

## REVIEW

# Interpretability and accessibility of machine learning in selected food processing, agriculture and health applications

N Ranasinghe<sup>1</sup>, A Ramanan<sup>2</sup>, S Fernando<sup>3</sup>, PN Hameed<sup>4</sup>, D Herath<sup>5</sup>, T Malepathirana<sup>1</sup>, P Suganthan<sup>6,7</sup>, M Niranjan<sup>8</sup> and S Halgamuge<sup>1\*</sup>

<sup>1</sup> Department of Mechanical Engineering, University of Melbourne, Australia.

<sup>2</sup> Department of Computer Science, Faculty of Science, University of Jaffna, Sri Lanka.

<sup>3</sup> Department of Computational Mathematics, Faculty of Information Technology, University of Moratuwa, Sri Lanka.

<sup>4</sup> Department of Computer Science, Faculty of Science, University of Ruhuna, Sri Lanka.

<sup>5</sup> Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Sri Lanka.

<sup>6</sup> KINDI Center for Computing Research, College of Engineering, Qatar University, Doha, Qatar.

<sup>7</sup> School of Electrical Electronic Engineering, Nanyang Technological University, Singapore.

<sup>8</sup> Department of Electronics and Computer Science, Faculty of Engineering and Physical Sciences, University of Southampton, United Kingdom.

**Summary:** Artificial Intelligence (AI) and its data-centric branch of machine learning (ML) have greatly evolved over the last few decades. However, as AI is used increasingly in real world use cases, the importance of the interpretability of and accessibility to AI systems have become major research areas. The lack of interpretability of ML based systems is a major hindrance to widespread adoption of these powerful algorithms. This is due to many reasons including ethical and regulatory concerns, which have resulted in poorer adoption of ML in some areas. The recent past has seen a surge in research on interpretable ML. Generally, designing a ML system requires good domain understanding combined with expert knowledge. New techniques are emerging to improve ML accessibility through automated model design. This paper provides a review of the work done to improve interpretability and accessibility of machine learning in the context of global problems while also being relevant to developing countries. We review work under multiple levels of interpretability including scientific and mathematical interpretation, statistical interpretation and partial semantic interpretation. This review includes applications in three areas, namely food processing, agriculture and health.

**Keywords:** Disease detection in agriculture, drug repositioning, food processing, interpretation of neural networks, metagenomics.

## INTRODUCTION

Artificial intelligence (AI) has seen an explosive growth over the last 20 years, largely through recent advances in machine learning (ML) – the data-centric branch of AI. A data-centric AI system consists of an AI model (a structure or architecture) and a method or learning algorithm that enables that model to derive usable information from data. Sometimes the data are exploratory, like the genomic data arriving from different parts of the world about constantly mutating viruses. To discover the presence of new variants or labels, we can feed an AI model with such uninterpreted data, so that researchers will be able to use this AI model to assign labels. Such AI models need unsupervised learning (UL) algorithms to extract information from unlabelled and uninterpreted data. We could also ask those researchers themselves to label the data with appropriate variant labels, and feed both labels and genomic data to an AI model that can then use a supervised learning algorithm like deep learning (DL), so that it can serve as a predictor for known variants of the virus. If such an AI model of sufficient strength requires it to be large, deep and complex, we call it a deep neural network (DNN). Shallow neural networks,

\* Corresponding author (saman@unimelb.edu.au;  <https://orcid.org/0000-0002-2536-4930>)



commonly referred to as neural networks (NNs) are data driven mathematical models consisting of about three layers of artificial neurons or nodes (several linear and nonlinear processing elements) which are interconnected through weighted connections.

Popular models of AI, in particular, ML based models have significant deficiencies preventing its broader application. For example, they are mostly uninterpretable (Halgamuge, 2021). This paper addresses some of these in the context of several major global problems in the post-pandemic world with relevance to developing countries.

Achieving interpretability in ML will promote broader and effective use of ML, by answering core questions about traceability, accountability, ethical compliance, and inherent biases. Serious concerns remain about the reliability of ML systems in certain contexts. For example, the claim about achieving self-driving status for vehicles has not allayed public concerns. Explainable AI (XAI) methods can help in some applications although may not provide a full explanation about the decision process of the ML model.

There are two main strategies that exist for achieving higher accessibility through automated ML model design. In “growing the ML model from scratch” the ML model starts with a default simple model and grows until its capacity is enough to solve the problem. This approach has been used in both supervised and unsupervised learning-based ML systems. In the “search for the best among selected candidates” approach, also called neural architecture search (NAS), the best ML model is selected from a list of candidate solutions. The former is applicable to both labelled and unlabelled data, whereas the latter is generally confined to labelled data.

Most ML research on increasing interpretability of ML systems on the Y-axis and accessibility through increased AI model design automation on the X-axis (Figure 1) lies along or close to the axes. Research into interpretable neural networks designed with minimal expert intervention (FAIR AI) will increasingly close a significant knowledge gap, also informed by relevant studies, for example, ML with continuous and life-long learning capability (Senanayake et al., 2021).

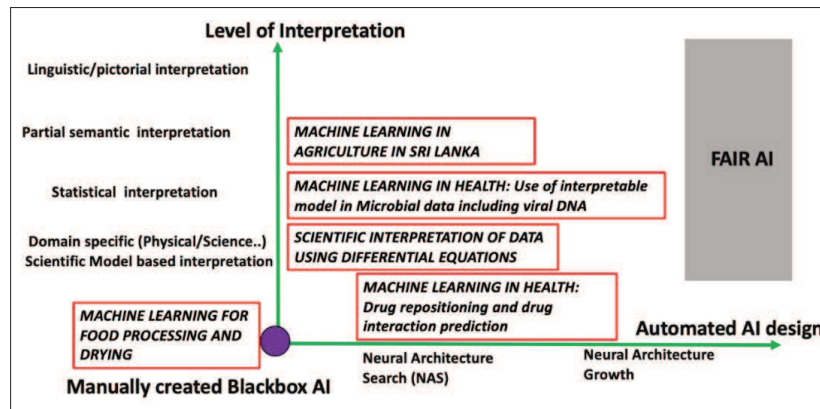


Figure 1: Contributions of the Paper and Fair, Accessible, Interpretable and Reproducible (FAIR) AI adapted from Halgamuge (2021)

This paper is organised with an example of scientific interpretation capability of ML models using differential equations followed by relevant applications of ML in three areas of importance, namely Food Processing, Agriculture and Health: examples of uninterpretable ML models applied to food drying, partially interpretable convolutional neural networks (CNN) and XAI in plant

disease detection including in rice cultivation, safety of taking multiple pharmaceutical drugs and reuse of existing drugs for new diseases using semi-automated unsupervised ML model construction and shedding some light into yet largely unexplored world of microbes including viruses using semi-automated ML model construction.

## SCIENTIFIC INTERPRETATION OF DATA USING DIFFERENTIAL EQUATIONS

Supervised learning problems generally focus on learning a relationship that maps a given input space to an output space based on input-output pairs. This involves training a model to learn this relationship by looking at a labelled dataset. Although most trained models including neural networks can be represented as a mathematical function, interpreting this function would typically be quite challenging. Only a very few models such as linear regression models will have an interpretable mathematical equation. CNNs are ML models with some (pictorial) interpretability. It can be noted that most equations modelling real-world problems do not have a very large number of terms (Brunton *et al.*, 2016). Furthermore, differential equations (DE) are commonly seen as governing equations of dynamical systems. These DEs are generally derived using first principles. Recently there has been a focus on recovering governing differential equations from observation data of dynamical systems.

