CS4182

Machine Learning

https://github.com/TTLawlor/CS4182Project

Authors: Tamara Lawlor, Luke O'Loughlin, Mary Tease, Seanie Lambe and Rachel O'Donoghue

Introduction – Mary Tease 19256434

Machine Learning is a subset of artificial intelligence. For a machine to be deemed 'intelligent' in an ever-changing world, it is very important for it to be able to learn, process data and adapt. While artificial intelligence covers the broad base of machines carrying out tasks on their own accord, machine learning focuses specifically on machines learning and adapting from data collected. To do this, machine learning consists of algorithms, statistics, and patterns. Algorithms are devised for machines and can be used to find and analyse patterns in different types of data. These findings can then be used to create a model if desired. Since a machine can very quickly sort through large amounts of data, a machine that has been given the correct algorithm can do any manner of sorting, analysing, and optimising faster and far more effectively than a human. Machine learning is commonly used in many areas such as by finance sectors for creating models to detect fraud and analyse stocks, in healthcare to help diagnosis and by the internet to provide users with relevant information.[1]

Supervised Learning – Rachel O'Donoghue 19274505

What is Supervised Learning?

Supervised learning is one of the most common types of Machine Learning. It "typically begins with an established set of data and an understanding of how that data is classified".[2] It uses labelled datasets to make predictions for new sets of unpredictable data collected later on using patterns and information previously found in old collections of data. Algorithms are shaped using pre-processed examples and are also assessed using test data. If the algorithm can identify this data successfully, then it has learned. If not, the proper adjustments must be made to the algorithm so that it can learn correctly. For example, you could give the algorithm some pictures of animals, each with their own description of what they are and their title. The algorithm, if working correctly, will sort through all of this new information and eventually be able to differentiate each different species and the different breeds of each animal using a picture.[2]

Sometimes, however, there can be patterns recognized in a subset of the given data. This may cause some problems, as it may not be identified in the larger group of data. This particular problem is called Overfitting, meaning the algorithm is only suited towards the 'test data' you have given it and not suited to larger amounts of unfamiliar data. To prevent this, one must calculate against unknown variables that may appear.[2]

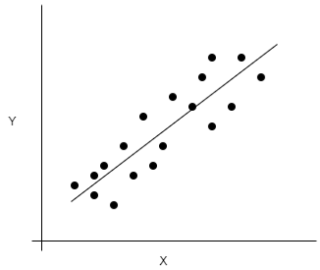
Types of Supervised learning

Supervised learning can be divided into two types: Regression and Classification. Regression usually involves numerical values, and is a continuous set of data, while Classification is derived from a limited set of values.

Regression "helps you understand the correlation between variables".[2] An example of regression analysis is weather forecasting. Regression is seemingly the more popular type of supervised learning used. It uses known weather patterns that have been collected throughout the years and uses it to present a prediction based on the current weather conditions.[2]

Classification algorithms output discrete, unordered values[3] meaning that their applicable use is extremely limited.

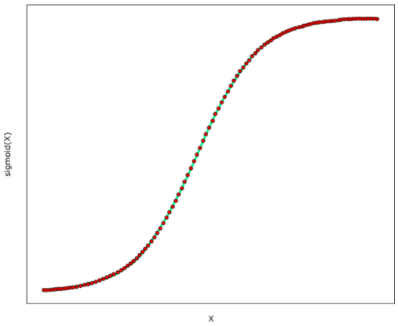
Regression Algorithms



One type of supervised learning algorithm is Linear Regression. This is an algorithm that presumes that there is a linear relationship between two variables: the independent variable and the dependent variable. Data is put through the algorithm and is then calculated using the function, $y = mx + c$.

A subset of Linear Regression is Clustered Linear Regression. This type of algorithm "makes some assumptions about the domain and the data set".[4] It improves Linear Regression by breaking down large amounts of data into subspaces. Because of this, it exceeds all other reputable machine-learning algorithms. However, it requires a lot of data which in turn can cause overfitting.

