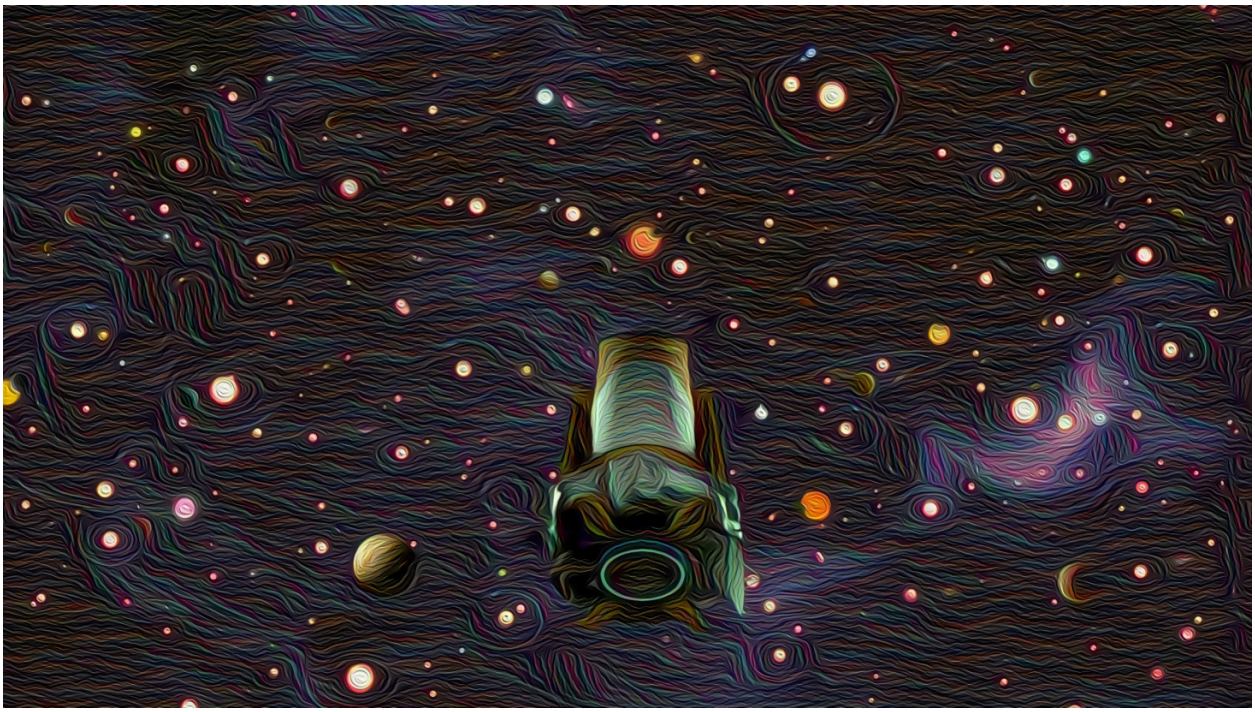
Machine Learning Pour la détection d'exoplanètes

Angelo Charry & Louis Delloye

3 février 2020



SCIENCES
SORBONNE
UNIVERSITÉ

1 Positionnement du problème

1.1 La mission Kepler [4]

La mission Kepler a été lancée en mars 2009 et terminée en octobre 2018. Elle a pour but la détection par transit d'exoplanètes de tailles comparables à celle de la Terre dans des régions proches de la zone habitable. Cette mission a permis d'estimer le nombres d'étoiles dans la voie lactée possédant une planète de ce type.

1.2 La méthode du transit [6] [9] [10]

Cette méthode est conceptuellement simple. On observe les variations de flux d'une étoile. Si un autre objet passe devant pendant l'observation, le flux diminue. Cette variation peut donner une estimation de la taille de l'objet. Si on observe un transit complet, on peut déterminer la période de révolution de l'objet, ce qui permet, à l'aide des lois de Kepler, de déterminer le demi-grand axe de révolution. Pour déterminer la masse de l'objet il faudra mesurer la vitesse radiale de l'étoile (induite par les perturbations gravitationnelles de l'objet étudié).