Udrescu and Tegmark (2020) propose a recursive multi-step method to search through the possible space of equations that fit a given dataset. It uses properties commonly found in real-world physics equations to reduce the search space. These include symmetry, separability, compositionality and more simplifying properties. This paper uses a neural network as a function approximator to discover some of these simplifying properties. The final equation is discovered using a brute force search across the simplified solution space. Udrescu *et al.* (2020) improve upon this approach to discover Pareto-optimal formulae (complexity vs accuracy).

Brunton *et al.* (2016) approach the same problem by using sparse regression techniques to circumvent the problem of searching through a large space of possible solutions. Here, the differential equation is assumed to be in the form of,

$$\frac{d}{dt}X(t) = f(X(t))$$

Where  $X$  is the independent variable and  $f(X(t))$  is assumed to be a linear combination of non-linear functions of  $X$ . A library  $\theta(x)$  of possible terms are constructed using these non-linear functions. The resultant formulation is,

$$\frac{d}{dt}X(t) = \theta(X)E$$

Where,  $E$  includes the coefficients corresponding to each library term. The problem is now in the form of a linear regression problem. The paper uses a sparse regression algorithm to solve for the coefficients since real life differential equations typically do not have many terms in the right-hand side. This idea is improved and applied to learn partial differential equations (PDE) in Rudy *et al.* (2017). Zhang and Lin (2018) propose a method to use Bayesian sparse regression to increase the robustness of the learned equation and quantify the uncertainty of the solution.

Neural networks are also used in recent work to recover governing differential equations from observed data. Martius and Lampert (2016) used neural network algebra to learn equations through backpropagation. A shallow neural network is used with custom activation functions which include multiplications, trigonometric functions and identity functions. Sparsity promoting L1 regularization is used to promote learning a simple equation with a low number of terms. Sahoo *et al.* (2018) proposes a method to extend the class of the learnable equations using an equation learning network to include divisions. Long *et al.* (2019) propose a deep neural network architecture to discover time-dependent PDEs from observed data. This architecture proposes the use of a  $\delta$ -block, which uses convolutions to approximate differential operators and a symbolic neural network to approximate the non-linear response function. The approximation framework is as follows:

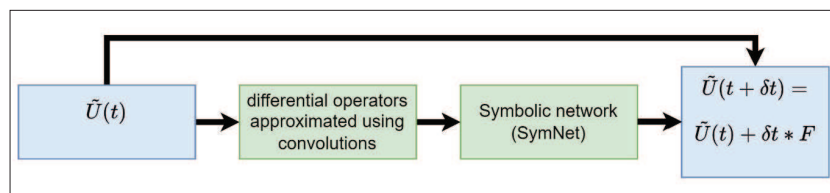


Figure 2: PDE-NET 2.0 (Long *et al.*, 2019)

$$\tilde{U}(t + \delta t) \approx \tilde{U}(t) + \delta t * F$$

where,  $\tilde{U}(t)$  is the predicted value at time  $t$  and  $F$  is the PDE-NET approximation. The architecture of a  $\delta$ -block is shown in Figure 2. A single-block can only approximate one-step dynamics, meaning that it is prone to error accumulation over time. Multiple-blocks with shared parameters are stacked to approximate multi-step dynamics and reduce error accumulation.

Most works in this area focus on learning a certain type of ordinary differential equation (ODE) or PDE. However, when applying these methods to real scenarios, we may not know the type of equation in advance. Furthermore, most of the approaches in the literature can only learn 1<sup>st</sup> order differential equations. Therefore, future work could focus on learning higher order DEs. Scalability is also a challenge in current methods and is especially evident in NN based equation learning algorithms. These are quite resource intensive and in certain cases require GPUs to run. Moreover, only a very few works explore methods of uncertainty quantification of the solution. This is also a potential research direction.

---

## MACHINE LEARNING FOR FOOD PROCESSING AND DRYING

Food security is a major concern around the world as foods have high nutritional value and are an essential part of everyday life. However, most plant-based foods are highly perishable and about one-third of global food production is lost annually due to inadequate processing (Vilariño *et al.*, 2017). This issue is amplified in developing countries where about 40% of seasonal fruits and vegetables are wasted (Karim & Hawlader, 2005). Drying is a major food preservation and processing technique which aims to remove moisture, preventing microbial spoilage (Kumar *et al.*, 2015), thus adding value to a product, permitting early harvest, reducing shipping weights and costs, minimizing packaging requirements and increasing shelf-life (Zielinska *et al.*, 2013). However, drying is highly complex involving interconnected simultaneous momentum, heat and mass transfer with time-varying physiochemical and anisotropic structural changes which depend on the dynamic product-drying environment interactions occurring (Welsh *et al.*, 2018; Khan *et al.*, 2020). Additionally, the underlying mechanisms, which originate within the materials cells, are still not well understood. Due to these complexities, optimising the design of dryers and process conditions is challenging and expensive. AI, for example, uninterpretable ML,

has demonstrated great potential to facilitate innovation and optimisation for the processing of foods, providing a low-cost alternative to the current design modelling approaches and in-line process controllers. Converting these ML models to interpretable ML models will have significant advantages.

### Uninterpretable Machine Learning

ML (e.g. NN) is able to investigate, model and predict the nonlinear time-varying behaviour of food material during drying (Angermueller *et al.*, 2016). The three main parts of NN represent the input layer (material definition and/or process conditions), the hidden layer (consisting of nodes and weighted connections) and the output (drying rate, moisture content evolution and/or quality). The approach utilises experimental examples of the system which is being modelled (training data) to optimize the parameters including weights of the connections between neurons to predict the complex outputs. Over the years various NN models have been applied in food drying for predicting complex outputs for various materials (Sun *et al.*, 2019) and various drying technologies (Sarkar *et al.*, 2020). Generally, ML is applied to predict four types of outputs for food processing, the evolution of moisture and temperature, morphological/structural changes, transfer coefficients and quality changes.

Common complex outputs predicted through ML models include the moisture and temperature evolution of the material and the drying rate/transfer coefficients of a dryer. Predicting such complex outputs can provide insight into the drying time and the energy consumption of a dryer. Chasiotis *et al.* (2020) utilised a NN to predict the moisture content evolution of convective drying cylindrical quince slices. This work utilised 1372 experimental samples split between training data and cross validation data to consider three input variables (temperature, flow velocity and time) to predict the samples' moisture content. The results showed good agreement between the predicted and experimental values. Çerçi and Daş (2019) applied a NN and decision tree to predict the heat transfer coefficient for natural and force convection, concluding NN was more successful in estimating the heat transfer coefficient. Saraceno *et al.* (2012) investigated and compared three different modelling approaches, a thin-layer model, a NN and a hybrid neural model for two different vegetables with different characteristic dimensions for a wide range of process conditions. Their work demonstrated that pure neural models gave very accurate predictions when modelling/reproducing known data/process conditions. However, the accuracy of the NN model significantly



decreased when attempting to extrapolate/apply the model to unknown scenarios/process conditions.

The morphological and quality changes food material experiences during drying have also been predicted through ML techniques. Scala *et al.* (2013) constructed a NN model for predicting the quality characteristics of Granny Smiths apples during convective dehydration. The work investigated experimental data from three different drying temperatures (40, 60 and 80°C) drying at three air flow rates (0.5, 1.0 and 1.5 m/s) and effectively predicted the colour, water capacity and total phenolic content of the samples. Additionally, the work identified the optimal drying conditions within the experimental process conditions. Chen and Martynenko (2013) utilised computer vision to evaluate the drying rate, shrinkage and colour changes of two varieties of blueberry, though the accuracy of the shrinkage measurements was limited by the pixel resolution of the camera and the accuracy of the colour measurement was limited by the quality of illumination and colour reproduction. Recently, Sinha and Bhargav (2022) developed an ANN model to predict key properties related to shrinkage, specifically solid density, initial porosity and initial water saturation of a given food material, using temperature and moisture data from a set of simple experiments. The work demonstrates how a NN model can serve as an efficient indirect method of property estimation in food material.

