**Machine Learning Algorithms for Predicting Diagnosis and Dysplasia Status**

Alexis Delabre   
M1 MIAGE, Université Toulouse 1 Capitole, Toulouse, France  
Alexis.Delabre@ut-capitole.fr

Yassine Yazid   
M1 MIAGE, Université Toulouse 1 Capitole, Toulouse, France  
Yassine.Yazid@ut-capitole.fr

Bruno Santome Antolin   
M1 MIAGE, Université Toulouse 1 Capitole, Toulouse, France  
Bruno.Santome-Antolin@ut-capitole.fr

Youness Haraki   
M1 MIAGE, Université Toulouse 1 Capitole, Toulouse, France  
Youness.Haraki@ut-capitole.fr

Burhan Ariyanto   
M1 MIAGE, Université Toulouse 1 Capitole, Toulouse, France  
Burhan.Ariyanto@ut-capitole.fr

ABSTRACT

A term of dysplasia is used to describe when there is a presence of abnormal cells within a tissue or organ in human bodies. Before cancer cells form in tissues of the body, the cells go through abnormal changes : first is hyperplasia and then become dysplasia. In hyperplasia, under a microscope, there is an increase in the number of cells in an organ while the cells look abnormal in dysplasia. They are both not a cancer but in a worst case, they can become cancer. The development of digital pathology and progression of state‐of‐the‐art algorithms for computer vision have led to increasing interest in the use of artificial intelligence (AI), especially machine learning (ML)‐based AI, in prediction of diagnosis because it is a huge field which learns from past experiences and gives proper predictions. In this study, we want to predict both the diagnosis (multi-categorical prediction) and the dysplasia status (binary prediction). We will discuss 3 (three) algorithms like K-Nearest Neighbors, Decision Tree and Artificial Neural Network algorithms using the data set that composed of 2362 patients from the Centre Hopitalier Universitaire (CHU) of Toulouse. The accuracy of prediction of the mentioned algorithms will be discussed in the research study.

KEYWORDS

Dysplasia, Diagnosis prediction, Patient characteristics, Classifications, Machine learning, K-NN, Decision tree, Artificial neural networks.

# **INTRODUCTION**

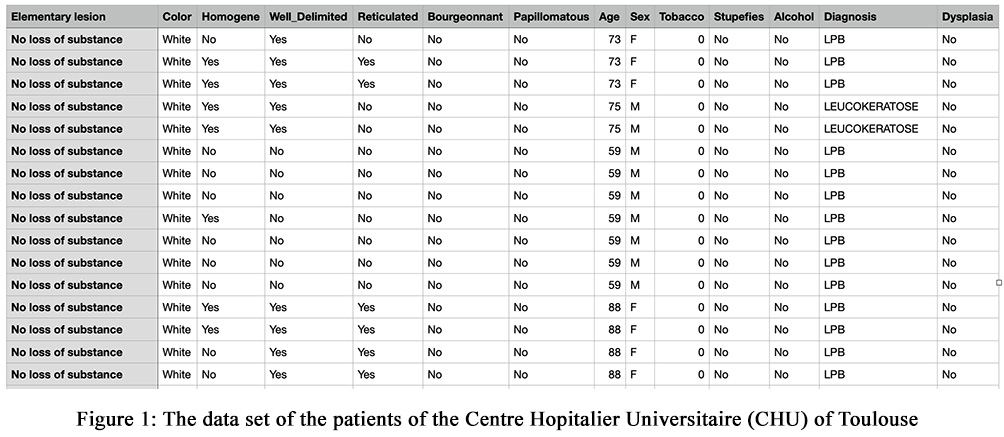
In medical practice, it is common for sick people to go to a doctor to be examined and get a diagnosis. If the doctor provides a diagnosis or recommends treatment, then the patient will need further explanation. When the doctor only answers, "I simply know" then it would be difficult for the patient to accept / digest the answer. Very likely, the patient will leave the doctor and go to another doctor who is more skilled and who can explain with certainty. Of course it would be a different case if the doctor answered the patient by saying, "I know because the computer told me." The patient can surely ask further question, "How did the computer find out?" Then the doctor will be able to explain further because he has access to model's output.

