Machine Learning for Reading Full Form Chinese Characters in Historical NZ Newspapers

ENGR489 2019: Charlene Leong

Team

- Charlene Leong - ENGR489 Student
  charlene.leong@ecs.vuw.ac.nz
- Marcus Frean - ECS Supervisor
  marcus.frean@ecs.vuw.ac.nz
- Dr Sydney Shep - Waiteata Press Reader in Book History
  sydney.shep@vuw.ac.nz
- Rhys Owen - Waiteata Press Technical Lead
  rhys.owen@vuw.ac.nz
- Ya-Wen Ho - Waiteata Press Publication Assistant and Language Expert
  ya-wen.ho@vuw.ac.nz

1 Problem Overview

The Wai-te-ata Press [1] is a letterpress printery established in 1962 at Victoria University of Wellington and acts as an interdisciplinary space to explore books and print in all their myriad forms. The Press' vision of preservation through production produces vibrant, living research that supports our creative heritage.

In 2016, The Press' gratefully accepted the transfer of kaitiakitanga of New Zealand's only surviving Chinese language printing typeface collection, a metric tonne of Chinese full-form character lead type [2] that had been in storage in a farmer's field south of Auckland (Fig 1). This type, ordered from Hong Kong, arrived new in Wellington in 1952 and was used to print the *New Zealand Chinese Growers Monthly Journal*, the organ of the Dominion Federation of New Zealand Chinese Commercial Growers. This journal continued for two decades till 1972, linked market gardeners around the country, published local and international news, featured articles on the civic careers of influential growers, and included Cantonese language lessons. At any one time, every Chinese household was said to be reading one of its 700 monthly copies. The journal was digitised by the Auckland City Libraries [3] in collaboration with the Alexander Turnbull Library, funded by the Chinese Poll Tax Heritage Trust and made publicly available. In 2017, those assets were transferred to the National Library of New Zealand for inclusion in their PapersPast digital collections. However, since the format did not conform to the existing PapersPast [4] standards there were questions around when this important cultural collection would again see the light of digital day. The print collection has now been restored, revitalised and rehoused at the Waiteata Press.

This project is built on lessons learnt from the Qilin project [5], a previous collaboration between the Press' and ECS, using machine learning techniques to perform full form Chinese Optical Character Recognition [OCR] and reveal a window into the twists and turns of lives which have shaped the Chinese communities in New Zealand today. One of the core challenges that arose in getting accurate OCR translation was inconsistency and disfiguration in the printed characters due to deterioration of the types over time, trouble translating different fonts and trouble recognising variants of characters. Throughout the process, the outputs were tested and corrected by a human expert, Ya-Wen Ho, in order to refine the printed type to Unicode translation accuracy. However, this is a very time-intensive task. Therefore, this project aims to prototype a user interface (UI) or user experience (UX) to enable our language expert, Ya-Wen, to validate the transcription accuracy and enable us to obtain an accurate set of labels for the corpus.

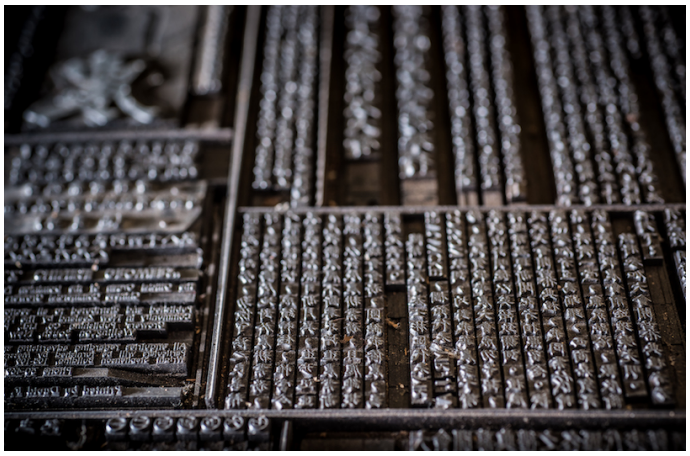


Fig. 1: The typeset for the Growers Journal August 1982 Issue [2]

2 Related Work/Background

The first type restoration project [6] headed by Rhys Owen, technical lead at Waiteata Press, uses OCRopus [7], a collection of document analysis programs for image pre-processing and Google Tesseract [8], Google's open-source OCR engine, for character recognition. This solution was built to run in a Docker machine. However these pre-built solutions had poor accuracy.