ML techniques have also been applied to calculate key transport properties. Mariani *et al.* (2008) developed an NN-based inverse method to estimate the apparent diffusivity of bananas at different drying temperatures. The work found that a small change in drying temperature and moisture content caused a significant change in bananas' apparent diffusivity. Sablani and Rahman (2003) developed an NN model for predicting the thermal conductivity of various foods in terms of moisture content, temperature and apparent porosity. NN models have demonstrated great predictive capabilities in comparison to other statistical approaches.

Although AI and ML techniques provide a low-cost alternative to facilitate innovation and optimisation for the processing of foods, applying ML to food processing does have some challenges. ANN is considered to be a 'black box' approach (i.e. uninterpretable) where the user cannot see what is happening during the simulation which is not ideal for understanding what is occurring during drying. Additionally, ML is a data-driven approach where the accuracy of the model heavily depends on acquiring a large comprehensive training dataset. However, food processing data is often scarce

and obtaining large datasets is expensive due to the resources and costs involved in measuring this data experimentally. As a result, most NN models in literature have been constructed using small datasets and therefore their accuracy significantly decreases when applying the model to data outside the training observatory data (Saraceno *et al.*, 2012). Recently, a new class of NN or deep learning, physics-informed neural networks (PINN), has emerged that can seamlessly integrate training data and complex mathematics to optimise a loss function. By incorporating prior knowledge, it minimises the need to have a large observatory training dataset for maintaining the high accuracy of the predictive model. Hence, overcoming the main challenges for AI in food processing. A PINN model can be trained with additional information obtained by enforcing physical laws with mathematical governing equations (Karniadakis *et al.*, 2021). PINN models have been constructed for the deformation of elastic plates (Li *et al.*, 2021), engineering heat transfer applications (Zobeiry & Humfeld, 2021), modelling fluid mechanics (Raissi *et al.*, 2020), understanding permeability and viscoelastic modulus properties (Yin *et al.*, 2021) and to solve forward and inverse problems (Zhang *et al.*, 2019). Though demonstrating great potential, the PINN modelling approach has not been applied to food drying yet. For additional information on AI for the application of food processing, readers are directed to the comprehensive reviews of Sun *et al.* (2019), Khan *et al.* (2020) and Nayak *et al.* (2020). More insights into PINN modelling approaches can be found in Karniadakis *et al.* (2021).

---

## MACHINE LEARNING IN AGRICULTURE IN SRI LANKA

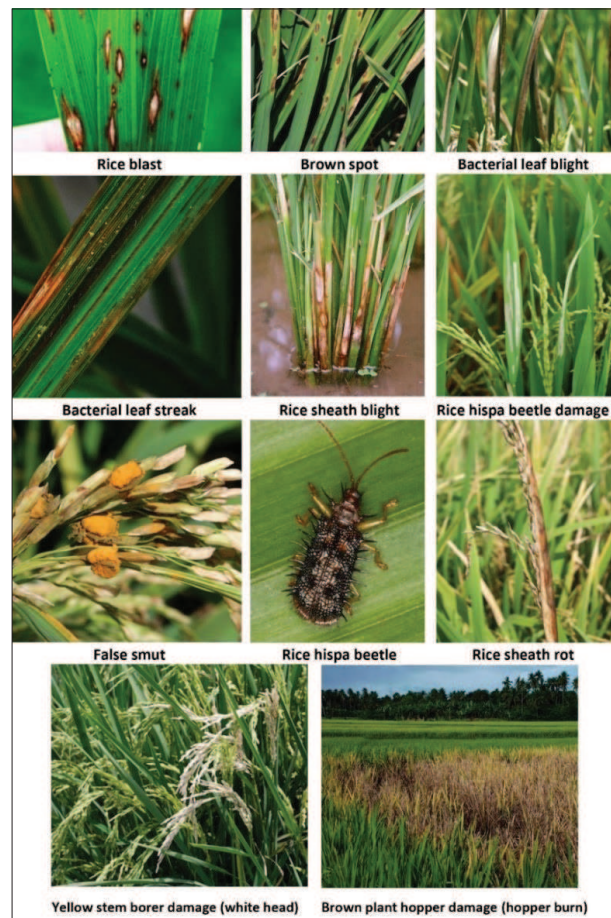
Sri Lanka is a tropical country with a high potential for cultivating and processing a variety of crops. Although the open-field agricultural system is the most prominent in Sri Lanka, the sector is increasingly enabled by the recent growth of IoT applications and AI systems in the field. Greenhouse agriculture is an emerging subsector today that attempts to maximize the harvests under a protected cultivation environment within the limited space. The recent government restrictions on chemical fertilizers have regulated the local agriculture industry to look for optimal fertilizer combinations that maximize yield in open-field systems and green-house environments. As the research of different permutated fertilizer combinations increases, numerous plant diseases and nutrient deficiencies receive increased attention from ML researchers.

Rice is one of the most popular food crops in the World and one of the main grains used in Asia including in Sri Lanka. Jaffna is located in the Northern tip of Sri Lanka at a longitude of  $79^{\circ} 45' - 80^{\circ} 20'$  and east latitude of  $9^{\circ} 30' - 9^{\circ} 50'$  with a population of around 700,000. The agriculture and fishery sectors play a crucial role in the gross production of Jaffna. Since the soil and climatic conditions are favourable to cultivate a wide range of crops including paddy, agriculture plays a significant role in the lives of people in Northern Sri Lanka. Various diseases in paddy agriculture have been seriously affecting rice production and constitute a big challenge for the agricultural community to ensure food security. These diseases include rice blast, bacterial leaf blight, bacterial leaf streak, sheath blight, seedling blight, false smut, rice hispa, sheath rot, root knot, leaf streak, yellow stem borer, brown spot, and brown planthopper (Figure 3) mainly caused by viruses, bacteria, fungi, etc. Diseases of rice plants could be affected by different factors, such as fertilizers, nutrients, water management, climatic conditions, lighting conditions, humidity, and farming conditions. The detection of such rice plant diseases is normally performed on visual assessment of the symptoms which is subjective, time-consuming even for well-experienced experts and is prone to error. Automating such visual assessment will provide information for the prevention and control of rice diseases through which the quality and the quantity of rice production can be increased by reducing the operation costs. Thus, the automation will contribute significantly to the economic growth of Sri Lanka.

### Explainable AI for plant disease classification

The diagnosis of plant nutrient deficiencies at early stages is essential as it indicates the impact of the applied fertilizer combination on plant growth. This deficiency could be a reason for overdose or lack of a particular nutrient or a combination. These deficiencies could be mainly seen in the plant's stems, leaves, or terminal buds (Figure 3). The association of these symptoms to different nutrient deficiencies requires expert knowledge. A data-centric approach is to use a classification model trained on labelled data. Image processing applications in the field of agriculture help to identify and diagnose crop diseases, pests and weeds, and the classification of agricultural products. CNN is one of the most effective methods of image processing which extracts image features, achieves high-level fusion of semantics and deep extraction of features through multilayer networks. Due to inherent learning capability, CNN is the prominently used machine learning model for

plant diseases classification. The convolution operation identifies the local features and through the processing in the subsequent layers of CNN, it hierarchically develops motifs and parts that could represent the symptoms of a particular disease. Despite the high classification accuracy, the CNN architecture is further explored to shed some light into the explanation or interpretation of its decision making. While techniques such as temperature scaling (Guo *et al.*, 2017) have been used to calibrate the CNN confidence scores, the explainable artificial intelligence models (XAI) have been hybridised with the CNN architectures (XAI-CNN) to provide explanations for such classifications. In general, these XAI models relate to or explain the class labelling by identifying the mostly contributed regions of the input image. The more commonly used XAI techniques are reviewed next emphasizing the main principle behind them.



