Kepler Star Classification with Convolutional Neural Network on Light Curve Frequency Power Spectra

Related Studies

1. Deep learning classification in asteroseismology
   1. <https://doi.org/10.1093/mnras/stx1174>
   2. Classified evolutionary state of red giants with 99% accuracy using CNN on folded power spectrums.
   3. Implications
      1. Automated methods do exist, however considerable effort is required into defining and acquiring features such as the observed period spacing ΔP (Bedding et al. 2011; Stello et al. 2013), the asymptotic period spacing ΔΠ1 (Bedding et al. 2011; Mosser et al. 2014; Vrard, Mosser & Samadi 2016) or the structure of mixed modes (Elsworth et al. 2016) in order to separate the populations. Furthermore, these methods require relatively high signal-to-noise data.
   4. Data
      1. Labeled Set
         1. 6008 Stars
         2. 1008 Testing
         3. 5000 Training
      2. Unclassified Set
         1. 8794 Kepler red giants
2. Machine Learning Kepler Exoplanet Validation
   1. <https://academic.oup.com/mnras/article/504/4/5327/5894933#248198424>
3. Deep Learning on Asteroseismology
   1. <https://arxiv.org/abs/1811.03639>
   2. <https://www.researchgate.net/publication/335376321_Deep_Learning_Applied_to_the_Asteroseismic_Modeling_of_Stars_with_Coherent_Oscillation_Modes>
   3. Modeled 7 pulsating stars, good agreement with past results
4. Machine Learning Techniques for Stellar Light Curve Classification
   1. <https://iopscience.iop.org/article/10.3847/1538-3881/aac16d/meta>
   2. 75% accuracy
   3. Simple Aperture Photometry (SAP) flux and Pre-Search Data Conditioning (PDC) SAP flux. The PDC SAP versions of the light curves remove instrumental variations and noise in the data while preserving both stellar and transiting exoplanet behavior. Therefore this is the flux measurement used to construct the light curves.
   4. Representation Learning
5. Kepler Star Classification Dataset
   1. <https://iopscience.iop.org/article/10.3847/1538-4365/229/2/30>
6. Kepler Dataset 2
   1. <https://iopscience.iop.org/article/10.3847/1538-4365/aab4f9>
7. Gaia Data
   1. <https://arxiv.org/pdf/1609.04172.pdf>