Artificial Intelligence is something that is often discussed nowadays because this kind of intelligence includes approaches and techniques such as machine learning, machine reasoning and robotics. In this paper, our main concern is machine learning because it is an approach that is applied using different techniques and algorithms in a variety of activities, including health care. The use of machine learning to solve clinical problems is called revolutionary clinical decision making, which implies that the system will look at a particular individual by collecting and interpreting data relevant to that individual's health. The result will suggest the best course of action that needs to be taken to maintain the individual's health. Usually, these algorithms are not very strong and concrete at first, but by performing repetitive tasks, the algorithms become stronger by learning from past experiences.

Machine learning approaches are used in computing decision-making in health care, where critical data analysis needs to be performed on medical data to reveal abnormalities that are not visible to humans. Implementing an algorithm for abnormality / disease detection is a difficult task, but what makes it more challenging is improving the algorithm's accuracy while reducing the time it takes to execute the algorithm. In the health and medical industry, the accuracy of predicting a disease is very important and requires an effective decision in taking an analysis and the accuracy of predicting a disease in a patient. At present, there are many diseases that need to be identified at an early stage in order to initiate relevant treatment. Otherwise, they are incurable and deadly. Therefore, complex analysis of medical data, medical reports, and medical images is required in a shorter time but with greater accuracy. An accurate prediction will lead to good prognosis (a forecast of the likely course of a disease or ailment) and good care. Also when discussing computational decision making in the health care sector, it is not always about detecting or predicting disease, analyzing biomedical images, etc., but also about how to conduct medical care research, patient care, allocate resources, manage hospital volumes, public health policy making, and much more. Thus, by conducting algorithms that can provide high accuracy in prediction, can save many lives in medical field. It is inevitable that the development of this machine learning system will continue to be endeavored to support progress in the field of healthcare.

# **STATE OF ART**

Computer science for medical diagnosis based on machine learning are not just science fiction. So many scientists are using and developing machine learning methods to support advances in the health sector. One really hot example is the prediction of a diagnosis of Covid-19 based on symptoms which combines several features to estimate how much risk of the infection (Zoabi, Deri-Rozov and Shomron, Jan. 2021). These researchers succeeded in predicting the results of the COVID-19 test with high accuracy. They used eight binary features: gender, age ≥ 60 years, known contact with an infected individual, and appearance of five baseline clinical symptoms. Another successful example is the application of machine learning in cancer prognosis and prediction (Kourou, Exarchos, Karamouzis and Fotiadis, 2015, p. 8- 7). Another example of the application of machine learning is the prediction of patient prognosis in psychiatry using neuroimaging and machine learning is being investigated (Arbabshirani et al., 2017; Bzdok & Meyer - Lindenberg, 2018; Janssen, Mourao - Miranda, & Schnack, 2018). In addition, machine learning algorithm for predicting early MRI-based Alzheimer's disease is being explored (Moradi et al., 2015).

Machine learning algorithms based on linear regression, including regulated approaches such as LASSO, Ridge Regression, or Linear Support Vector Machines are most reliable, because their decision value is a linear combination of these features. The decision tree (DT) is also considered to be at the same level as the above algorithms, along the accuracy trade-off. Although DT allows nonlinear interactions between features via a separation sequence, this feature is still well interpreted by humans even though inspection can be a challenge for deeper trees. On the other hand, the K-nearest neighbor (K-NN) algorithm is an example of instance-based learning or "lazy" learning, where the input is compared to the training data set and receives the most frequent label among the training observations it most closely resembles. While this trend outperforms the previous algorithms mentioned above, the fact of the "similarities" in high-dimensional spaces are difficult to explaine from a human perspective.

The same trend is found in kernel regression, which can be conceptualized as a weighted form from the previous approach, which each observation in training contributes to the prediction according to its similarity to the test sample. In this case, most of the interpretability is usually lower than the nearest neighbor approach. The ensemble-based approach, which repeatedly adjusts simpler models (such as DT and linear models) combines the results into a final prediction has enjoyed its popularity and success in recent years, as this trend shows excellent accuracy even in a medium-sized training set. Finally, artificial neural networks (ANNs) have demonstrated their ability to provide the best accuracy in many fields, but also represent the class of tools, which have the lowest transparency, because decisions are made based on nonlinear interactions involving millions of parameters.