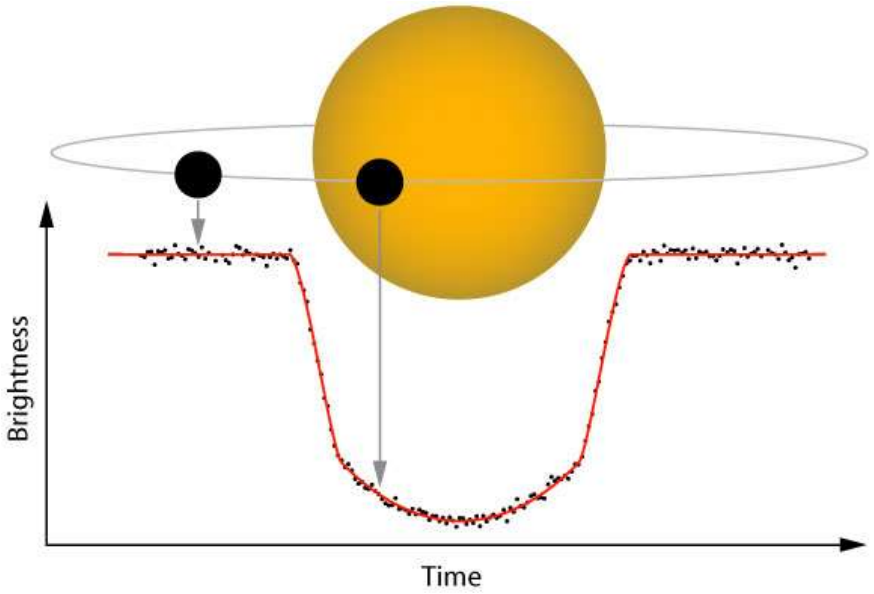


Figure 1 – Schéma d'un transit planétaire © Institute for Astronomy-University of Hawaï

1.3 Collecte des données [5]

Le télescope est muni de 42 caméras CCD (Charge-Coupled Device) ce qui lui permet d'observer simultanément près de 150 000 étoiles (sélectionnées pour leur luminosité). L'acquisition a lieu toutes les 30min environ.

La mission Kepler est divisée en 19 campagnes durant lesquelles le télescope observe une partie du ciel pour approximativement 80 jours.

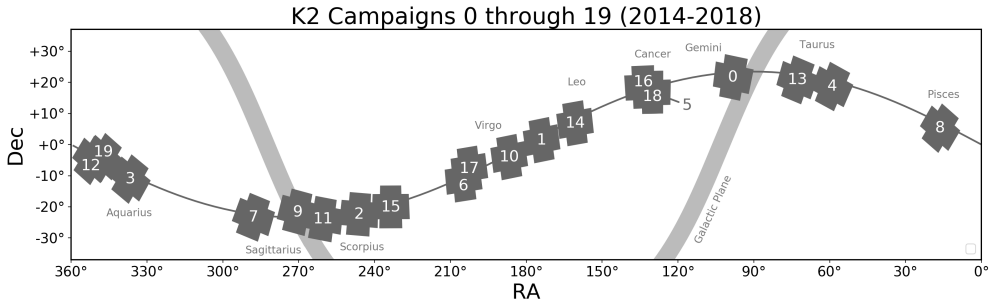


Figure 2 – Campagnes de la mission Kepler

1.4 Confirmation des exoplanètes [3]

Kepler est capable de récolter beaucoup de données simultanément, mais ses mesures ne sont pas suffisamment précises pour conclure sur la nature des observations. Les causes de variations dans la mesure du flux de l'étoile peuvent être nombreuses (poussière cosmique, astéroïdes, système binaire, naine brune, etc).
De plus, nous avons vu que le transit photométrique seul peut déterminer uniquement la taille de l'objet orbitant autour de l'étoile. Les mesures de Kepler ne sont donc pas auto-suffisantes, elles doivent être reprises par des observations approfondies sur Terre. En fait les observations de Kepler permettent de déterminer ce qu'il est nécessaire d'étudier.