Star Classification using Deep Learning and Asteroseismology

For incomplete light curves

Dataset

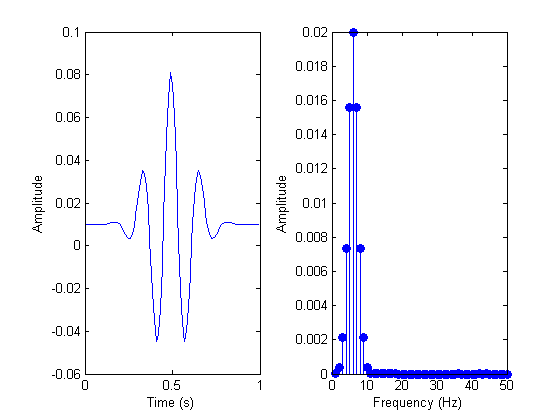
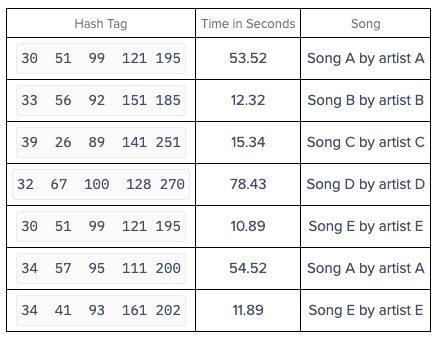
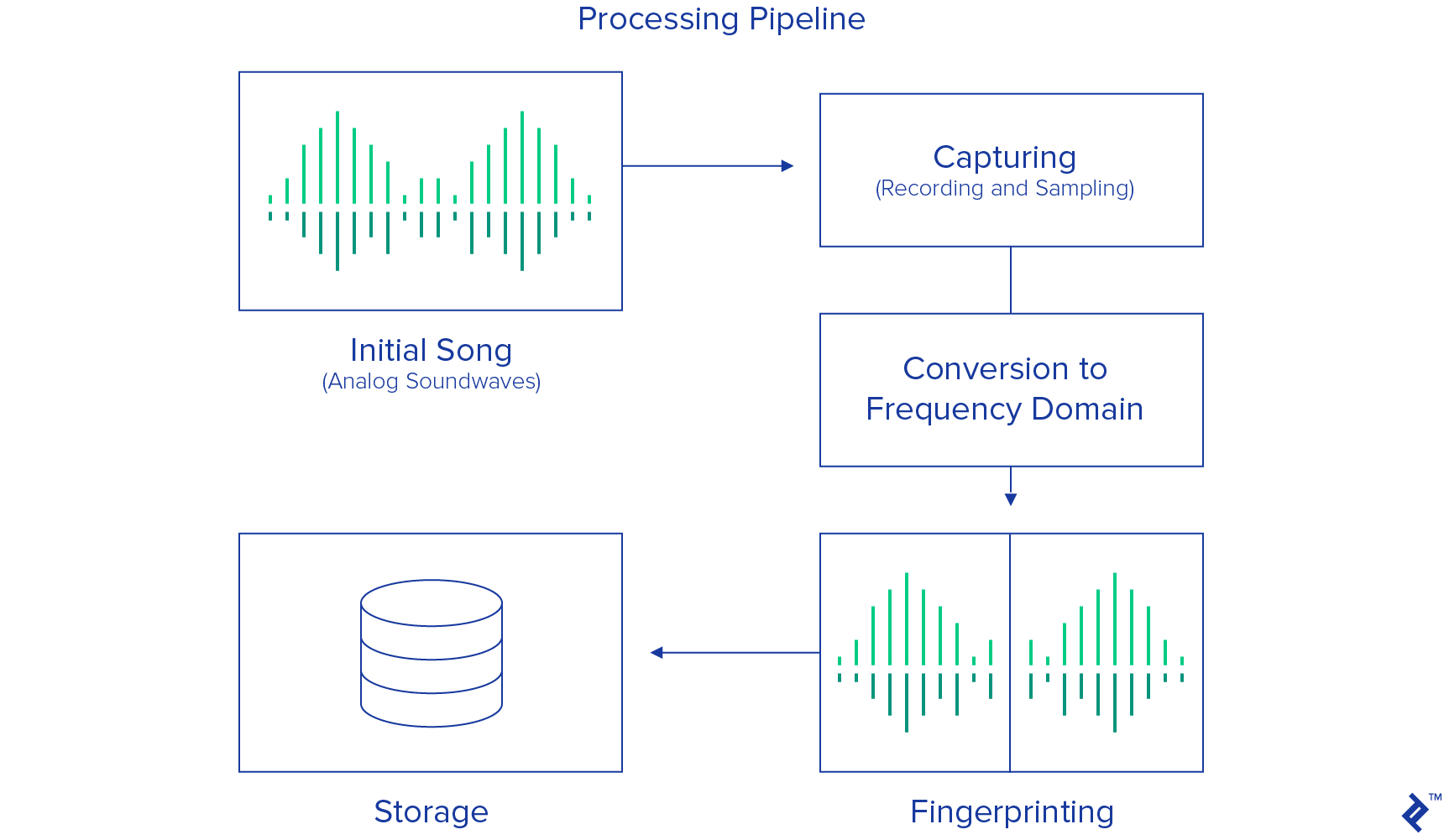
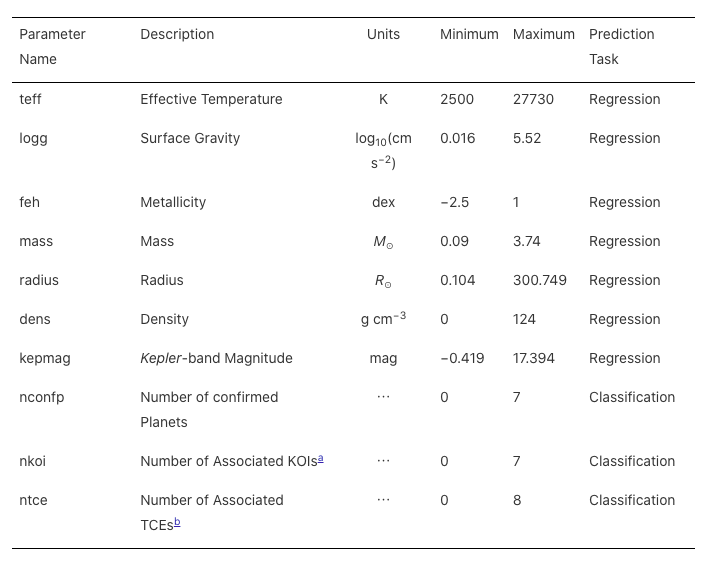
<https://exoplanetarchive.ipac.caltech.edu/bulk_data_download/>