# **THE DATA SET**

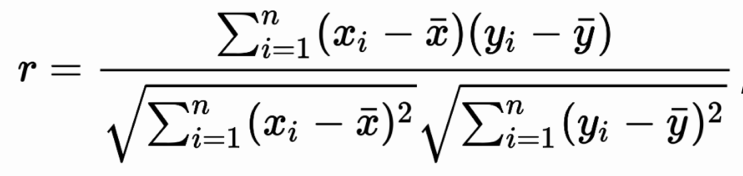
Machine Learning (ML) depends heavily on data. This is the most important aspect that makes algorithm training possible and explains why machine learning has become so popular in recent years. That is why data preparation is such an important step in the machine learning process. In short, data preparation is a set of procedures that help make your data set more suitable for machine learning.

In this project, we are trying to predict the accuracy from the data set. The dataset we are trying to process contains data of patients from Centre Hopitalier Universitaire (CHU) of Toulouse (Figure 1). It contains 2362 patients, where the 116 patients have dysplasia status (yes/no status) and also multi-categories diagnosis. Knowing what to predict will help in deciding which data might be more valuable to collect. When formulating a problem, it is inseparable from data exploration and categories of *classification, clustering and regression.*

In terms of *classification*, there is a need for data sets to be classified. You want an algorithm to answer yes-or-no binary questions (bird or fish, healthy or not, car or bike, etc.) or you want to create a multiclass classification (cancer, asthma, or stroke; cat, dog or bird, etc.). You also need the correct answers labeled, so the algorithm can learn from them. In *clustering*, we want an algorithm to find the classification rule and the number of classes. The main difference from classification tasks is that we do not really know what groups are and the principle of division. in this case, for example, we need to segment the patient and adapt a specific approach to each segment depending on the quality. In *regression*, we want the algorithm to return multiple numeric values. For example, if you spend too much time trying to find an accurate diagnosis because it depends on many factors, a regression algorithm can help estimate this value.

Each feature, or column, in a data set represents a measurable piece of data that can be used for analysis: Name, Age, Gender, Status, and so on. Features are also sometimes referred to as "variables" or "attributes." The quality of the features in the data set has a major impact on the quality of the insights you will gain when you use that dataset for machine learning.

We selected 3 (three) algorithms to make predictions on the status of diagnosis and dysplasia, and for the features we chose to make these predictions differ from one algorithm to another. For the K-Nearest Neighbors (K-NN) algorithm, we used all the features data: Lession, Color, Homogene, Well Delimited, Reticulated, Bourgeonnant, Papillomatus, Age, Sex, Tobacco, Stupefies and Alcohol. And then we used the Pearson Correlation in this K-NN algorithm to measure of a linear correlation between two variables, as shown below:



where:

* r = Pearson Coefficient
* n= number of the pairs of the stock
* ∑xy = sum of products of the paired stocks
* ∑x = sum of the x scores
* ∑y= sum of the y scores
* ∑x2= sum of the squared x scores
* ∑y2= sum of the squared y scores

If the correlation coefficient is -1, it indicates a strong negative relationship. It implies a perfect negative relationship between the variables.

If the correlation coefficient is 0, it indicates no relationship. And if the correlation coefficient is 1, it indicates a strong positive relationship. It implies a perfect positive relationship between the variables.

While in DT algorithm, we used only three features to make the prediction: Lession, Color and Homogene. And last but not least, we used some features to project our prediction for the diagnosis and the dysplasia status: Homogene, Reticulated, Age, Sex, Tobacco, Stupefies and Alcohol.

In order to display in 2-D or 3-D the data, dimensionality reduction is needed. The simplest method is the Principal Component Analysis (PCA), which perform an orthogonal linear projection on the principal axsis (eigenvector) of the covariance matrix. We applied PCA to extract the relevant information from the data set and used a nonparametric K-NN analysis for classification. Then we plot the PCA (Figure 3) based on the Pearson Correlation we have produced above.