Another type of regression algorithm is Logistic Regression. It is used when the variable is able to be categorized. It uses the logit method that has been implemented to map data. It then predicts the probability of new data between a 0 and 1 probability scale, using the sigmoid formula.



$$\phi(z) = \frac{1}{1 + e^{-z}}$$

It gives the probability of many cases, such as pass/fail or, for a more practical example, healthy/sick. Logistic Regression is "one of the most widely used models in biomedicine."[5]

This model can be used to determine whether or not an image contains a certain object. Objects that have been detected in the image would be assigned a probability value between 0 and 1.

As shown above, the differences between the two forms of regression algorithms are that linear regression has infinite potential values, while logistic regression has exact outcomes.

Examples

There are many examples of Supervised learning in the world, speech recognition being one of them. This algorithm is trained how to recognize not only when a person is speaking, but what that person is saying, until it is able to understand a person's voice. Popular applications of speech recognition are Siri and Google Home. These applications both have a recognized keyword that activates them.

Another example of supervised learning is using it with fraud detection in telecommunications. This is when telecommunication services are accessed by users who have not paid for said service. A study shows that using both supervised and unsupervised learning showed positive results when being used to detect against a fraud detection problem. In this study of *"An application of supervised and unsupervised learning approaches to telecommunications and fraud detection" by Constantinos S. Hilas*[6]*,* it reads:

> From both analyses it is concluded that accumulated, in time, characteristics of a user yield better discrimination results. Aggregating user's behavior for periods larger than a week was avoided in order to preserve some level of on-line detection ability. Both approaches gave better results with Profile1.

This data was used to further evolve an application suited to recognize fraud in the area of telecommunications.

Unsupervised learning – Tamara Lawlor 19276494

Unsupervised learning is a specific type of machine learning that is suited to dealing with large amounts of unlabelled data. Larger datasets with multitudes of variables are often too difficult and time consuming to comprehend and categorize. Therefore, algorithms are used to interpret the structure and classify this data by attempting to identify different patterns or clusters without any human supervision. [7] Two of the main techniques used to uncover the patterns in a dataset are cluster analysis and principal components analysis.[8]

Cluster Analysis

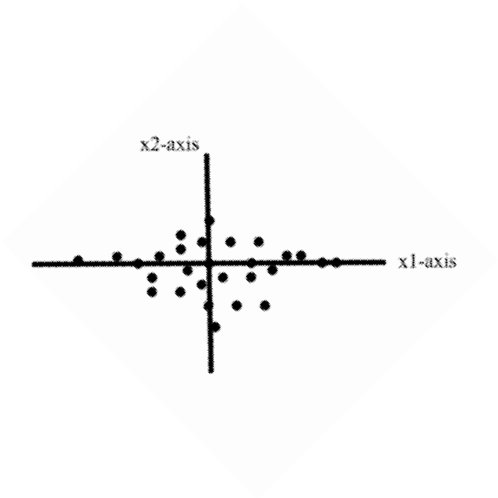
Cluster analysis takes multivariate sets and identifies characteristics that are common to groups of the inputted data. When the data is inputted the number of groups is unknown until the machine creates such groups in which the variables can be considered to possess characteristic patterns that are the same and as such the individuals in the group will be treated similarly. This method is useful as it can detect the relationships between data points and the patterns in the dataset allowing important information to be efficiently recognised and retrieved.[9]

K-means clustering is the most popular method that is used in clustering analysis to partition data. This method takes several random data points as the cluster centres, calculates the distance between the data points and their respective cluster centres using a loss function, and partitions or assigns the data to classes of individuals with similar characteristics, minimizing the dissimilarity between them and the centroid of their clusters. Loss functions, most commonly the Euclidean squared function, are used to calculate and measure the dissimilarity between individuals and their cluster centroids.