N.B : Ainsi, dans le dataset, si l'étoile est classée comme ne présentant pas d'exoplanète l'orbitant, cela n'implique pas nécessairement qu'aucun autre objet n'orbite cette étoile.

2 Les Données

2.1 Aperçu général

Les données ont été collectées par WΔ [11] et sont disponibles sur Kaggle. Les données labellisées comme "sans exoplanètes" proviennent de la campagne 3 car c'est la campagne la moins susceptible de contenir des exoplanètes mal classées. Les données labellisées comme "exoplanète confirmée" proviennent de plusieurs campagnes différentes afin d'augmenter le nombres d'exoplanètes dans le dataset. Le dataset contient 42 exoplanètes confirmées pour 5615 non-exoplanètes. On repère donc immédiatement une difficulté : la classe exoplanète est sous-représentée (moins de 1% du dataset) et donc le dataset est très mal "équilibré".

Les données se présentent comme suit :

exoTest.csv (27.57 MB) — 20 of 100 columns — Views

	# LABEL	# FLUX.1	# FLUX.2	# FLUX.3	# FLUX.4	# FLUX.5	# FLUX.6	# FLUX.7	# FLUX.8	# FLUX.9
1	2	119.88	100.21	86.46	48.68	46.12	39.39	18.57	6.98	6.63
2	2	5736.59	5699.98	5717.16	5692.73	5663.83	5631.16	5626.39	5569.47	5550.44
3	2	844.48	817.49	770.07	675.01	605.52	499.45	440.77	362.95	287.27
4	2	-826	-827.31	-846.12	-836.03	-745.5	-784.69	-791.22	-746.5	-709.53
5	2	-39.57	-15.88	-9.16	-6.37	-16.13	-24.05	-0.9	-45.2	-5.04
6	1	14.28	10.62999999999999	14.55999999999999	12.41999999999998	12.06999999999999	12.91999999999998	12.27	3.190000000000005	8.470000000000003
7	1	-150.479999999999996	-141.720000000000001	-157.599999999999999	-184.599999999999999	-164.889999999999999	-173.869999999999995	-162.909999999999996	-167.040000000000001	-172.759999999999995
8	1	-10.06	-12.78	-13.16	-9.81	-18.91	-20.33	-22.85	-19.17	-17.97
9	1	454.660000000000003	440.599999999999977	382.289999999999979	361.629999999999976	298.629999999999976	253.289999999999979	155.859999999999986	110.379999999999976	31.7099999999999919
10	1	187.399999999999994	209.599999999999991	199.909999999999989	179.619999999999995	171.209999999999992	161.839999999999997	163.019999999999999	171.610000000000001	113.529999999999999
11	1	205.069999999999992	177.979999999999996	163.410000000000003	159.699999999999997	157.709999999999992	167.569999999999992	191.279999999999999	196.910000000000003	187.229999999999996
12	1	335.739999999999991	330.209999999999992	290.659999999999989	274.179999999999993	271.239999999999991	176.419999999999998	176.75	132.329999999999987	85.5099999999999948
13	1	-8.789999999999996	0.040000000000000773	0.660000000000000082	-2.63	4.420000000000000007	-2.949999999999993	-12.17	0.670000000000000073	2.210000000000000004
14	1	449.159999999999989	419.779999999999999	357.449999999999997	355.949999999999997	284.409999999999989	251.739999999999991	192.789999999999994	128.349999999999991	110.339999999999997
15	1	154.579999999999994	127.269999999999997	128.859999999999993	122.529999999999999	94.8099999999999977	78.9399999999999951	76.1899999999999951	66.6899999999999951	40.0699999999999997
16	1	-9.180000000000000006	-9.430000000000000006	-11.42000000000000001	-8.889999999999999999	-3.480000000000000002	-13.66000000000000001	-8.87	-7.830000000000000004	-1.019999999999999998
17	1	53.849999999999999985	24.779999999999999988	67.639999999999999994	46.139999999999999994	23.639999999999999994	30.489999999999999998	31.209999999999999991	34.25	60.470000000000000012
18	1	3.119999999999999989	-2.320000000000000016	1.329999999999999993	-9.670000000000000007	-5.070000000000000016	-11.99000000000000002	-10.96	-8.210000000000000004	-1.510000000000000022
19	1	-14.459999999999999999	-9.689999999999999994	-9.279999999999999997	-7.349999999999999991	-12.489999999999999999	2.220000000000000003	-4.37	-1.139999999999999999	2.040000000000000008
20	1	-109.52	-85.900000000000000001	-87.96	-75.349999999999999999	-74.369999999999999999	-61.360000000000000001	-46.219999999999999998	-34.159999999999999999	-23.690000000000000001