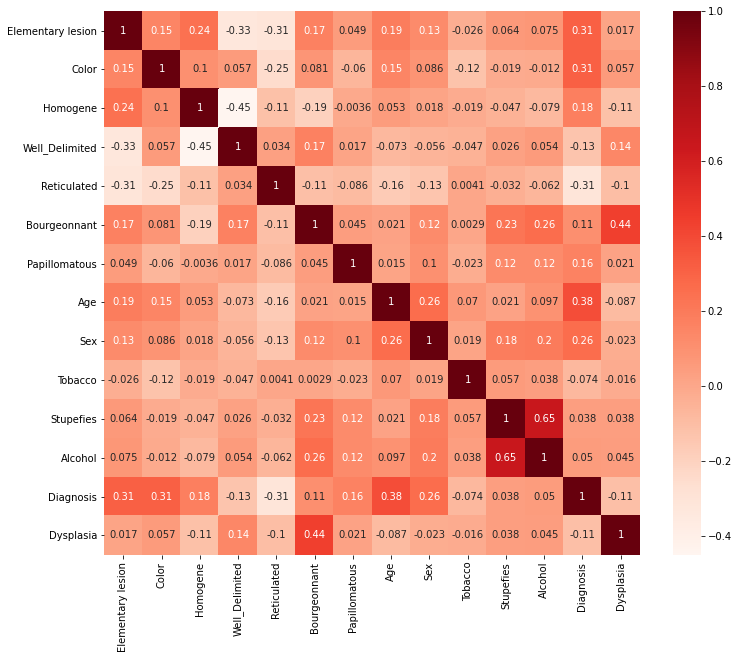


Figure 2: Formula of Pearson Correlation and its plot from the data set of the patients at CHU of Toulouse.

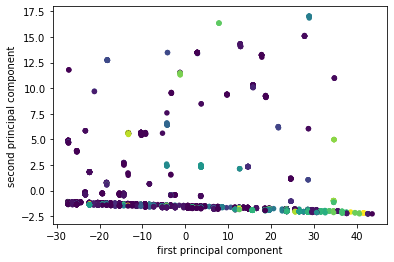


Figure 3: Plot of PCA based on Pearson Correlation.

# **K-NEAREST NEIGHBORS (K-NN) ALGORITHM**

## Overview about K-NN Algorithm

The K-NN is considered as a supervised machine learning algorithm. A supervised learning algorithm analyzes the training data which are labeled as an input data, to be able to produce a function which determines correctly new output when given new unlabeled data. Apparently, within the data science industry, it is widely used to solve classification predictive problems, although this algorithm can be used also for regression problems.

K-NN is a method for classifying objects based on the learning data that is closest to the object. The learning data is plotted into a multi-dimensional space with each dimension representing each feature of the data. New data classification is done by looking for the k label of the nearest neighbor. The most labels that appear are the new data labels. When k = 1, the new data is labeled with the closest neighbor label. The distance commonly used is the Euclidean distance. The Euclidean or Euclidean metric distance is the "usual" straight line distance between two points in Euclidean space. At this distance, the Euclidean space becomes a metric space.