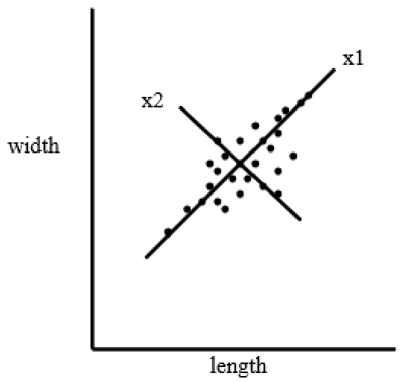
Take a dataset $x = (x1, x2, \dots, xn)$. The k-means clustering algorithm separates the data into K sets $S = \{S1, S2, \dots, Sk\}$ of individuals with similar characteristics. The objective is to find $T = \Sigma_{k=1}^{K} \Sigma_{x \epsilon S_k} L(x, C_k)$, where C_k is the cluster centroid and L is the Euclidean distance between the data points x ϵ S_k and C_k. The individuals of the dataset are delegated to the nearest C_k and T is calculated. The algorithm continues unless it falls below a certain determined threshold causing it to stop, as it continues calculating it re-assigns C_k as the mean of S, until there is no change between the current and previous centroids. As the execution of this algorithm is dependent on the number of clusters and centre point of the clusters, this algorithm must be repeated many times with alternating cluster centroids until there are no individual outliers outside a cluster.[10]

Principal Component Analysis

Principal components analysis (PCA) is the second main technique used in unsupervised learning, it is less commonly used than cluster analysis but serves the same purpose in allowing us to easily understand multivariate datasets. This method can reduce the dimensionality of the variables which makes analysing the dataset easier. Principal component analysis replaces the original s variables in the data set by a much smaller t. The premise of PCA is to be able to explain the variations between the variables of a multidimensional observation using $t(t < s)$.

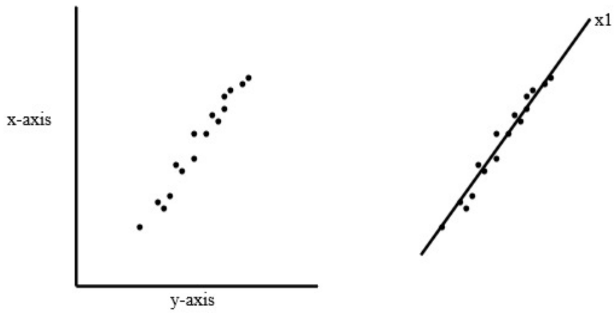


If we graph 2 characteristics of each individual e.g. length and width. You can observe that both characteristics or variables have in general around the same variance and are highly correlated with each other. We would then pass a vector, the first principal component $(x1)$, through the data points in the direction with the greatest number of variants and a second vector, also the second principal component $(x2)$, at a right angle to the original vector, maximising the variance in all directions, with their intersection marking the data points centroid.



Using the coordinates of the datapoints in relation to the two vectors, we could then re-plot the data on a new graph. The data on the graph is merely rotated and as such the spatial relationships of each data point has not been altered. From looking at the graph one can note that the variance of the data is greater along the $x1 - axis$ than that of the $x2 - axis$ both of which axis are, unlike the graph above, uncorrelated to each other. The individuals can be explained using these new variables e.g. $x1$ could represent size and $x2$ the shape. This method allows one to quickly detect the relationships between data points.

In cases where the variance of the axis is very small, the axis can be ignored, which will reduce the dimensionality of the dataset as demonstrated in the following graph.[11] [12]



Machine Learning Bias and Ethical Issues – Luke O'Loughlin 19231326

Biased Data and Creators

All predications made by algorithms are done using the data it is given. This raises the question; what happens when the algorithm is fed biased data? It will do what is most logical. It will find the pattern in the bias and continue it unless corrected. The algorithm cannot differentiate between biased and neutral data. An algorithm doesn't understand the concept of being biased. It can be presumed (not guaranteed) that the majority of biased data given to algorithms is done so unintentionally. An investigation that looked at in Yapo, A. and Weiss, J. W. (2018)[13] showed that an algorithm designed by Northpointe (a for-profit company)

> is twice as likely to mistakenly identify white defendants as a low risk for committing future crimes, and twice as likely to erroneously tag black defendants as a high risk.