**Figure 3:** Examples of common rice diseases (courtesy of <http://www.knowledgebank.irri.org/>)

Generally, XAI techniques that explain single disease predictions are of great interest in plant disease classification. CAM (Zhou *et al.*, 2016), LIME (Ribeiro *et al.*, 2016), and SHAP (Lundberg & Lee, 2017) are commonly used XAI models which provide justifications by generating local attention maps or by modelling feature importance. CAM (Class Activation Map) and its derivatives such as GRAD-CAM (Selvaraju *et al.*, 2017) and GRAD-CAM++ (Chattopadhyay *et al.*, 2018) are visual explanation techniques that can identify discriminative image areas by locating the pixel groups responsible for influencing the association between a given input image and a particular class label. These pixels are determined by projecting classification layer weights or classification layer ‘gradients of activation’ to the convolution feature maps. In contrast, LIME (Local Interpretable Model - Agnostic Explanations) attempts to develop linear (interpretable) models over the complex perceptron model by identifying significant features of a data sample. It modifies a single data sample by tweaking the feature values (or identifying super-pixels) to understand the relationship between input and the predicted output. By doing so, LIME provides a list of explanations reflecting the contribution of each feature in the prediction of the disease class of a given data sample. In SHAP (Shapley Additive Explanations), the feature values of a data sample act as players in a coalition. Shapley value is the average marginal contribution of a feature value. These values calculate the feature importance and explain the association of the data sample to a particular class label. When applying these XAI models, several challenges are recognized. In LIME, it is hard to model non-linear relationships with linear models even by tweaking features or identifying the super-pixels. The perturbations of a single data point with random noises would not cover the variations in the dataset. In SHAP, the Shapley values provide additive contributions of explanatory features, which might be misleading if the model is not additive. Further, it is not mathematically feasible to calculate the Shapley values for all the classification problems. Therefore, mainly CAM-based XAI techniques have been applied in plant disease classification (Liu *et al.*, 2020) (Ennadifi *et al.*, 2020) demonstrating significant improvements in the explainability of the class-label predictions.

When employing CNN together with XAI models, few challenges are recognised. First, publicly available datasets are small and often imbalanced leading to models getting overfitted even if the data are augmented with various augmentation techniques. Therefore, when applying XAI techniques on small and imbalanced datasets, CAM and GRAD-CAM approaches do not

perform well because it becomes difficult to locate the correct region of the convolution feature maps just by activating the classification layer of these models. This problem gets severe as the number of classes increases. Therefore, developing large balanced datasets is vital.

Second, a deficiency could be associated with multiple stages: early-stage, mild-stage, and severe stage, and symptoms of several deficiencies could be overlapped making the classification challenging. For example, there could be both Ca<sup>+</sup> and B deficiency; their symptoms could not be separable at their early stages but separable when they reach the mild stage. Therefore, applying CNN models at different stages of the deficiency could result in different predictions with a low confidence score. Using XAI models on poorly calibrated models further decays the explanations’ accuracy. Therefore, reasons based on feature activations or feature importance are insufficient for plant disease classification. It requires a much richer analysis of the features to map their activation to more probable two or three classes (including the predicted) based on the stages of the deficiency and growth stage of the plant.

Although existing XAI techniques give explanations in terms of feature activation and importance that can be interpreted by machine learning researchers they cannot fully explain the decision-making process in terms of agricultural technology and therefore of less use in real applications in disease recognition in agriculture. Therefore, these illustrations must map into human-understandable explanations for example by integrating the concepts of expert systems on the activated convolution features of XAI.

### **Rice disease detection using convolutional neural networks**

Due to the inherent ability of the CNN model to show layer by layer pictorial explanation of how an input image is processed to reach the classification outcome, it is a good candidate to use in detecting rice diseases even without hybridizing with XAI.

Jiang *et al.* (2020) proposed a rice disease recognition model that combines CNN and support vector machine (SVM) to identify four types of rice diseases. Images were pre-processed by applying mean-shift algorithm to segment lesions from rice leaf disease. From the images of leaf lesions shape features such as area, roundness and shape complexity were extracted, and the colour features were extracted using CNNs by converting RGB colour space into HSI and YCbCr colour space. The extracted



features were then classified by an SVM with RBF kernel due to the small number of sample images. They collected a total of 8911 images from rural farmland and from rice leaf disease atlas in which 6637 images of disease rice were manually cropped. A classification rate of 96.8% was reported.

Anandhan and Singh (2021) proposed a rice disease recognition model by comparing Mask R-CNN and Faster R-CNN algorithms to identify five types of rice diseases. Images were pre-processed using a multistage median filter to reduce noise. They implemented the model using ResNet-50 (Residual Neural Network) as the backbone and optimised the model using a stochastic gradient descent optimiser. In each iteration, a mini batch of size 60 was used. The learning rate was set to 0.01 with weight decay fixed to 0.0001. They collected about 1500 images from rice plants in the region under different weather conditions and images were manually annotated for diseases. A classification rate of 87.5% was reported for Mask R-CNN which outperforms Faster R-CNN.

Liang *et al.* (2019) proposed a rice blast recognition method using CNN. In their work, high-level features extracted by CNN were compared with handcrafted features: Harr-wavelet Transform and local binary patterns histograms (LBPH) in rice blast recognition. In addition to using CNN in the recognition task, the extracted features were also classified by SVM using an RBF kernel. Moreover, they created a rice blast disease image set that is made publicly available (<http://www.51agritech.com/zdataset.data.zip>). Images encompassing 2902 negative and 2906 positive samples were considered for testing and training a CNN. Quantitative analysis results show that CNN with Softmax and CNN with SVM have almost the same performance. A classification rate of 95.83% was reported for CNN showing better performance for the said recognition task.

Zhou *et al.* (2019) proposed an approach to identify three types of rice diseases based on FCM-KM and Faster R-CNN fusion techniques. Images were pre-processed using a weighted multilevel median filter for noise removal. Thereafter, Otsu threshold segmentation algorithm (Otsu, 1979) was applied to segment lesions from leaf disease images. In order to extract rice disease characteristics, the FCM-KM was used to reset the bounding box size in Faster R-CNN for the convergence to be accelerated. The optimal value of K in K-means was selected with the maximum and minimum distance algorithm and where those initial cluster centres should be positioned. In addition, the dynamic population firefly algorithm based on the chaos theory was applied to the clustering process to jump out of the local

optimum and obtain faster convergence. The FCM-KM and Faster R-CNN were applied to extract the disease characteristics and classify the images for pests and diseases. An image set consisting of 3010 images was tested and a classification rate of 97.2% was reported for the proposed method.

Sreevallabhadev (2020) proposed a method to identify rice blast disease by using CNN for feature extraction and SVM for classification. Experiments were carried out with 60 experimental configurations that vary in the choice of deep learning architecture, training mechanism, and dataset type. They used approximately 60,000 images of plant leaves in their original dataset (<https://www.kaggle.com/abdallahalidev/plantvillage-dataset>) and created three different versions of the dataset by considering colour, grayscale and leaf segmented images. An AlexNet based on CNN was used in extracting the features that are then fed to an SVM classifier. A classification rate of 96.8% was reported.

It is worth noting that, a majority of the work reported in the literature uses the following experimental setups and points out that there is a huge room for carrying out research in this domain using deep learning approaches.

- Evaluate on small scale datasets consisting of 500 to 5000 images beyond the work in (Sreevallabhadev, 2020). The small sized datasets directly impact the quality of the mapping function approximated by neural networks.
- Develop models to recognise a few rice diseases though there are several known diseases that greatly affect the quality and quantity of rice production.
- Propose methods mainly using features such as shape, texture and colour by the use of histogram-oriented gradient (HOG), scale-invariant feature transform (SIFT), wavelet, local binary pattern (LBP), HSI and YCbCr, CNN, etc.