Sur la première colonne se trouvent les labels (2 pour les exoplanètes, 1 pour les non-exoplanètes) puis les 3197 autres colonnes représentent les valeurs de flux (en e^-/sec) prises à environ 30min d'intervalle.

Pour traiter les données on utilise une formule proche de la RPN (Relative Power Noise)

$$S_{RPN}(t) = \frac{S(t) - \langle S(t) \rangle}{max\left(|S(t)|\right)} \tag{1}$$

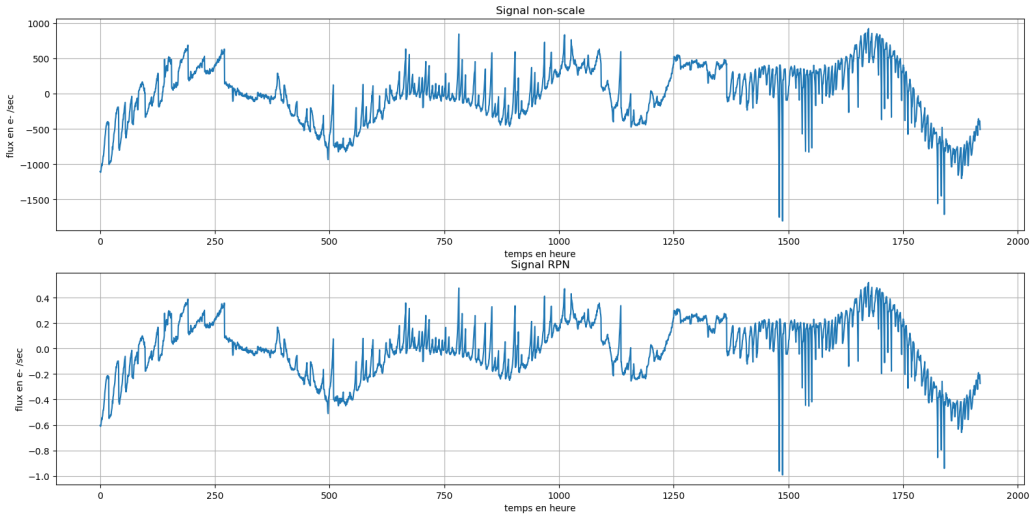


Figure 3 – Signal d'une étoile contenant une exoplanète brute (en haut) VS RPN (en bas)

On remarque que cette méthode permet de préserver parfaitement le signal tout en le restreignant entre -1 et 1 et en le recentrant autour de 0.

2.2 Bootstrapping [12]

Pour rééquilibrer le dataset nous avons utilisé une méthode de bootstrapping, ie que l'on copie de manière aléatoire les données minoritaires (ici les étoiles comportant au moins une exoplanète orbitant autour d'elle) afin de rééquilibrer le dataset. Nous utilisons la symétrie par renversement du temps pour augmenter le nombre de data labellisée exoplanète.