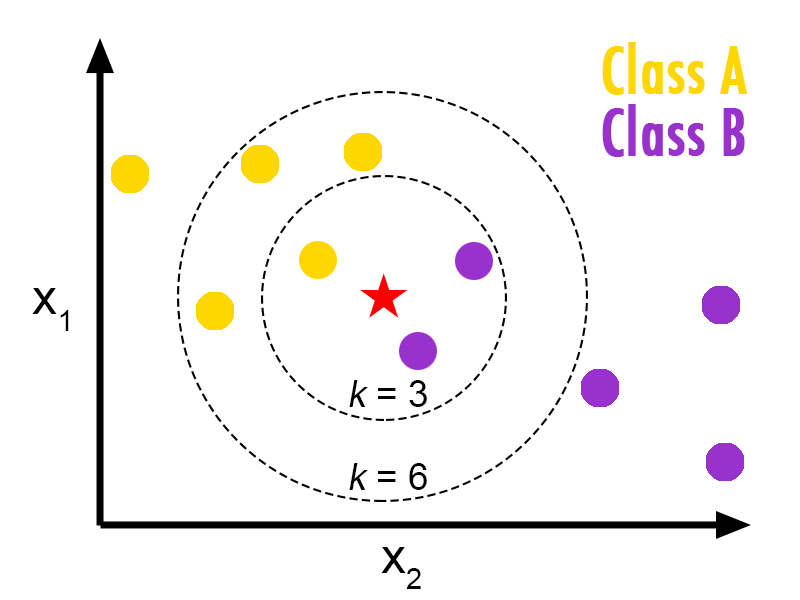


Figure 4: The way it works in K-NN algorithm.

## How It Works

Before being able to apply the K-NN algorithm, the data set has to be prepared for it. There are a few constraints that we have taken into account. Firstly the information has to be in a table format. Secondly, all the columns must only contain numerical data (binary) except for the last column, which has to be composed by labels that are able to confirm some functions are working.

Once the data set is cleaned (set) and in the correct structure, we used seaborn library and matplotlib to be able to produce Pearson Correlation for the data set. Then we print the features that appear to have the correlation, which are: Lession, Color, Homogene, Well Delimited, Reticulated, Bourgeonnant, Papillomatous, Age, Sex, Diagnosis and Dysplasia. We found out also that are 2 (two) features that have weak correlation: Stupefies and Alcohol. Using one of the Scikit-learn library, called PCA (Principal Component Analysis) we displayed the 2D plot for the model.

With the database compressed, we can start applying the KNN algorithm. To do so we used the following libraries:

1) *KNeighborsClassifier*: it is main library for the algorithm, it includes all the methods needed to archive the successful application of the algorithm.

2) *Train\_test\_split*: to be able to split the dataset into two.

Since we are doing supervised machine learning, we have to divide the data set into two parts; On one hand we have the training data set with the features and the labels. On the other hand we have the test data set that will help to validate the training data set, with their features and labels. While splitting the data, we used a test\_size of "0.25" and a random state "1234" using first Y1 as target and then we repeat the process to check Y2 as a target.

We can see that in the first target, the accuracy score calculated is 0.8747, lower than the one with the second target, which is 0.9881. The closest the score is to 1, the better the classification and the accuracy of the prediction of the data.

# **DECISION TREE (DT) ALGORITHM**

## Overview about DT Algorithm

As a supervised algorithm, decision tree can be used for both Classification and Regression problems, although mostly it is preferably to be used for Classification problems. This algorithm is a tree-structured classifier, where decisions, consequences and outcomes are being considered.

There are two basic nodes in decision tree (Figure 5): *(1) the Decision Node, and (2) Leaf Node*. The multiple branches exist in the Decision nodes and they are used to make any decision. While the output for those decision are called Leaf nodes and they do not contain any branches.

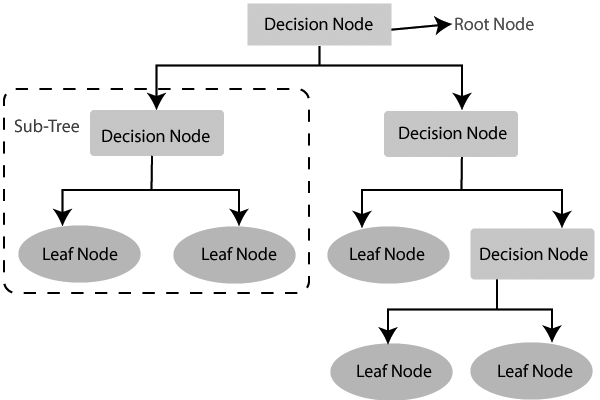


Figure 5: Two basic nodes of decision tree.

This algorithm is suitable for simple problem solving and if the data set is small. Although this algorithm is easy to understand, it has certain problems such as overfitting problems and biased results when working with unbalanced data sets. But on the other hand, decision tree is able to map linear and nonlinear relationships.