---

## MACHINE LEARNING IN HEALTH

### Drug repositioning and drug interaction

Drug repositioning and drug interaction prediction are two fundamental applications of drug development and clinical care that have significant benefits in pharmacology. Repositioning of existing drugs can be classified as single drug-based repositioning and drug combination-based repositioning. Drug combinational treatments are identified to be much effective for treating some diseases. Moreover, drug-drug interactions (DDIs) are likely to occur when a pair of drugs or a combination



of drugs are co-administered. Investigating harmful DDIs is essential to enhance the effects of clinical care.

In these contexts, the features of existing drugs, such as chemical structures, gene expressions, target proteins, side effects, indications, etc., are considered to compare the pairwise drug similarity (Sun *et al.*, 2016; Hameed *et al.*, 2017; 2018) Further, chemical structural data, gene expression data, side effect related data and the transcriptional responses have been used for predicting drug mechanism of action (MoA), i.e., the prediction of molecular targets for a particular drug. ‘Connectivity Map’ (Lamb *et al.*, 2006) resource, which is frequently used in related studies, can be used to find connections among small molecules sharing an MoA, chemicals and physiological processes, and diseases and drugs.

Over the past two decades, machine learning and network analysis-based drug repositioning have gained popularity. Machine learning approaches incorporate clustering, classification and deep learning techniques with statistical concepts, whereas network analysis approaches represent pharmacological knowledge as networks with nodes and edges. In network-based analysis, features such as genes, proteins, molecules and phenotypes can be used as the nodes. Their functional similarities, mode of actions and relationships can be represented by the edges which interconnect the nodes.

In network-based analysis, techniques such as random walk (Zhang *et al.*, 2017), node embedding (Su *et al.*, 2021), matrix perturbation (Zhang *et al.*, 2017), and Steiner-tree based algorithms (Sun *et al.*, 2016) were used. They focus on building a network using drug features, actions, and characteristics to predict the novel drug-repositioning candidates via drug-drug, disease-disease, and drug-disease relationships. Zhu *et al.* (2020) proposed a drug-centric graph model, extracted and integrated six drug knowledge bases and constructed the drug knowledge graph. They have used a path-based data representation method and embedding-based data representation for comprehensive analysis for drug repositioning. Both methods applied to the drug knowledge graph evidenced better predictive performance on diabetes mellitus treatments. In matrix perturbation, novel drug-disease prediction can be transformed into a missing link prediction problem. In Sun *et al.* (2016), the prize-collecting Steiner Tree approach has proven to be a promising subnetwork identification method for drug repositioning. Their Physarum-inspired subnetwork identification algorithm employed on drug similarity networks has inferred useful repositioning candidates for cardiovascular diseases.

Positive Unlabelled Learning (PUL) is an emerging topic in the field of computational drug repositioning and drug interaction prediction. The application of PUL for single-drug repositioning (Mordelet & Vert, 2011; Yang *et al.*, 2012) and DDI prediction (Hameed *et al.*, 2017; Zheng *et al.*, 2019) has shown improved predictive performance. PUL enables prioritizing plausible negatives from the unlabelled data and improves performance compared to randomly selecting negatives from the unlabelled data. Mordelet and Vert (2011) introduced a scoring function and assigned a specific score to each data pair through which the data are sorted in descending order to distinguish positives and negatives. Similarly, Yang *et al.* (2012) have classified unlabelled data as reliable negatives, likely positives, likely negatives and weakly negatives. Hameed *et al.* (2017) also identified likely negatives from the unlabelled DDI pairs to treat as negatives when developing the binary classification models. These PUL approaches have shown significant improvements in final predictions.

Drug combination-based repositioning is emerging research in computational drug repositioning. Some recent studies focused on computational drug combination-based repositioning, exploring both therapeutic uses and adverse effects of drug combinations. Since there exist approximately 16,000 approved drugs on the market (Wishart *et al.*, 2018), millions of drug combinations can be formed. However, only a small number of drug combinations is confirmed as useful with experimental research. Therefore, there is a need for accurate and reliable approaches to infer repositioning candidates from those millions of unlabelled drug combinations. PUL can produce significant improvements in drug combination-based repositioning as well. Further, awareness of harmful DDIs is beneficial to extract the most suitable drug combinations for drug repositioning.

Imbalanced data is another challenge arising in the drug repositioning and drug interaction prediction domains. Mostly balanced training sets are preferred in machine learning approaches and in model training which can improve the generalizability of the trained model by preventing the model from being biased towards a particular class. Wei *et al.* (2013) discovered that regardless of the rates of positives and negatives in human genome data, support vector machines trained on balanced datasets have performed well in binary classification. Leevy *et al.* (2018) have also emphasized the issues that may arise with high-class imbalanced datasets and have presented the possible solutions that can be applied to avoid such potential issues. Further, Hameed *et al.* (2017) emphasized using balanced

datasets and employed multiple balanced training sets to strengthen the final prediction using ensemble learning which can reduce the variance of the final outputs.

### Use of interpretable models in microbial data including viral DNA

Microbiology is the study of the structure and functions of microorganisms and their interactions with other microorganisms, species, or environments (Moitinho-Silva *et al.*, 2017; DiMucci *et al.*, 2018; Xie *et al.*, 2019). Due to the advancements in microbial sequencing technologies, for instance, 16S rRNA sequencing, microbial studies generate a massive amount of data with a large number of samples and variables. Since most microorganisms are not present in most samples, the microbial data are also known to be highly sparse (Martino *et al.*, 2019), adding to its challenges. Therefore, ML has been employed as a powerful tool that can analyse and identify significant patterns in microbial communities (Ghannam & Techtmann, 2021).

ML algorithms can appear in many forms in microbial data analysis. Dimensionality reduction and visualization leading to better interpretation of data is one such significant manifestation. To better comprehend the underlying patterns in the high dimensional microbial data, it is necessary to reduce the number of dimensions either to visually interpretable two or three dimensions or to a reasonable number of dimensions that can be used as the input to another ML model. Principal component analysis (PCA) (Wold *et al.*, 1987) and principal coordinate analysis (PCoA) (Kruskal & Joseph, 1978) are two techniques commonly used for the task. However, they fail to capture the highly non-linear relationships in some microbiome data (Xu *et al.*, 2016). In contrast, manifold learning techniques such as Isomap (Tenenbaum *et al.*, 2000), Locally Linear Embedding (LLE) (Roweis & Saul, 2000), and t-Distributed Stochastic Non-linear Embedding (t-SNE) (Van der Maaten & Hinton, 2008) can overcome this limitation and therefore are commonly used in exploratory data visualization. t-SNE was also considered regularly for data visualization in the recent past on account of its capability in revealing the local structure in high-dimensional data, also bringing its attention to microbiome data (Kostic *et al.*, 2015). Moving forward, Xu *et al.* (2020) proposed a t-SNE based classification method for compositional microbiome data using Aitchison distances as the conditional probabilities. By this method, authors were able to achieve better classification performance compared with the classifiers built in the original high-dimensional space. However, t-SNE's ability in preserving the global structure of data

has been questioned. Therefore, a recently proposed technique, Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al.*, 2018) has gained popularity due to its capability in preserving the local structure while being superior in capturing the global structure compared to t-SNE. UMAP's usefulness in revealing composite patterns in microbiome data is demonstrated by Armstrong *et al.* (2021) with their application of the algorithm on three different microbial datasets. Autoencoders are artificial neural network-based algorithms that learn a compressed representation of the high dimensional input by minimizing its reconstruction error. Recently, autoencoders and their variations (e.g. variational autoencoders) have demonstrated their usefulness in learning meaningful latent features from data (Wang *et al.*, 2014), showing potential to be used with microbial data. Therefore, the research in this domain would benefit from the concept of interpretable neural networks when optimally identifying a suited architecture for a particular problem and mapping the process of identifying latent patterns to human-understandable explanations.