Light curves were generated by the Kepler pipeline and are archived at MAST. Here, we provide wget scripts to retrieve all of the light curves from the NASA Exoplanet Archive. While the wget scripts are relatively small (script file size are given in Description), the volume of data retrieved is significant. **Most Kepler Quarterly scripts will retrieve 175 GB and the full set of Kepler light curves approach 3 TB.**

Light curve files are produced for each target using simple aperture photometry. For any particular quarter, there are more than 160,000 long cadence targets and up to 512 short cadence targets being archived. Short cadence targets always have a corresponding long cadence light curve. Long cadence targets will be observed for at least a quarter and short cadence targets will be observed for at least a month (except for Q4 where targets on module 3 were lost due to a hardware failure). In the case where a target is observed at both long and short cadence, there will be one long cadence light curve each quarter and up

to three short cadence light curves, one for each month. A light curve file contains time series data. Each data point corresponds to a measurement from a long or short cadence. Long and short cadence data are not mixed in a given light curve file. For each data point there are multiple flux and centroid values along with uncertainties. The value NaN is specified for any

missing data values. The light curves are packaged as FITS binary table files with a primary header, a light curve extension and an aperture extension. The long cadence FITS headers are listed in Appendix A.1.

1. Background
   1. Kepler Space Telescope
      1. <https://archive.stsci.edu/files/live/sites/mast/files/home/missions-and-data/k2/_documents/MAST_Kepler_Archive_Manual_2020.pdf>
      2. The Kepler mission was designed to survey a region of the Milky Way galaxy to detect and characterize Earth-size and smaller planets in or near the habitable zone by using the transit method of planetary detection. This was accomplished by observing changes in the brightness of stars in the same patch of sky for 4 years between May 2009 and May 2013.
      3. 200000 distant stars
         1. Mostly labeled
         2. Unlabeled - around 2000 stars
            1. Mostly due to short light curves or missing data
      4. Present Researches
         1. Machine Learning Techniques for Stellar Light Curve Classification
         2. LSTM-RNN and feature detection
         3. 74% accuracy after numerous improvements
   2. Asteroseismology
   3. Machine Learning / Deep Neural Network
   4. Shazam
      1. Shazam analyses the captured sound of a song and finds the name of the song. based on an acoustic fingerprint in a database of millions of songs
      2. Demonstration
         1. See You Again - Charlie Puth
         2. <https://www.youtube.com/watch?v=RgKAFK5djSk&ab_channel=WizKhalifa>
         3. Scriabin Op. 11 No. 11
         4. <https://www.youtube.com/watch?v=nhcuJOt9jJE&ab_channel=newFFL3>
      3. How does it work?
         1. <https://www.toptal.com/algorithms/shazam-it-music-processing-fingerprinting-and-recognition>
         2. Samples digital signal of sound
         3. Splits song into about 40 sections, Fourier Transform
         4. 
         5. Within each interval, identify the frequency with the highest magnitude.
         6. This information forms a signature for this chunk of the song, which when combined, becomes the fingerprint of the song as a whole.
         7. 
         8. Analyze the relative timing of the signatures and compare them to the dataset.
         9. 
2. Purpose/Implication
   1. Speed/Huge Influx of Data
      1. Hubble Space Telescope produced approximately 3 GB per day. James Webb Space Telescope (JWST) expected to produce approximately 57.5 GB per day (Beichman et al. 2014). Square Kilometer Array, which will be online in 2027, is predicted to produce on the order of 109 GB per day It will generate data streams far beyond the total Internet traffic worldwide.
      2. Impossible for humans to pick out and sort through
   2. Hidden Features
      1. As shown by Shazam, many features might not be recognizable by human researchers
   3. Incomplete/Wasted Data
      1. As shown by Shazam, it’s hard for humans or models to recognize incomplete datasets. However, specifically-trained neural networks can do that.
   4. Machine Learning Application in Kepler
      1. Not much ML research has been applied on Kepler datasets. Many related works in supernovae ML projects.
3. Data
   1. Light Curves from MAST (~200000 stars)
   2. Matching labels from kepler stellar 17.csv.gz (~200000 stars)
      1. 
   3. Unlabeled light curves (~2000 stars)
4. Procedure
   1. Part 1: Replicate (Hinners et al., 2018)
   2. Part 2: Improve Classification with Starzam
      1. Fourier Transform + Asteroseismology Feature Extraction
   3. Part 3: Complete Unfinished Kepler Datasets (3200 out of 200000 stars)