The second project, named the Qilin project [5], headed by an ENGR301-302 team, were tasked with categorising characters according to their font, size and type as well as perform frequency analysis of each character in the given data. To process a page of the newspaper there are a number of steps required; finding the characters printed by the type, turning the characters into a machine-readable format and counting the frequency of each occurrence of character. The first challenge in finding the characters is the layout analysis. Due to the newspapers containing a mixture of both Chinese and Latin letter forms, a custom machine learning layout solution was built using a CNN based on the GoogLeNet architecture [9] and class activation mapping [CAM] to classify sections of a newspaper (Fig 2, 3, 4). A subset of images were broken into section classes, to differentiate headings from subheadings from body text in order to generate the training data.

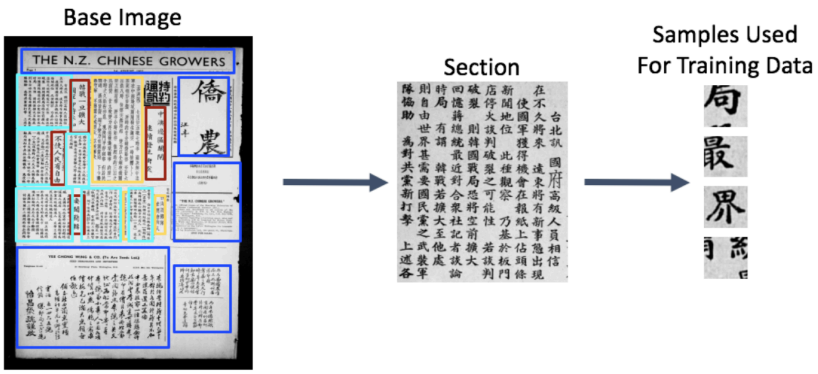


Fig. 2: From Newspaper Scans to Training Data [10]

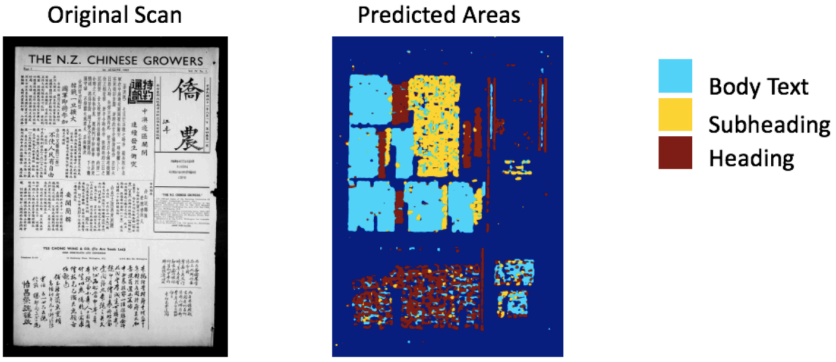


Fig. 3: Combining Prediction Mask from CAM [10]

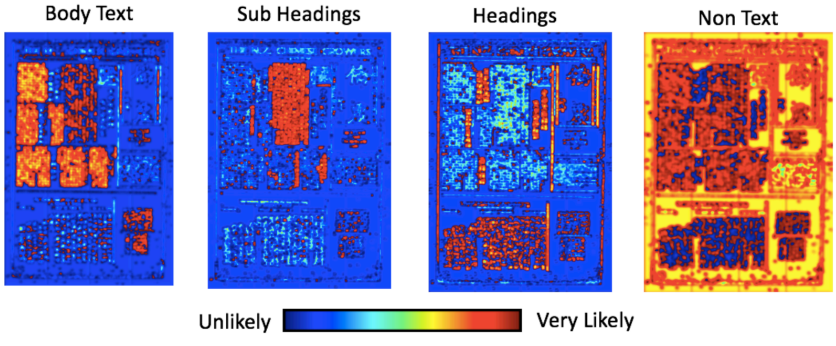


Fig. 4: Individual Prediction Maps from CAM [10]

