitoba and has just recently completed her M.Sc. in Biosystems Engineering at the University of Manitoba. She has worked in manufacturing and agricultural research, and currently works in industry for a software consultancy as a project manager.

Nioosha E. Khorasani obtained her B.Sc. in Computer Engineering from the Amirkabir University of Technology (aka Tehran Polytechnic). She worked in telecommunication industry as a hardware test engineer for two years and currently is a M.Sc. student at the Electrical and Computer Engineering Department, University of Manitoba.

Reviewers' comments:

Reviewer #1:
This work discusses the use of deep learning methods for agricultural applications. The paper is well framed in the magazine's scope overviewing the problem and its context. The use of deep learning-based structures such as CNN is then addressed highlighting the convolutional and classification layers. To restate the advantages of deep learning, a case example is finally given comparing NN and CNN for audio processing.

Overall, the paper represents a good match for the SI in machine learning and signal processing. It is well written; it is easy to read, and to follow. As IMM targets generalist papers including comprehensive descriptions of the topics addressed, I recommend:

1. To expand the SoA on deep learning projects addressing agricultural problems.

We thank the reviewer for the comments. We have included a discussion as well as more references regarding the agricultural problems and solutions [19, 20 and 21].

2. I was surprised of not finding details of Fast and Faster R-CNN in this review (Even though I acknowledge they can be grouped in CNNs). In particular, the former is widely used in real-time applications.

The reviewer mentions an important type of network and we added the following:

Moreover, in the last few years a novel network architecture called R-CNN has been proposed for object detection in real-time applications. The initial convolutional pipeline was proposed by Girshick et et al. [16] in 2014 and won the Pascal VOC challenge. The author successively proposed an updated version of the architecture (so called Fast R-CNN) [17] which jointly trained the CNN, classifier, and bounding box regressor using fully connected layers. The network was successively improved by Shaoqing Ren (Faster R-CNN) [18] by replacing the Region Proposal Network with a fully convolutional region extraction technique. Such network is particularly used in agricultural applications since it allowed a reliable and real-time processing of the images, as shown in [19, 20, 21].

3. Just as the convolutional layer was explained (Table 1), a comparative evaluation of the most popular training databases is recommended (COCO, PASCAL VOC, SUN, etc.).

We included the following:

Computer vision applications require the training on large amount of images. A popular training strategy is to exploit training databases such as COCO, PASCAL VOC, SUN. Microsoft COCO is a large-scale object detection, segmentation, and captioning dataset for object segmentation and recognition in context including 330K images [34]. Pascal VOC

[35] provides standardized image data sets for object class recognition. SUN [36] provides a benchmark for scene categorization. The Places [37] database introduced by MIT provide a novel scene-centric database with 205 scene categories and 2.5 millions of images with a category label. The importance of this dataset is not only to provide large data-set for network training, but also international challenges allowing researchers to compare novel network architectures. However, training on such large data-bases sometimes requires hardware and GPU resources not accessible to all research groups. For this reason a popular training strategy is to use transfer learning techniques for adapting networks pre-trained on public databases to custom image data-sets [38].

4. For agricultural related applications, the limitations of the aforementioned databases must be highlighted and training with custom images should be discussed.

We added:

However, training on such large data-bases sometimes requires hardware and GPU resources not accessible to all research groups. For this reason a popular training strategy is to use transfer learning techniques for adapting networks pre-trained on public databases to custom image data-sets [38].

Reviewer #2: This is an important topic and the authors summarize very well the latest in the field. However, to benefit scientists and engineers involved in agricultural research the authors need to organize better their paper and provide the reader with each of the techniques presented in the paper in a separate section including pros and cons and an example(s) for illustration.

We moved the hardware discussion into a consecutive part as it was discussed in two different sections. In addition, the longest section was broken down into two more sections in order to help the organization of the paper.

A conclusion section has been included to emphasize the pros and cons of the study. Unfortunately, we were forced to reduce the amount of additional text in order to fulfill the maximum length of the paper allowed for the magazine

Conclusions

In this paper the recent techniques for image and signal classification are reviewed with particular emphasis to agricultural applications. The experiments presented in this paper highlight the limitation of classical machine learning techniques vs neural network approaches. As demonstrated, classical machine learning approaches reaches lower performances compared to NN and CNN architectures. However, the need of a larger amount of training data labeled by an expert may not always be available in real world technologies, and the hardware requirements (GPUs) for computation may limit the use of such system in real-time applications.