This algorithm was designed to not consider ethnicity or race, as one may expect, but it still managed to discriminate. How this happened is not completely known but it can be largely assumed that the data given to the algorithm was unintentionally biased. Even so, enough independent patterns were found linking black people together, causing the algorithm to become biased against them. This is a huge ethical problem. This is why, especially in the field of justice, algorithms must be able to explain and justify their decisions. However, this does also affect the transparency of the system, as looked at later.

There is also the issue of bias in the creators of the algorithms.[14] All humans are inherently biased by the things that have shaped them in their lives. Those creating the algorithms are not above this. Whether consciously or not, a designer can create an algorithm with a slight bias against anyone or anything. With no regulations to stop them creating biased algorithms, it becomes all the more important that transparency and explainability of decisions made by the algorithm become a crucial factor when developing them. Otherwise there can be no justification for poor, unethical decisions.

Lack of regulation and transparency

Though AI development may feel like a very modern endeavour, it has been underway since 1956. However, in recent times AI development is growing at an exponential rate.[9] And so, it isn't surprising to discover that there are no laws in place directly concerning AI. The problem is that AI should be exceptional in this regard. So much of our modern world is reliant on algorithms that it should not be the case that their creation is completely unchecked and unregulated. These lack of regulations could have serious consequences. The lack of accountability of companies, the funnelling of data to a few select places and the ability do whatever they like is terrifying. The only thing currently stopping companies creating biased algorithms is the bad PR they would face as a consequence if found out.

One of the biggest problems in relation to the lack of forced regulations on said organisations is their lack of transparency of how their systems work. This is why these algorithms are so often referred to as "black boxes"[10]. If companies open up and reveal how their algorithm works, they fear their work being stolen and that people will figure out how to exploit the system. However, the issue with keeping the "black box" is that people are given extremely hazy guidelines on what the algorithm wants, leaving them to blindly guess to figure out what the system wants.

Solutions

- Transparency and justification: Something I have mentioned that is severely lacking is transparency into how these systems actually work and the explainability of their decisions. With this being made clearer, people would suffer far less from anxiety induced by the uncertainty and stop frustration with the system. It would also allow companies to be held accountable for poor decisions. This may also sway public opinion on the idea of machine learning algorithms as they would be able to understand them, something that, up until this, has not been done.

- Algorithms creating biased predications is a serious problem with major ethical issues. One way to tackle this is through Algorithmic Impact Assessments.[15] Canada has implemented this in the form of a questionnaire that allow the engineers of these algorithms to take into account for many more issues than they may have otherwise thought of thanks to the public. It also gives the public much more power as to how these systems will operate.[16]

- The EU published a paper which sets out ways in which research and development can be improved between European countries as well as a framework for policies to regulate AI development going forward.[17]

- This proposal will only affect the EU. That's why I feel that the setting up of self-regulating bodies is the best option. They will be able to make changes quickly and decisively to AI regulation and, provided they are abided and not corrupted, would allow AI to flourish without over correction from government laws. This has already happened with the setting up of *The Partnership on Artificial Intelligence to Benefit People and Society* and *The Ethics and Governance of Artificial Intelligence Fund.*[9]

Agriculture and Machine Learning – Mary Tease 19256434

Agriculture is one of the most vital industries in the world. The demands and pressures faced by farms of all scales is only increasing due to the ever-growing world population. This pressure requires farms to be able to maximise their output and efficiency, which is where machine learning can be extremely useful. It can be applied to many different areas of agriculture such as crops[18] and predicted yield[19], livestock and the breeding of dairy cows.[20]