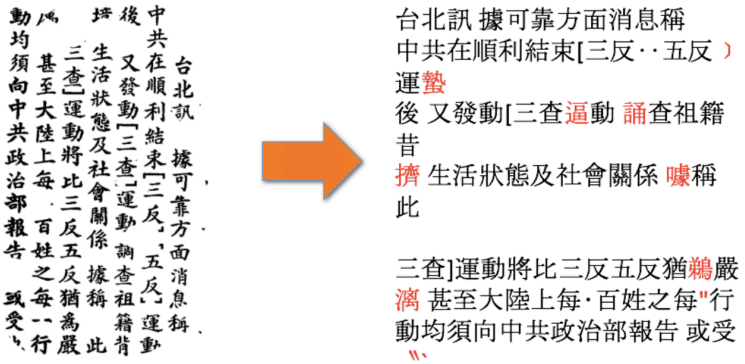


Fig. 5: OCR Results [10]

For the OCR component, Google Tesseract was used to predict 13,000 different characters using training data from Unicode fonts via the type catalogue used to order the original metal types (Fig 6). However, this was only able to achieve achieve 50% accuracy. Potentially, higher accuracy scores could have been achieved by reducing the number of characters selected from the newspaper. However, given Tesseract delivered approximately 80% to 90% accuracy on sections from the newspaper, it was unlikely this was the case.

Since different type sizes and sometimes even fonts were used for different sections, this was an important stage in the workflow. However, while it could recognise the different sizes of characters well, it was not as adept at recognising the characters that were in a different font. This ended up not being that much of an issue given the masking causes the outlier characters to become unrecognisable, meaning the OCR solution would not recognise these as characters.

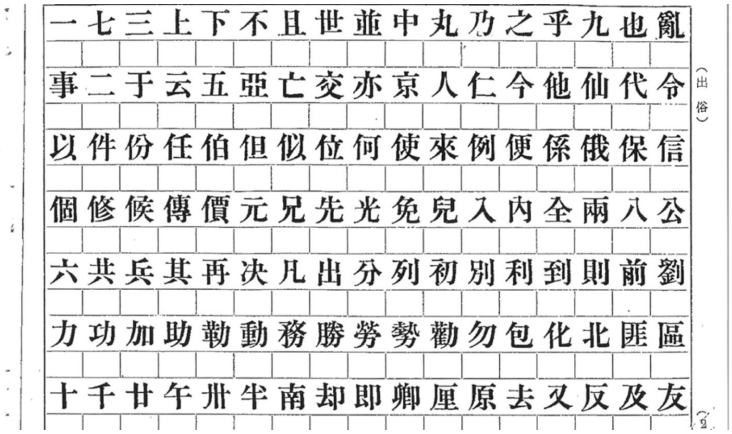
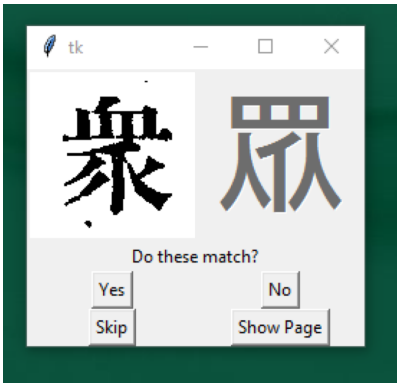
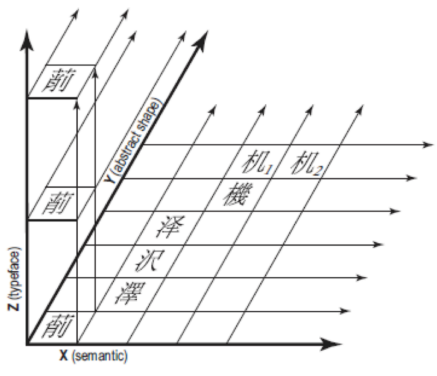


Fig. 6: Type Catalogue [10]

Another major challenge for the team's character checker the existence of multiple variants of a character. For example, the Taiwanese Ministry of Education dictionary of variant forms documents the character for 眾 zhong [a crowd] as historically having 40 variant forms.