3 Méthode de classification

3.1 Architecture du modèle

Nous avons utilisé pour classifier les données un réseau séquentiel composé de réseaux convolutifs (1D) [1] et de LSTM [7] (cf fig.4). En effet, les données à notre disposition étant des time-series, il est presque indispensable d'utiliser un type de réseau récurrent. Nous utiliserons donc ici les couches Long Short Term Memory (LSTM).

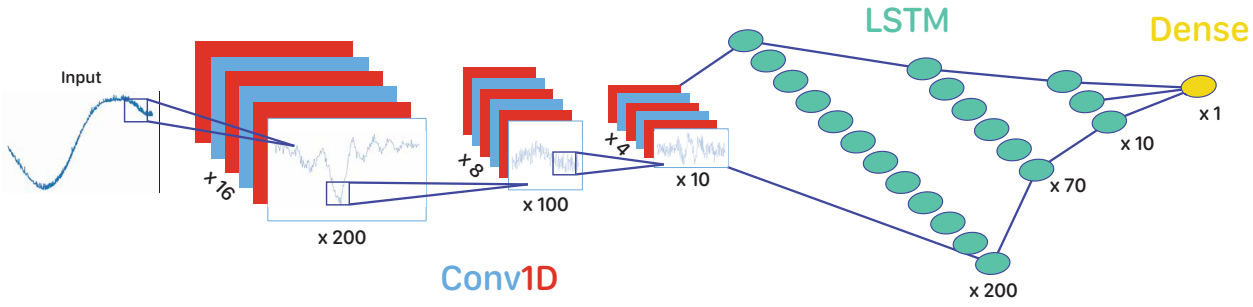


Figure 4 – Architecture du réseau de neurones

3.2 Évaluation du modèle

3.2.1 Les bonnes métriques [8] [2]

D'une manière un peu naïve, on pourrait utiliser la métrique "accuracy" pour évaluer notre modèle. En procédant ainsi et sans "bootstrapper" nos données, on obtient d'emblée environ 99% d'accuracy : le modèle apprend à tout classer comme "étoile sans exoplanète l'orbitant" et ainsi il minimise son erreur puisque cette dernière catégorie est largement majoritaire. Il est donc primordial d'utiliser des métriques adaptées. En somme, la métrique "accuracy" est adéquate pour un dataset équilibré. Pour nous la métrique la plus pertinente est recall car nous ne voulons pas manquer d'exoplanètes. Quoiqu'il en soit, nous comparerons tout de même les différentes métriques :

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+FP+FN+TN} \\ \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{f1} &= \frac{2*(\text{Recall}*\text{Precision})}{(\text{Recall}+\text{Precision})} \end{aligned} \tag{2}$$

où on choisit la convention suivante : TN (True Negatives) correspond aux étoiles sans exoplanète bien classées, TP (True Positives) correspond aux étoiles avec exoplanète(s) bien classées, FP (False Positives) correspond aux étoiles sans exoplanète mal classées, FN (False Negatives) correspond aux étoiles avec exoplanètes mal classées.

3.2.2 Validation croisée

Afin d'évaluer le modèle de la manière la plus fidèle possible, on effectue une validation croisée. Celle-ci se déroule comme suit : on découpe aléatoirement toutes les données en un ensemble d'entraînement et un ensemble de test, puis on entraîne le modèle à chaque découpage avec le nouvel ensemble d'entraînement et on l'évalue grâce au nouvel ensemble test.

Bien entendu, cela n'est pas aussi simple que de découper à l'aveuglette les données en k parties puis d'en sélectionner une pour le test et de garder le reste pour l'entraînement. En effet, en raison du "déséquilibre" des données, il faut être plus vigilant. On commence donc par séparer les étoiles autour desquelles orbite au moins une exoplanète, de celles qui n'en possèdent pas. On commence par effectuer le découpage sur chacun de ces sous-ensembles. Une fois ceci fait on peut bootstrapper les deux sous-ensembles (entraînement et test) correspondant aux exoplanètes. On concatène par la suite la partie entraînement sélectionnée pour les étoiles et celles pour les exoplanètes, et on procède de même pour le test. On peut à présent évaluer notre modèle sur chacun de ces différents découpages.