Use of Machine Learning in breeding of dairy cows

Deciding whether to breed an animal or not is a big decision on any farm, but it is an even heavier decision on a large-scale dairy farm. Knowing if it is worthwhile to go through with insemination of a cow is extremely useful information for a farmer. A study from the Journal of Dairy Science Volume 97, Issue 2, February 2014, took 10 years' worth of data from 26 farms via Alta Genetics, Watertown, Wisconsin. This study implemented many forms of machine learning algorithms to predict the reproductive performance of Holstein cows. No one type of algorithm was best suited to collect the required data, so the study used the most common types of algorithms including decision trees, Naïve Bayes algorithms and bagging[16]. This study analysed many different health traits of each cow, the most important traits proved to be lameness, ketosis and retained placenta. The data was compiled into tables which revealed mean statistics and traits. The study was able to provide information as to which health traits in cows lead to them being highly fertile, which is a great benefit to farmers.

Use of Machine Learning in crop yield

There have been numerous studies of how machine learning can assist the important topic of yield mapping. Knowing what the approximate yield of a certain crop is, is vital information for a farmer. A study from the *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing* used Artificial Neural Networks, which are machine learning models that are adept at learning, adapting and reading the more complex data found in studies like this, to create a model grass biomass prediction. This study was done here in Ireland, where grassland makes up over 80% of all agricultural land[15], and since very few similar studies on grass had taken place before, the results proved very useful.

Further afield, machine learning has been used to help automatically count coffee fruits on branches. This study[21] used a machine vision system (MVS), which included an image processing algorithm. The system was shown fruit branches with variables like the number of fruits on it, different stages of the fruits development and fruits of different harvest dates. With this data, three sets of models were created for the categories: harvestable, un-harvestable and "*fruits whose maturation stage were disregarded".* The data gained in this study proved to be extremely useful and accurate, with the authors finding;

"that MVS neither overestimates nor underestimates the number of fruits and that it shows a correlation higher than 0.90"

The MVS created in this study allowed for the estimation of fruit numbers, and wound up being able to estimate their weight and how matured each fruit was. The MVS was also in use in four separate Variedad Castillo® coffee plots and is said to be an efficient and non-harmful way to provide very useful information to coffee growers. [21]

Healthcare and Machine Learning – Seanie Lambe 19264267

Machine learning has the potential to revolutionise the medical industry at rates unseen before. The availability of computational techniques that can handle large complex datasets has opened the floodgates for what is possible in the healthcare sphere, and the explosive growth of health-related data has presented unprecedented opportunities for improving health of a patient[22]

Big Data

Big data refers to the voluminous and complex amount of data collected from sources like web enterprise applications, mobile devices and digital repositories which cannot be easily managed using traditional tools.[23] Big data enables an organization to store and manage large amount of data at high speed to get the right insight from data and streamline certain aspects of the healthcare industry, making procedures more efficient and cost effective. Rather than coding specific sets of instructions to accomplish a task, the machine is 'trained' using algorithms, allowing it to learn how to perform specific tasks. Machine learning thrives in a repetitive environment, learning intricate patterns from even high-dimensional raw data with little guidance and continuously improving. Medicine seems particularly amenable to machine learning.[24]

Impact in Healthcare

Machine learning aids clinicians in delivering care to patients in the United Kingdom. Using large digital databases of Electronic Health Records, machine learning has enabled data-driven prediction of drug effects and interactions, identification of type 2 diabetes subgroups, and discovery of comorbidity clusters in autism spectrum disorders. Similarly, In the United States, the IBM Watson Health cognitive computing system (IBM Corp., Armonk, New York) has used machine learning to develop techniques to improve diagnostic accuracy and reduce costs put on cancer patients using large amounts of patient cases and over one million scholarly articles.[25]