We are all aware that decision tree algorithm can be robust to analyze data set and provide prediction with a reasonable accuracy. Decision tree has been used for the past decades in medicine (healthcare) with the purpose to anticipate disease by analyzing patient's symptoms. If it can give prediction with strong accuracy that a patient will be sick, the sooner they can be treated and the better the chances to keep them healthy.

Nowadays, decision tree allows the scientists to analyze a set of patients, with different symptoms, and the disease they have. One of the main strengths of decision tree is that we can generate a tree graph to better see how it is working in the background.

## How It Works

When diving into the analysis, it is necessary to pre-process the data set before using it. So, we only keep the features that are relevant to our model. As previously explained, we focus on cell analysis with these features: Lesion, Color and Homogeneous. The data set contains 2362 patients, where the 116 patients have dysplasia status.

The value of the feature is a string, for example color is white, red or mixed. Decision Tree algorithm cannot work with strings, therefore we need to convert them into integers, in other words converting string values to numbers. So, one of the Scikit-learn python package called "Label Encoder" is used to do this task.

After the data set is converted into integers, the next step will be creating the model. Firstly, it will be done without training the data set. Like previously, we use function from Scikit-learn library, called the "Decision Tree Classifier", with the purpose to create the model that is easily to be treated. Next, we will fit the model passing as arguments, our inputs (features) and our target (dysplasia result).

The range score obtained is between 0.95 and 0.97. We can easily say that the accuracy of the model is pretty good and allows us to try some predictions.

For example, we try to predict if a patient is likely to have dysplasia, if they have the following feature values: No substance loss / White / No. The result shown is "No". We can expect a wrong prediction around only 5%. These results can be applied to real life only if our model is well constructed and takes into account everything we need to provide such predictions. The mean value increases slightly as we train the model. In most cases, it would be useful to divide the data set into two parts and test our model on the unused portion of the data set, but in our case, it seems irrelevant to do so.

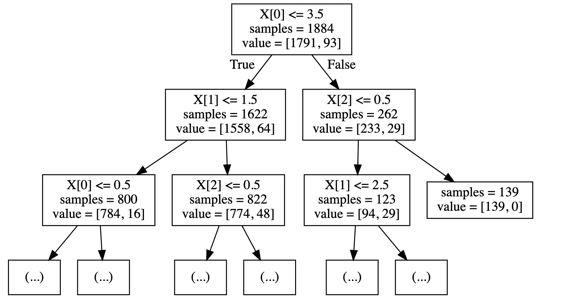


Figure 6: Decision Tree algorithm in classifying the data.

Looking at the graph (Figure 6), we can see how the DT algorithm classifies patients based on the feature values. To explain it, X [0] means we take the first feature in the feature array "X"; the sample is the number of patients we focus on; value is an array representing the repartition of the output (in this case "yes" = 1791 and "no" = 93, for the dysplasia evaluation).

The root node is about feature X [0] "base lesion", and separates between patients who have "volume augmentation" (coded as 4) or not. The tree will separate the values as: vertices where X [0] < 3 and vertices where X [0] > 3. This shows that only 262 patients had "volume augmentation" while the others had one of the other values.

We did also the prediction to see if the diagnosis has the high accuracy or not, or whether the prediction is still relevant. The result achieved was not so high compared to the dysplasia status, because it ranges between 0.60 to 0.65 that in other words, the accuracy is still good but it is not high enough seen from the side of accuracy. Besides, the values within the diagnosis are multi-categories which make it harder to provide prediction with high accuracy value.

# **ARTIFICIAL NEURAL NETWORK (ANN) ALGORITHM**

## Overview about ANN Algorithm

Artificial Neural Network is a popular supervised learning algorithm with the initial goal of simulating biological neural networks in humans. This artificial network is actually a simplified simulation of the neural network in the brain, but the simulations that are carried out are unable to describe the complexity of the human biological tissue.