4 Résultats

4.1 Des résultats prometteurs

À première vue notre modèle est prometteur. Il classe de manière générale avec un recall de l'ordre de 95%. Comme on peut le remarquer sur les matrices de confusion suivantes :

$$\begin{pmatrix} 725 & 78 \\ 0 & 803 \end{pmatrix} \quad \begin{pmatrix} 775 & 27 \\ 134 & 668 \end{pmatrix} \quad \begin{pmatrix} 745 & 57 \\ 0 & 802 \end{pmatrix} \quad \begin{pmatrix} 753 & 49 \\ 274 & 528 \end{pmatrix} \quad \begin{pmatrix} 779 & 23 \\ 272 & 530 \end{pmatrix} \quad \begin{pmatrix} 789 & 13 \\ 442 & 360 \end{pmatrix} \quad \begin{pmatrix} 759 & 43 \\ 0 & 802 \end{pmatrix}$$

Elles se lisent de la façon suivante : $\begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}$

Les AUC ROC (Receiver operating characteristic) scores sont : 0.98, 0.98, 0.98, 0.88, 0.86, 0.88, 0.98

On peut établir la matrice de confusion moyenne ainsi que son écart type :

$$\overline{confusion\_matrix} \quad \pm \quad \Delta\text{confusion\_matrix} = \begin{pmatrix} 641.85714286 & 160.28571429 \\ 41.42857143 & 760.71428571 \end{pmatrix} \pm \begin{pmatrix} 161.62604862 & 161.48393488 \\ 20.68717438 & 20.43606256 \end{pmatrix}$$

De même, on peut établir, à partir de ces différentes matrices, les moyennes de chaque métrique :

$$Accuracy = 0.87 \pm 0.0915 \tag{3}$$
$$Precision = 0.94 \pm 0.0205 \tag{4}$$
$$Recall = 0.8 \pm 0.201 \tag{5}$$
$$f1 = 0.90 \pm 0.0487 \tag{6}$$
$$AUC = 0.93 \pm 0.0535 \tag{7}$$

Le modèle présente donc des résultats qui semblent concluants en validation croisée, bien que sur deux découpages près d'un tiers des étoiles "avec exoplanètes" sont mal classées.

4.2 Prise de recul sur les résultat

En étudiant les données mal classées, on remarque que certaines données posent problème, qu'elles soient dans l'ensemble d'entraînement ou dans le test. En effet, certaines données sont mal classées à chaque découpage, peu importe si le modèle s'est entraîné avec ou non. On a donc identifié quelques-unes de ces données (cf fig.5) :