Metagenomics which is the analysis of DNA sequences of multiple species is effectively being used for studying microbes, especially viruses (Herath *et al.*, 2017). A metagenomic sample would consist of a large number of DNA sequence reads of multiple species making it inherently complex to study. Sample preparation including the DNA sequencing, reads assembly, annotation, and analysis of the data can be identified as the key steps in a metagenomic experiment and ML models are being used in all the mentioned steps (Krause *et al.*, 2020). Metagenomic data can be used to infer the microbes' interactions with their host environments, and thereby are considered in microbiome biomarker discovery for disease diagnosis and monitoring. However, due to their complex nature, using metagenomic data in precision medicine as a decision support system demands interpretable models with conciseness and readability by non-experts (Prifti *et al.*, 2020). Predomics, an ML approach inspired by microbial ecosystem interactions attempts to develop an interpretable and accurate model that can be used to analyse any type of data; especially microbiome data. Its effectiveness has been demonstrated on liver cirrhosis data (Prifti *et al.*, 2020). Furthermore, the effectiveness of an interpretable ML approach has been demonstrated by a meta-analysis of 1042 fecal metagenomic samples. It suggests the use of interpretable models to obtain non-intrusive predictive disease biomarkers for colorectal cancer using metagenomics data of the gut microbiome (Casimiro-Soriguer *et al.*, 2022). Furthermore,

another recent study suggests that easily accessible microbiome samples and their metagenomic analysis using interpretable models have the potential for non-invasive diagnosis of diseases (Carrieri *et al.*, 2021). The mentioned study demonstrates it by generating explanations for the predictions of skin hydration, age, menopausal status, and smoking habits based on leg skin microbiome (Carrieri *et al.*, 2021). As such, as the sequencing cost is becoming low, low and middle-income countries can benefit from interpretable models for metagenomics data analysis that can be used for disease surveillance and anti-viral therapy.

## DISCUSSION

Achieving interpretability in AI will promote broader and effective use of AI, by answering core questions about traceability, accountability, ethical compliance, and inherent biases. Several applications of AI reviewed in the paper provide different levels of interpretability from non to explainable AI. Interpretability can be further expanded to achieve linguistic interpretation (e.g. Halgamuge, 1997; Cao *et al.*, 2020).

AI models have narrow accessibility as they are generally manually designed by experienced AI experts, which limits access for fields in which AI experts have no knowledge or interest. Some automation in the design of AI is used in the drug repositioning and drug interaction prediction application described. Other AI methods, for example, evolutionary algorithms, can be used to design AI systems automatically (Cao *et al.*, 2020). There is vast untapped potential: for medical breakthroughs, distribution and use of scarce resources, forecasting with unheard-of accuracy in natural and economic domains, etc, when both interpretability and accessibility of AI systems can be combined as shown for small scale neural networks in Halgamuge (1997).

## Acknowledgement

Authors acknowledge the contribution of A/Prof Azharul Karim and Dr Zachary Welsh on Food Processing section and various grants including ARC DP220101035 and ARC DP210101135, University of Melbourne Scholarships of PhD students Nisal Ranasinghe and Tamasha Malepathirana.

## REFERENCES

Anandhan K. & Singh A.S. (2021). Detection of paddy crops diseases and early diagnosis using faster regional convolutional neural networks. *International Conference*

- on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 898–902.  
DOI: <https://doi.org/10.1109/ICACITE51222.2021.9404759>
- Angermueller C., Pärnamaa T., Parts L. & Stegle O. (2016). Deep learning for computational biology. *Molecular Systems Biology* **12**: 878.  
DOI: <https://doi.org/10.15252/msb.20156651>
- Armstrong G., Martino C., Rahman G., Gonzalez A., Vázquez-Baeza Y., Mishne G. & Knight R. (2021). Uniform manifold approximation and projection (UMAP) reveals composite patterns and resolves visualization artifacts in microbiome data. *Msystems* **6**: e00691.  
DOI: <https://doi.org/10.1128/mSystems.00691-21>
- Brunton S., Proctor J. & Kutz J. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **113**: 3932–3937.  
DOI: <https://doi.org/10.1073/pnas.1517384113>
- Cao B., Zhao J., Lv Z., Gu Y., Yang P. & Halgamuge S. (2020). Multiobjective evolution of fuzzy rough neural network via distributed parallelism for stock prediction. *IEEE Transactions on Fuzzy Systems* **28**: 939–952.  
DOI: <https://doi.org/10.1109/TFUZZ.2020.2972207>
- Carrieri A. *et al.* (19 authors) (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Scientific Reports* **11**: 1–18.  
DOI: <https://doi.org/10.1038/s41598-021-83922-6>
- Casimiro-Soriguer C., Loucera C., Peña-Chilet M. & Dopazo J. (2022). Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer. *Scientific Reports* **12**: 1–14.  
DOI: <https://doi.org/10.1038/s41598-021-04182-y>
- Çerçi K.N. & Daş M. (2019). Modeling of heat transfer coefficient in solar greenhouse type drying systems. *Sustainability* **11**: 5127.  
DOI: <https://doi.org/10.3390/su11185127>
- Chasiotis V.K., Tzempelikos D.A., Filios A.E. & Moustris K.P. (2020). Artificial neural network modelling of moisture content evolution for convective drying of cylindrical quince slices. *Computers and Electronics in Agriculture* **172**: 105074.  
DOI: <https://doi.org/10.1016/j.compag.2019.105074>
- Chattopadhyay A., Sarkar A., Howlader P. & Balasubramanian V.N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847.  
DOI: <https://doi.org/10.1109/WACV.2018.00097>
- Chen Y. & Martynenko A. (2013). Computer vision for real-time measurements of shrinkage and color changes in blueberry convective drying. *Drying Technology* **31**: 1114–1123.  
DOI: <https://doi.org/10.1080/07373937.2013.775587>
- DiMucci D., Kon M. & Segrè D. (2018). Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. *Msystems* **3**: e00181–18.  
DOI: <https://doi.org/10.1128/mSystems.00181-18>
- Ennadifi E., Laraba S., Vincke D., Mercatoris B. & Gosselin