One sector of healthcare that benefits greatly from machine learning is Epidemiology, the study and analysing of the distribution, patterns and determinants of health and disease conditions in defined populations The use of supervised machine learning and neural networks will assist in in both identifying risk factors for disease and implementing preventive healthcare, meaning the actions taken to prevent diseases from spreading. Epidemiologists must process and interpret large amounts of complex data and using said data can lead to an improved understanding of risk factors for development of healthcare-associated infections allowing for targeted prevention approaches. The applications of ML in infectious disease are diverse and include risk stratification for specific infections, identifying the relative contribution of specific risk factors to overall risk, understanding pathogen–host interactions, and predicting the emergence and spread of infectious diseases.[18]

Deep Learning

Deep learning has achieved remarkable success in various artificial intelligence research areas, including healthcare. Using deep learning, machines can learn from large amounts of data that is unstructured or unlabelled, creating a very deep and complex neural network. This technology has shown superior performance to other machine learning algorithms in areas such as image and voice recognition. However most interestingly, it has aided in drug discovery. Deep learning has a much more flexible architecture, so it is possible to create a neural network architecture tailor-made for a specific problem. Reaction predictions, a crucial element of drug discovery, have been reported to be enhanced by deep learning technologies with performance being on-par with, or superior to, the rulebased methods, speeding up the process of drug discovery, an ever more important aspect of the healthcare industry given our current climate.[26]

The Negatives

The late Prof. Stephen Hawking in 2016 declared AI to be "either the best, or the worst thing, ever to happen to humanity".[27] As it makes things more efficient and streamlined in the healthcare industry, it may come at the expense of jobs. Researchers at Stanford University have created an AI algorithm that can identify skin cancer as well as a professional doctor. The program was trained on nearly 130,000 images of moles, rashes, and lesions using machine learning. It was then tested head-to-head against 21 human dermatologists, where it performed with an accuracy on par with humans ("at least" 91 percent as good). In the future, they suggest it could be used to create a mobile app for spotting skin cancer at home. As this technology improves at rapid pace, dermatologists may become redundant in due time.[28]

BIBLIOGRAPHY

---

[1] Alpaydin, E. (2004) *Introduction to machine learning.* Cambridge, Mass.; MIT Press. Available here.

[2] Mueller, J., & Massaron, L. (2016). *Machine learning for dummies.* Hoboken, New Jersey: John Wiley & Sons, Inc.

[3] Maglogiannis, I., Karpouzis, K., Wallace, B.A., Soldatos, J. (2007). *emerging artificial intelligence applications in computer engineering: real word ai systems with applications in eHealth, HCI, information retrieval and pervasive technologies.* IOS Press, Amsterdam, Netherlands.

[4] Betran A. and H. Altay, G.(2002). *Clustered linear regression.* Knowledge-Based Systems 15, 169-175, available: https://doi.org/10.1016/S0950-7051(01)00154-X

[5] Dreiseitl, S. and Ohno-Machado, L. (2002). *Logistic regression and artificial neural network classification models: a methodology review.* Journal of Biomedical Informatics 35, 352–359, available: https://doi.org/10.1016/S1532-0464(03)00034-0

[6] Hilas, C. S. and Mastorocostas, P. A. (2008). *An application of supervised and unsupervised learning approaches totelecommunications fraud detection.* Knowledge-Based Systems 21 , 721–726, available: https://doi.org/10.1016/j.knosys.2008.03.026

[7] Mueller, J., & Massaron, L. (2016). *Machine Learning for Dummies.* Hoboken, New Jersey: John Wiley & Sons, Inc. Available here.

[8] Upton, Graham, and Ian Cook. "Unsupervised Learning." *A Dictionary of Statistics* 1 Jan. 2014. Web. Available here.

[9] Aldenderfer, M. S. & Blashfield, R. K. (1985) *Cluster analysis.* London: Sage Publications. Available here.