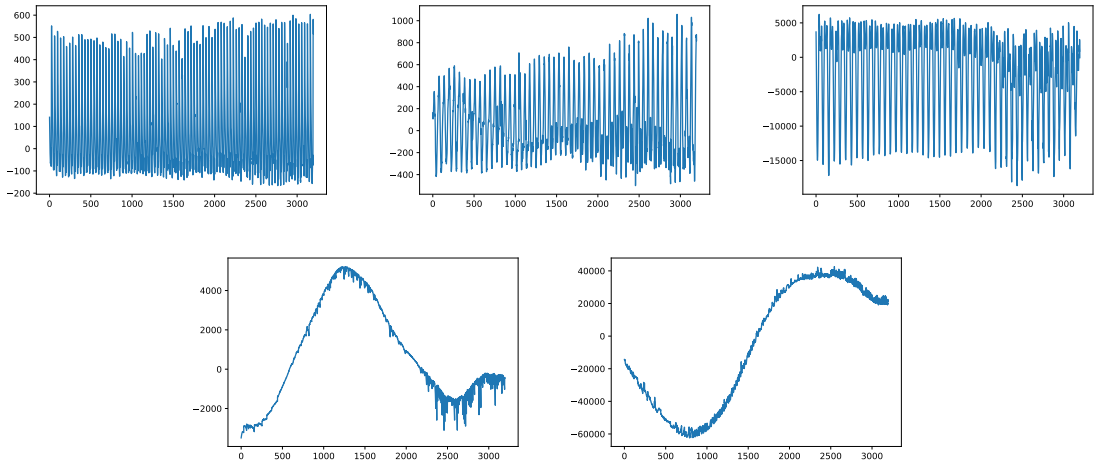


Figure 5 – Exemples de données problématiques : ces cinq exemples correspondent à des étoiles labellisées comme "sans exoplanète".

On remarque assez aisément pourquoi le modèle a du mal à classer les trois premiers exemples. Ce n'est en effet pratiquement que du bruit, ou du moins, on ne distingue rien de significatif permettant d'identifier un véritable flux.

Cependant, pour les deux exemples suivants le problème semble être d'une toute autre nature. On distingue deux courbes très marquées avec ce qui ressemble fortement à des courbes traduisant le "transit" d'une exoplanète devant les étoiles en question. Le fait qu'une étoile soit classée comme "sans exoplanète" n'implique pas qu'un autre astre n'orbite pas cette étoile. En effet, on pourrait être dans le cas présenté plus haut où la méthode du transit traduit la présence d'un astre mais d'autres méthodes ont par la suite éliminé la possibilité que cet astre soit une exoplanète. Il y a donc un véritable problème avec ce dataset puisque l'on doit être en mesure de distinguer des systèmes binaires à partir de données photométriques, ce qui n'est pas possible en pratique. Il faudrait soit un troisième label pour les systèmes binaires, soit une feature de plus indiquant si l'étoile observée est connue comme possédant un compagnon.

5 Axes d'amélioration

L'un des principaux défauts de notre modèle est qu'il ne respecte pas la symétrie par renversement du temps. Les couches LSTM et convolutionnelles brisent inévitablement cette symétrie. Pour le vérifier, il suffit de comparer les prédictions du modèle sur les données normales et renversées.

$$\text{matrice de confusion données normales} = \begin{pmatrix} 533 & 29 \\ 1 & 4 \end{pmatrix}$$

$$\text{matrice de confusion données reversées} = \begin{pmatrix} 514 & 48 \\ 1 & 4 \end{pmatrix}$$

On remarque que le renversement du temps perturbe légèrement le modèle surtout pour les données labellisées "sans exoplanètes". On pourrait imaginer un modèle où l'on renverse le signal tout en conservant le signal d'origine (cf fig.6). Les deux signaux passent ensuite parallèlement dans le même réseau. Il faudra faire attention à traiter les deux signaux exactement de la même façon. Il suffira alors de faire la moyenne des prédictions en sortie. On pourrait espérer avec cette méthode avoir un peu moins de faux négatifs et améliorer légèrement les prédictions.