- B. (2020). Wheat diseases classification and localization using convolutional neural networks and GradCAM visualization. *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1–5.  
DOI: <https://doi.org/10.1109/ISCV49265.2020.9204258>
- Ghannam R.B. & Techtmann S.M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal* **19**: 1092–1107.  
DOI: <https://doi.org/10.1016/j.csbj.2021.01.028>
- Guo C., Pleiss G., Sun Y. & Weinberger K.Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, pp. 1321–1330.
- Halgamuge S. (1997). Self evolving neural networks for rule based data processing. *IEEE Transactions on Signal Processing* **45**: 2766–2773.  
DOI: <https://doi.org/10.1109/78.650103>
- Halgamuge S. (2021). FAIR AI: A Conceptual Framework for Democratization of 21st Century AI. *International Conference on Instrumentation, Control, and Automation (ICA)*, pp. 1–3.  
DOI: <https://doi.org/10.1109/ICA52848.2021.9625672>
- Hameed P., Verspoor K., Kusljic S. & Halgamuge S. (2017). Positive-unlabeled learning for inferring drug interactions based on heterogeneous attributes. *BMC Bioinformatics* **18**: 1–15.  
DOI: <https://doi.org/10.1186/s12859-017-1546-7>
- Hameed P., Verspoor K., Kusljic S. & Halgamuge S. (2018). A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration. *BMC Bioinformatics* **19**: 1–18.  
DOI: <https://doi.org/10.1186/s12859-018-2123-4>
- Herath D., Jayasundara D., Ackland D., Saeed I., Tang S. & Halgamuge S. (2017). Assessing species diversity using Metavirome data: methods and challenges. *Computational and Structural Biotechnology Journal* **15**: 447–455.  
DOI: <https://doi.org/10.1016/j.csbj.2017.09.001>
- Jiang F., Lu Y., Chen Y., Cai D. & Li G. (2020). Image recognition of four rice leaf diseases based on deep learning and support vector machine. *Computers and Electronics in Agriculture* **179**: 105824  
DOI: <https://doi.org/10.1016/j.compag.2020.105824>
- Karim M.A. & Hawlader M.N. (2005). Mathematical modelling and experimental investigation of tropical fruits drying. *International Journal of Heat and Mass Transfer* **48**: 4914–4925.  
DOI: <https://doi.org/10.1016/j.ijheatmasstransfer.2005.04.035>
- Karniadakis G.E., Kevrekidis I.G., Lu L., Perdikaris P., Wang S. & Yang L. (2021). Physics-informed machine learning. *Nature Reviews Physics* **3**: 422–440.  
DOI: <https://doi.org/10.1038/s42254-021-00314-5>
- Khan M.I., Sablani S.S., Joardder M.U. & Karim M.A. (2020). Application of machine learning-based approach in food drying: Opportunities and challenges. *Drying Technology*: 1–17.  
DOI: <https://doi.org/10.1080/07373937.2020.1853152>
- Khan M.I., Welsh Z., Gu Y., Karim M.A. & Bhandari B. (2020). Modelling of simultaneous heat and mass transfer considering the spatial distribution of air velocity during intermittent microwave convective drying. *International Journal of Heat and Mass Transfer* **153**: 119668.  
DOI: <https://doi.org/10.1016/j.ijheatmasstransfer.2020.119668>
- Kostic A.D. et al. (22 authors) (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host and Microbe* **17**: 260–273.  
DOI: <https://doi.org/10.1016/j.chom.2015.01.001>
- Krause T., Andrade B., Afi H., Wang H., Zheng H. & Hemmje M. (2020). Understanding the role of (advanced) machine learning in metagenomic workflows. *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications* 56–82.  
DOI: [https://doi.org/10.1007/978-3-030-68007-7\\_4](https://doi.org/10.1007/978-3-030-68007-7_4)
- Kruskal & Joseph B. (1978). Multidimensional scaling. In: *Handbook of Data Visualization*. Sage, New York, USA.  
DOI: <https://doi.org/10.4135/9781412985130>
- Kumar C., Millar G.J. & Karim M.A. (2015). Effective diffusivity and evaporative cooling in convective drying of food material. *Drying Technology* **33**: 227–237.  
DOI: <https://doi.org/10.1080/07373937.2014.947512>
- Lamb J., Crawford E., Peck D., Modell J., Blat I., Wrobel M. & Golub T. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**: 1929–1935.  
DOI: <https://doi.org/10.1126/science.1132939>
- Leevy J., Khoshgoftar T., Bauder R. & Seliya N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data* **5**: 1–30.  
DOI: <https://doi.org/10.1186/s40537-018-0151-6>
- Li W., Bazant M.Z. & Zhu J. (2021). A physics-guided neural network framework for elastic plates: Comparison of governing equations-based and energy-based approaches. *Computer Methods in Applied Mechanics and Engineering* **383**: 113933.  
DOI: <https://doi.org/10.1016/j.cma.2021.113933>
- Liang W., Zhang H., Zhang G. & Cao H. (2019). Rice blast disease recognition using a deep convolutional neural network. *Scientific Reports* **9**: Article number 2869.  
DOI: <https://doi.org/10.1038/s41598-019-38966-0>
- Liu B., Ding Z., Tian L., He D., Li S. & Wang H. (2020). Grape leaf disease identification using improved deep convolutional neural networks. *Frontiers in Plant Science* **11**: 1082.  
DOI: <https://doi.org/10.3389/fpls.2020.01082>
- Long Z., Lu Y. & Dong B. (2019). PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics* **399**: 108925.  
DOI: <https://doi.org/10.1016/j.jcp.2019.108925>
- Lundberg S.M. & Lee S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30**: 4765–4774.
- Mariani V.C., De Lima A.G. & dos Santos Coelho L. (2008). Apparent thermal diffusivity estimation of the banana during drying using inverse method. *Journal of Food Engineering* **85**: 569–579.  
DOI: <https://doi.org/10.1016/j.jfoodeng.2007.08.018>



- Martino C., Morton J.T., Marotz C.A., Thompson L.R., Tripathi A., Knight R. & Zengler K. (2019). A novel sparse compositional technique reveals microbial perturbations. *MSystems* **4**: e00016–19.  
DOI: <https://doi.org/10.1128/mSystems.00016-19>
- Martius G. & Lampert C. (2016). Extrapolation and learning equations. *arXiv preprint arXiv:1610.02995*.
- McInnes L., Healy J. & Melville J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.  
DOI: <https://doi.org/10.21105/joss.00861>
- Moitinho-Silva L. *et al.* (11 authors) (2017). Predicting the HMA-LMA status in marine sponges by machine learning. *Frontiers in Microbiology* **8**: 752.  
DOI: <https://doi.org/10.3389/fmicb.2017.00752>
- Mordelet F. & Vert J. (2011). Prodiges: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics* **12**: 1–15.  
DOI: <https://doi.org/10.1186/1471-2105-12-389>
- Nayak J., Vakula K., Dinesh P., Naik B. & Pelusi D. (2020). Intelligent food processing: Journey from artificial neural network to deep learning. *Computer Science Review* **38**: 100297.  
DOI: <https://doi.org/10.1016/j.cosrev.2020.100297>
- Otsu N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetic* **9**(1): 62–66  
DOI: <https://doi.org/10.1109/TSMC.1979.4310076>
- Prifti E., Chevaleyre Y., Hanczar B., Belda E., Danchin A., Clement K. & Zucker J. (2020). Interpretable and accurate prediction models for metagenomics data. *GigaScience* **9**(3).  
DOI: <https://doi.org/10.1093/gigascience/giaa010>
- Raissi M., Yazdani A. & Karniadakis G.E. (2020). Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. *Science* **367**: 1026–1030.  
DOI: <https://doi.org/10.1126/science.aaw4741>
- Ribeiro M.T., Singh S. & Guestrin C. (2016). ‘Why should I trust you?’ Explaining the predictions of any classifier. *Proceedings of the 22<sup>nd</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.  
DOI: <https://doi.org/10.1145/2939672.2939778>
- Roweis S.T. & Saul L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**: 2323–2326.  
DOI: <https://doi.org/10.1126/science.290.5500.2323>
- Rudy S., Brunton S., Proctor J. & Kutz J. (2017). Data-driven discovery of partial differential equations. *Science Advances* **3**(4).  
DOI: <https://doi.org/10.1126/sciadv.1602614>
- Sablani S.S. & Rahman M.S. (2003). Using neural networks to predict thermal conductivity of food as a function of moisture content, temperature and apparent porosity. *Food Research International* **36**: 617–623.  
DOI: [https://doi.org/10.1016/S0963-9969\(03\)00012-7](https://doi.org/10.1016/S0963-9969(03)00012-7)
- Sahoo S., Lampert C. & Martius G. (2018). Learning equations for extrapolation and control. *International Conference on Machine Learning*, pp. 4442–4450.
- Saraceno A., Aversa M. & Curcio S. (2012). Advanced modeling of food convective drying: a comparison between artificial neural networks and hybrid approaches. *Food and Bioprocess Technology* **5**: 1694–1705.  
DOI: <https://doi.org/10.1007/s11947-010-0477-3>
- Sarkar T., Salauddin M., Hazra S.K. & Chakraborty R. (2020). Artificial neural network modelling approach of drying kinetics evolution for hot air oven, microwave, microwave convective and freeze dried pineapple. *SN Applied Sciences* **2**: 1–8.  
DOI: <https://doi.org/10.1007/s42452-020-03455-x>
- Scala K.D., Meschino G., Vega-Galvez A., Lemus-Mondaca R., Roura S. & Mascheroni R. (2013). An artificial neural network model for prediction of quality characteristics of apples during convective dehydration. *Food Science and Technology* **33**: 411–416.  
DOI: <https://doi.org/10.1590/S0101-20612013005000064>
- Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D. & Batra D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626.  
DOI: <https://doi.org/10.1109/ICCV.2017.74>
- Senanayake D., Wang W., Naik S. & Halgamuge S. (2021). Self-organizing nebulous growths for robust and incremental data visualization. *IEEE Transactions on Neural Networks and Learning Systems* **32**: 4588–4602.  
DOI: <https://doi.org/10.1109/TNNLS.2020.3023941>
- Sinha A. & Bhargav A. (2022). A simplified modelling approach for predicting shrinkage and sensitive material properties during low temperature air drying of porous food materials. *Journal of Food Engineering* **317**: 110732.  
DOI: <https://doi.org/10.1016/j.jfoodeng.2021.110732>
- Sreevallabhadev R. (2020). An improved machine learning algorithm for predicting blast disease in paddy crop. *Materials Today* **33**: 682–686.  
DOI: <https://doi.org/10.1016/j.matpr.2020.05.802>
- Su X., You Z., Wang L., Hu L., Wong L., Ji B. & Zhao B. (2021). SANE: a sequence combined attentive network embedding model for COVID-19 drug repositioning. *Applied Soft Computing* **111**: 107831.  
DOI: <https://doi.org/10.1016/j.asoc.2021.107831>
- Sun Q., Zhang M. & Mujumdar A.S. (2019). Recent developments of artificial intelligence in drying of fresh food: A review. *Critical Reviews in Food Science and Nutrition* **59**: 2258–2275.  
DOI: <https://doi.org/10.1080/10408398.2018.1446900>
- Sun Y., Hameed P., Verspoor K. & Halgamuge S. (2016). A physarum-inspired prize-collecting steiner tree approach to identify subnetworks for drug repositioning. *BMC Systems Biology* **10**: 25–38.  
DOI: <https://doi.org/10.1186/s12918-016-0371-3>
- Tenenbaum J.B., Silva V.D. & Langford J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290**: 2319–2323.  
DOI: <https://doi.org/10.1126/science.290.5500.2319>
- Udrescu S.M. & Tegmark M. (2020). AI Feynman: A physics-