<https://solarflare.njit.edu/datasources.html>

Shazam Research Paper

<https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf>

[avery\_wang@alumni.stanford.org](mailto:avery_wang@alumni.stanford.org)

[avery@shazamteam.com](mailto:avery@shazamteam.com)

<https://astronomy.as.virginia.edu/people/profile/jmangum>

Jeff Mangum

[jmangum@nrao.edu](mailto:jmangum@nrao.edu)

434-296-0347

Tuesday November 16, 2021

Contacted Astronomer Jeff Mangum, provided three additional contacts

Michael Lam ([mtlsps@rit.edu](mailto:mtlsps@rit.edu)), Rochester Institute of Technology

Paul Demorest ([pdemores@nrao.edu](mailto:pdemores@nrao.edu)), NRAO Socorro

Scott Ransom ([sransom@nrao.edu](mailto:sransom@nrao.edu)), NRAO Charlottesville

Started constructing/playing with data pipeline on Google Colab

All KIC Stars: <https://archive.stsci.edu/kepler/kepler_fov/search.php>

Search Help: <https://archive.stsci.edu/kepler/kepler_fov/help/search_help.html#kic_seasons>

All Light curve Files: <https://archive.stsci.edu/kepler/data_search/search.php>

Search Help: <https://archive.stsci.edu/search_fields.php?mission=kepler_fov>

Lightkurve Data Pipeline Guide: <http://docs.lightkurve.org/tutorials/1-getting-started/searching-for-data-products.html>

Simbad

<http://simbad.u-strasbg.fr/simbad/sim-fsam>

Querying using astroquery

<https://astroquery.readthedocs.io/en/latest/simbad/simbad.html>

Simbad Classification Table

​​<http://simbad.u-strasbg.fr/simbad/sim-display?data=otypes>

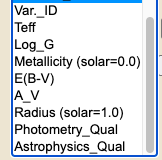
2021 ML Variable Star Paper, 500 Samples using Random Forest, 86% Accuracy

<https://www.ej-physics.org/index.php/ejphysics/article/view/93/62>

Part 1 - Variable Star Classification

1. Find variable star names from simbad with data
2. Find non-variable star names from simbad with data
3. At every star name, download light curve files from MAST KIC >= 1 quarter
4. Split light curves to 1 quarter long, label, save data
5. Train test split 80 20
6. Find

Part 2 - Star characteristic prediction

1. In KIC - download all star statistics
   1. Flags
      1. Data Availability = 2 (Has Archived Data)
      2. Star/Gal\_ID = 0 (Star)
   2. 
   3. Found:
      1. 191449 with labels (Astrophysics Qual = 6)
      2. 14693 w/o labels (Astrophysics Qual = 0)
   4. Photometry Quality
      1. 0 - 24
      2. 1 - 111
      3. 2 - 427
      4. 3 - 638
      5. 4 - 2515
      6. 5 - 191751
      7. 6 - 6930
      8. 7 - 1433
      9. 8 - 2313
2. In LC - download all light curves of max quarter length
   1. Split into each quarter, normalize, remove outliers
   2. Fast fourier transform for frequencies 1-48/day, in increments of 0.01 and binned by mean in stacks of 10
      1. 470 frequency-power data points per quarter
3. Train test split

<https://www.stat.cmu.edu/~cshalizi/350/lectures/19/lecture-19.pdf>

Evaluating Predictive Models

Training Results

1. M1.1
   1. Epoch: 114
   2. Mean Sq. Error, Mean Abs. Error
   3. 0.1885125 0.18122323
2. M1.2
   1. Epoch: 139
   2. Loss: 279649
   3. Mean Sq. Error, Mean Abs. Error
   4. 0.15384701 0.14505906