[10] Narges Ahmadzadehgoli et al. (2018) LINEX K-Means: Clustering by an Asymmetric Dissimilarity Measure. *Journal of Statistical Theory and Applications (JSTA).* [Online] 17 (1), [online]. Available from: https://doaj.org/article/2c04bd19d38b49c881fbfa12c14cddb5.

[11] Upton, Graham, and Ian Cook. "principal components analysis (PCA)." *A Dictionary of Statistics* 1 Jan. 2014. Web. Available here.

[12] Holland, Steven M. "Principal components analysis (PCA)." *Department of Geology, University of Georgia, Athens, GA* (2008): 30602-2501. Available here.

[13] Yapo, A. and Weiss, J. W. (2018). 'Ethical implications of bias in machine learning', *Hawaii International Conference on System Sciences*, Hawaii, 03 Jan, available: http://hdl.handle.net/10125/50557

[14] The World Economic Forum's Global Future Council on Human Rights (2018). *How to Prevent Discriminatory Outcomes in Machine Learning*[White paper], from The World Economic Forum: http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

[15] Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*, AI Now Institute, available: https://ainowinstitute.org/aiareport2018.pdf

[16] *Algorithmic Impact Assessment v0.8* (n.d.) Government of Canada, available https://canada-ca.github.io/aia-eia-js/

[17] European Commission (2020). *White paper on artificial intelligence: a European approach to excellence and trust*[White paper], from the European Commission: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[18] Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. *Machine Learning in Agriculture: A Review. Sensors 2018*, *18*, 2674. Available: https://www.mdpi.com/1424-8220/18/8/2674/htm#

[19] Ali, F. Cawkwell, E. Dwyer and S. Green, *"Modeling Managed Grassland Biomass Estimation by Using Multitemporal Remote Sensing Data—A Machine Learning Approach,"* in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 7, pp. 3254-3264, July 2017*, doi: 10.1109/JSTARS.2016.2561618. Available: https://cora.ucc.ie/bitstream/handle/10468/6463/1430.pdf?sequence=1

[20] Saleh Shahinfar, David Page, Jerry Guenther, Victor Cabrera, Paul Fricke, Kent Weigel, *Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms, Journal of Dairy Science, Volume 97, Issue 2, 2014*, Pages 731-742, ISSN 0022-0302, https://doi.org/10.3168/jds.2013-6693. (http://www.sciencedirect.com/science/article/pii/S0022030213008059)

[21] P.J. Ramos, F.A. Prieto, E.C. Montoya, C.E. Oliveros, *Automatic fruit count on coffee branches using computer vision, Computers and Electronics in Agriculture, Volume 137, 2017*, Pages 9-22, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2017.03.010. (http://www.sciencedirect.com/science/article/pii/S016816991630922X)

[22] Khare, A. et al. (2017) Machine Learning Theory and Applications for Healthcare. Journal of Healthcare Engineering. [Online] 20172.

[23] Wiens, J. & Shenoy, E. S. (2018) Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clinical Infectious Diseases. [Online] 66 (1), 149–153.

[24] Jones, L. D. et al. (2018) Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? Bone & joint research. [Online] 7 (3), 223–225.

[25] Wang Z, Shah AD, Tate AR, et al. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One 2012;7:e30412.

[26] Applications of Deep Learning in Biomedicine Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov Molecular Pharmaceutics. Available here. 2016, 13(5),1445-1454

[27] Hern A. Stephen Hawking: AI will be 'either best or worst thing' for humanity. https://www.theguardian.com/science/2016/oct/19/stephen-hawking-ai-best-or-worst-thing-for-humanity-cambridge(date last accessed 20 December 2018).

[28] Esteva A, Kuprel B, Novoa R A, Ko J, Swetter S M, Blau H M and Thrun S 2017 Dermatologist-level classification of skin cancer with deep neural networks Nature 542 115–8