- inspired method for symbolic regression. *Science Advances* **6** (16).  
DOI: <https://doi.org/10.1126/sciadv.aay2631>
- Udrescu S.-M., Tan A., Feng J., Neto O., Wu T. & Tegmark M. (2020). AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Advances in Neural Information Processing Systems* **33**: 4860–4871.
- Van der Maaten L. & Hinton G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**: 1–48.
- Vilarinho M.V., Franco C. & Quarrington C. (2017). Food loss and waste reduction as an integral part of a circular economy. *Frontiers in Environmental Science* **5**: 21.  
DOI: <https://doi.org/10.3389/fenvs.2017.00021>
- Wang W., Huang Y., Wang Y. & Wang L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 490–497.  
DOI: <https://doi.org/10.1109/CVPRW.2014.79>
- Wei Q., Dunbrack Jr & Roland L. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS ONE* **8**: e67863.  
DOI: <https://doi.org/10.1371/journal.pone.0067863>
- Welsh Z., Simpson M.J., Khan M.I. & Karim M.A. (2018). Multiscale modeling for food drying: State of the art. *Comprehensive Reviews in Food Science and Food Safety* **17**: 1293–1308.  
DOI: <https://doi.org/10.1111/1541-4337.12380>
- Wishart D. et al. (22 authors) (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**: D1074–D1082.  
DOI: <https://doi.org/10.1093/nar/gkx1037>
- Wold S., Esbensen K. & Geladi P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**: 37–52.  
DOI: [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9)
- Xie K., Guo L., Bai Y., Liu W., Yan J. & Bucher M. (2019). Microbiomics and plant health: an interdisciplinary and international workshop on the plant microbiome. *Molecular Plant* **12**: 1–3.  
DOI: <https://doi.org/10.1016/j.molp.2018.11.004>
- Xu W., Schultz T. & Xie R. (2016). An efficient visualisation method for exploring latent patterns in large microbiome expression data sets. *International Journal of Data Mining and Bioinformatics* **15**: 47–58.  
DOI: <https://doi.org/10.1504/IJDMB.2016.076016>
- Xu X., Xie Z., Yang Z., Li D. & Xu X. (2020). A t-SNE based classification approach to compositional microbiome data. *Frontiers in Genetics* **11**: 620143.  
DOI: <https://doi.org/10.3389/fgene.2020.620143>
- Yang P., Li X., Mei J., Kwok C. & Ng S. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics* **28**: 2640–2647.  
DOI: <https://doi.org/10.1093/bioinformatics/bts504>
- Yin M., Zheng X., Humphrey J.D. & Karniadakis G.E. (2021). Non-invasive inference of thrombus material properties with physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering* **375**: 113603.  
DOI: <https://doi.org/10.1016/j.cma.2020.113603>
- Zhang D., Lu L., Guo L. & Karniadakis G.E. (2019). Quantifying total uncertainty in physics-informed neural networks for solving forward and inverse stochastic problems. *Journal of Computational Physics* **397**: 108850.  
DOI: <https://doi.org/10.1016/j.jcp.2019.07.048>
- Zhang S. & Lin G. (2018). Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **474**(2217).  
DOI: <https://doi.org/10.1098/rspa.2018.0305>
- Zhang W., Chen Y., Liu F., Luo F., Tian G. & Li X. (2017). Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinformatics* **18**: 1–12.  
DOI: <https://doi.org/10.1186/s12859-016-1415-9>
- Zheng Y., Peng H., Zhang X., Zhao Z., Gao X. & Li J. (2019). DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions. *BMC Bioinformatics* **20**: 1–12.  
DOI: <https://doi.org/10.1186/s12859-019-3214-6>
- Zhou B., Khosla A., Lapedriza A., Oliva A. & Torralba A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929.  
DOI: <https://doi.org/10.1109/CVPR.2016.319>
- Zhou G., Zhang W., Chen A., He M. & Ma X. (2019). Rapid detection of rice disease based on FCM-KM and faster R-CNN fusion. *IEEE Access* **7**: 143190–143206.  
DOI: <https://doi.org/10.1109/ACCESS.2019.2943454>
- Zhu Y., Che C., Jin B., Zhang N., Su C. & Wang F. (2020). Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal* **26**: 2737–2750.  
DOI: <https://doi.org/10.1177/14604582200937101>
- Zielinska M., Zapotoczny P., Alves-Filho O., Eikevik T.M. & Blaszcak W. (2013). A multi-stage combined heat pump and microwave vacuum drying of green peas. *Journal of Food Engineering* **115**: 347–356.  
DOI: <https://doi.org/10.1016/j.jfoodeng.2012.10.047>
- Zobeiry N. & Humfeld K.D. (2021). A physics-informed machine learning approach for solving heat transfer equation in advanced manufacturing and engineering applications. *Engineering Applications of Artificial Intelligence* **101**: 104232.  
DOI: <https://doi.org/10.1016/j.engappai.2021.104232>