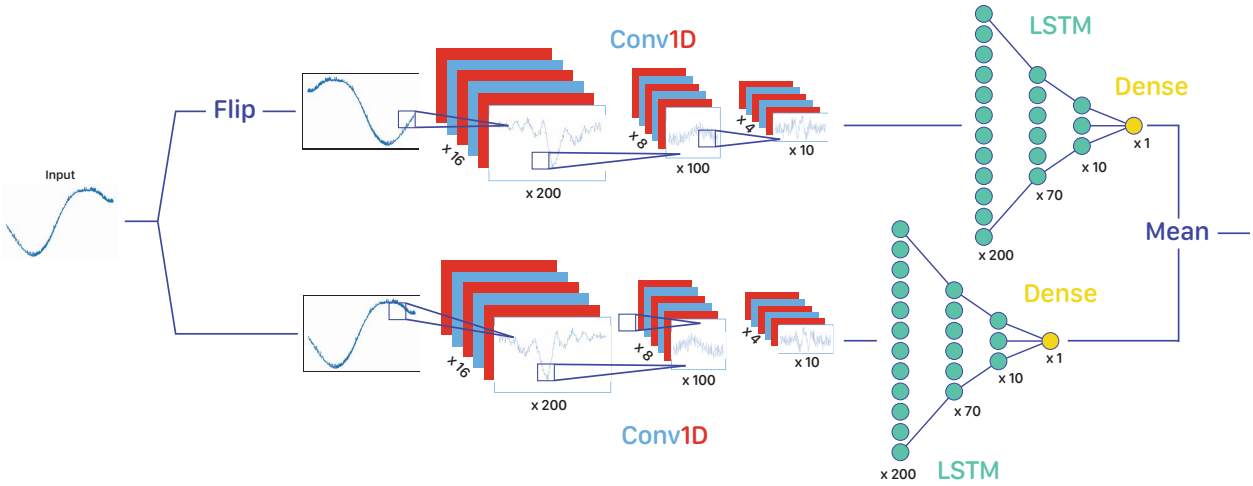


Figure 6 – Architecture d'un réseau symétrique par renversement du temps

Références

[1] *Convolutional neural network*. In : *Wikipedia*. Page Version ID : 929708718. 7 déc. 2019. URL : https://en.wikipedia.org/w/index.php?title=Convolutional_neural_network&oldid=929708718 (visité le 07/12/2019).

[2] Hugo Ferreira. *Confusion matrix and other metrics in machine learning*. Medium. 4 avr. 2018. URL : https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning-894688cb1c0a (visité le 07/12/2019).

[3] *How do you find – and confirm – a planet ? 10 things about the search for exoplanets – Exoplanet Exploration : Planets Beyond our Solar System*. URL : https://exoplanets.nasa.gov/news/1524/how-do-you-find-and-confirm-a-planet-10-things-about-the-search-for-exoplanets/ (visité le 07/12/2019).

[4] Michele Johnson. *Mission overview*. NASA. 13 avr. 2015. URL : http://www.nasa.gov/mission_pages/kepler/overview/index.html (visité le 07/12/2019).

[5] Michele Johnson. *Spacecraft and Instrument*. NASA. 13 avr. 2015. URL : http://www.nasa.gov/mission_pages/kepler/spacecraft/index.html (visité le 07/12/2019).

[6] *Méthodes de détection des exoplanètes*. In : *Wikipédia*. Page Version ID : 156190964. 26 jan. 2019. URL : https://fr.wikipedia.org/w/index.php?title=M%C3%A9thodes_de_d%C3%A9tection_des_exoplan%C3%A8tes&oldid=156190964 (visité le 07/12/2019).

[7] Aditi Mittal. *Understanding RNN and LSTM*. Medium. 12 oct. 2019. URL : https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e (visité le 07/12/2019).

[8] Marcos Silva. *Confusion Matrix — Deep Dive*. Medium. 15 nov. 2019. URL : https://towardsdatascience.com/confusion-matrix-deep-dive-8a028b005a97 (visité le 07/12/2019).

[9] *Transit (astronomie)*. In : *Wikipédia*. Page Version ID : 161628112. 8 août 2019. URL : https://fr.wikipedia.org/w/index.php?title=Transit_(astronomie)&oldid=161628112 (visité le 07/12/2019).

[10] *Transit Photometry*. URL : https://www.planetary.org/explore/space-topics/exoplanets/transit-photometry.html (visité le 07/12/2019).

[11] *winterdelta - Overview*. GitHub. URL : https://github.com/winterdelta (visité le 07/12/2019).

[12] Lorna Yen. *An Introduction to the Bootstrap Method*. Medium. 28 jan. 2019. URL : https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60 (visité le 07/12/2019).