ANN has neurons or nodes (vertices) and synapses (edges) that mimic the workings of the human brain network which is made up of cells called neurons. There are 3 layers in ANN, where each node in each layer is connected to all nodes in other layers. A deeper neural network can be created by increasing the number of the hidden layers. In fact, ANN receives multiple inputs and produces one output (Figure 7).

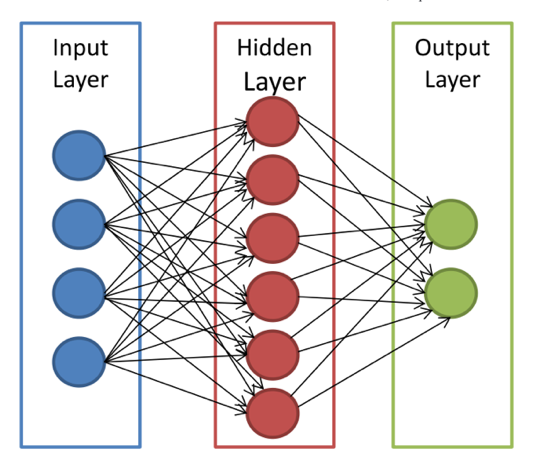


Figure 7: The ANN structure where arrows connect the output of one node to input of another.

The goal of this paper is to evaluate ANN in disease diagnosis, based on lesion and patient characteristics. And in the end, a prediction is provided for the diagnosis (multi-categorical prediction) and the dysplasia status (binary prediction). Each patient classified into multiple categories for the diagnosis: Leucokeratose, Diapneusie, Papillome, Carcinome, Mycose, etc. And also they are classified as dysplasia and non-dysplasia. Classification is an important tool in medical diagnosis for decision support. Feed-forward back propagation neural network is used as a classifier to distinguish between dysplasia or non-dysplasia person in both cases. In this study, the data were obtained from the Centre Hopitalier Universitaire (CHU) of Toulouse, in order to diagnose diseases. The data is separated into inputs and targets. The targets for the neural network will be identified with 1's as dysplasia and will be identified with 0's as non-dysplasia.

## How It Works

First of all, the necessary libraries (Figure 8) needed have to be imported to run this ANN algorithm in Python. They are:

### Glob: this module (short for global) is a useful Python library that is used to return all file paths matched with specific pattern.

### Keras: a neural network library to build and train model easily in machine learning, that provides high-level APIs, while TensorFlow provides both low and high level APIs.

### NumPy: an extension of Numeric and Numarray. It contains a multi-dimensional array and matrix that can be used to perform a number of mathematical operations on arrays: trigonometric, statistical and algebraic routines..

### Pandas: the most popular library in Python for data analysis and manipulation tool. It provides streamlined forms of data representation and it is fast, powerful and easy to use.

### Scikit-learn: the most useful library for machine learning in Python. It contains a lot of efficient tools not only for machine learning but also for statistic modeling including classification, clustering, regression and dimensionality reduction.

### Matplotlib: a plotting library in Python for 2D plots of arrays. It is a multi-platform data visualization and it generates high quality line plots, scatter plots, bar charts and histograms.

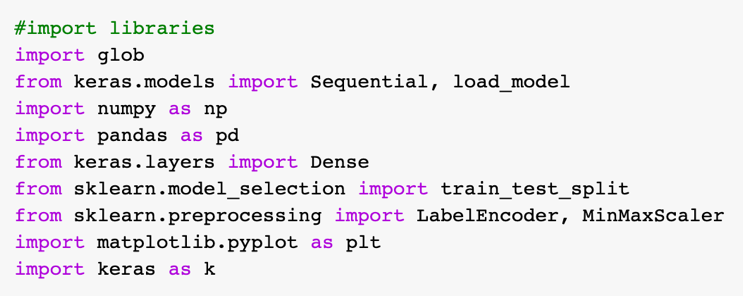


Figure 8: Libraries needed for running ANN algorithm.

Once we have imported the libraries needed to run this algorithm, we load the data set and then check if there is something unusual. We will use the features to produce the prediction for diagnosis and dysplasia status. We will create a list of column names we want to keep/to use (the features: Homogene, Reticulated, Age, Sex, Tobacco, Stupefies and Alcohol). Then we will drop the rows of the data with n/a or missing values.

First we will predict if the dysplasia status has the high accuracy using this ANN algorithm. As we notice that the status of the dysplasia contains yes and no, we will convert this into binary (0 and 1). Once it is converted, the data will be divided into independent (x) dataset (the features) and dependent (y) data set (the target). Then we used one of the Scikit- learn library, Min-MaxScaler to scale the data set so that all the input features will lie between 0 to 1. The next step will be splitting the data into 80% for training and 20% for testing purpose then we will shuffle the data.

The next step is building the model by using Sequential library and then we complied the model to gain insight about the loss and the accuracy of the prediction. And after, we trained the model by setting the epoch up to 2000. based on the training, we could have the information that the prediction of the data based on the dysplasia status is around 0.95 which we find very good and it has high accuracy. And the last step is we plotted the model loss vs accuracy and then showed the actual values versus the prediction values.

We tried to implement the same algorithm but with the target being “diagnosis”, the data set must be updated same as the target was "dysplasia", which is a binary classification problem, where we had two textual values that we encoded into 0 and 1. However, we cannot employ the identical method with the target "diagnosis" because we have more than two values. Consequently, we are facing a multi-class classification problem. To resolve this problem, we have to place every type of values of the column "diagnosis" in a new column and then encode the values into 1 or 0. We would be performing a binary classification with each unique value of the column diagnosis. The result that we obtained has shown us the score averagely 0.65 using this ANN algorithm, which has the similar score as the DT algorithm we ran before.

#### **CONCLUSION**

To conclude, if we compare the 3 (three) algorithms in general cases, the three algorithms are highly performant, but the question is which algorithm is more suitable for our data set, that can provide the highest accuracy score. Our results have shown that the K-Nearest Neighbors algorithm is the most appropriate one for this kind of data set. The prediction of the ANN algorithm and the Decision Tree algorithm was almost similar and slightly lower than the KNN's prediction.

We managed to show that 3 (three) algorithms in machine learning can be used to support the predictive accuracy of a data set. Despite the challenges and barriers above, the potential for ML-based AI approaches to digital diagnosis is promising because AI has strong feature representation learning capabilities made possible by improved algorithms, big data accumulation and increased computing power. People will have more confidence in AI algorithms once they have been validated using multi-center data and have increased interpretability. Collaboration between diagnostic experts and AI will promote more precise treatment for many patients in healthcare.

##### **REFERENCES**

1. F. Ahmed, “Artifical neural networks for diagnosis and survival prediction in colon cancer,” *Mol Cancer*, vol. 4, Aug. 2005.
2. L. Kuo-An, "Developmental dysplasia and dislocation of the hip in adults," *Springer Singap*., vol. 1, 2018.
3. T. Ankur, A. Muhammad Aurangzeb, E. Carly and K. Vikas, "Explainable machine learning models for healthcare AI," *ACM*, 2018.
4. J. Yahui, Y. Meng, W. Shuhao, L. Xiangchun and S. Yan, “Emerging role of deep learning-based artificial intelligence in tumor pathology,”in Cancer Commun. (London), p. 154-166, April 2020.
5. A. Filippo, L. Alberto, PM. Eladia, V. Petr, H. Ales, H. Josef, "Artifical nueral networks in medical diagnosis," *Journal of Applied Biomedicine*, 2013.
6. K. Konstantina, P.E. Themis, P.E. Konstantinos, V.K. Michalisand I.F. Dimitrios, "Machine learning applications in cancer prognosis and diagnosis," *Biotech. Journal*, vol. 13, p. 8-17, 2015.
7. Y. Yazeed, D. Shira and S. Noam, “Machone learning-based prediction of COVID-19 diagnosis based on symptoms,” *npj Digital Medicine*, Jan. 2021.
8. Y. Mingyu, l. Guo-Zheng, "Medical diagnosis by using machine learning techniques," *Research Gate*, Dec. 2013.
9. H. Bert, E. Simon, "*Your evidence? Machine learning algorithms for medical diagnosis and prediction*," Online library, Dec. 2019.