(a) Character Checker



(b) 3D Conceptual Model for Variant Forms

In order to classify these variants, the 3D conceptual model [11] developed by the Unicode Consortium for variant unification of Han ideographic character was used. The model expresses written elements in terms of three primary attributes: semantic (meaning, function), abstract shape (general form), and actual shape (instantiated, typeface form). As such, it provides a triangulated variable system that could, in future, assist in character disambiguation, particularly for worn types with their inherent legibility problems.

In conclusion, the outcome of the Qilin project a prototype of the infrastructure of the project, enabling research into Big Data analytic approaches in addition to helping Waiteata Press understand the complexity of the challenges in designing a custom full form Chinese OCR solution for these newspapers due to the nuances in their layout, font and wear and tear. The project helped the Press' resolve a specific problem, but the trajectory of the solution meant to other cultural data contexts and non-Latin languages.

3 Proposed Solution

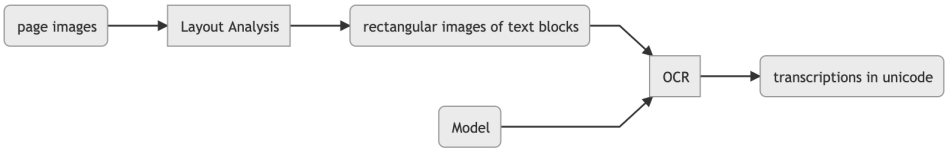


Fig. 7: OCR Pipeline [6]

In order to retrain for a more accurate OCR model, we must first obtain a good training set of Unicode labels and corresponding images. Due to the time intensive task of manual consulting unclassified and misclassified characters, we need to come up with more efficient way to build a training set of labelled data with a human-in-the-loop approach (Fig 7).

The proposed solution involves categorising all extracted characters from the newspapers into image buckets with their corresponding Unicode labels which are then used to train a new model (Fig 10a, 10b, 10c). Images from the newspapers will be obtained using the section classifier from the Qilin project. Further research will have to be done in further extracting individual characters from the sections. Depending on how accurate the buckets are, there is potential for some reordering of misclassified characters before retraining. We are then planning to use the feature maps from the retrained model and apply dimensionality reduction techniques such as *principal component analysis (PCA)* and *t-distributed stochastic neighbor embedding (t-SNE)* to visually explore the data by clustering images by feature similarity and therefore evaluating the quality of the classifier (Fig 10d). Each t-SNE cluster can then be explored further in 2D grid format where Ya-Wen can easily identify misclassified images in the cluster by implementing an appropriate tagging schema (Fig 10e). Misclassified images will then either be re-assigned to their appropriate cluster or a new cluster if unlabelled. The goal is use this technique in a iterative fashion to obtain enough accurately Unicode labelled characters to train a model that is capable of transcribing newspaper scans of the *New Zealand Chinese Growers Monthly Journal* with minimal errors (Fig 9).

It is intended that the user experience (UX) will dynamically change as the model and the clusters are recorrected which requires some further thought into the infrastructure capable of the compute and storage required for this application. The current intention is to use Victoria University's high performance computing cluster Rāpoi [12] intended for research purposes. Other options include the Google Cloud Platform [13] or Amazon Web Services as Rāpoi has ability to connect to popular cloud providers via their SDKs.

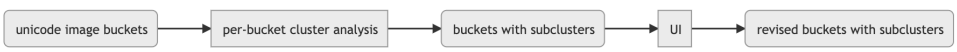
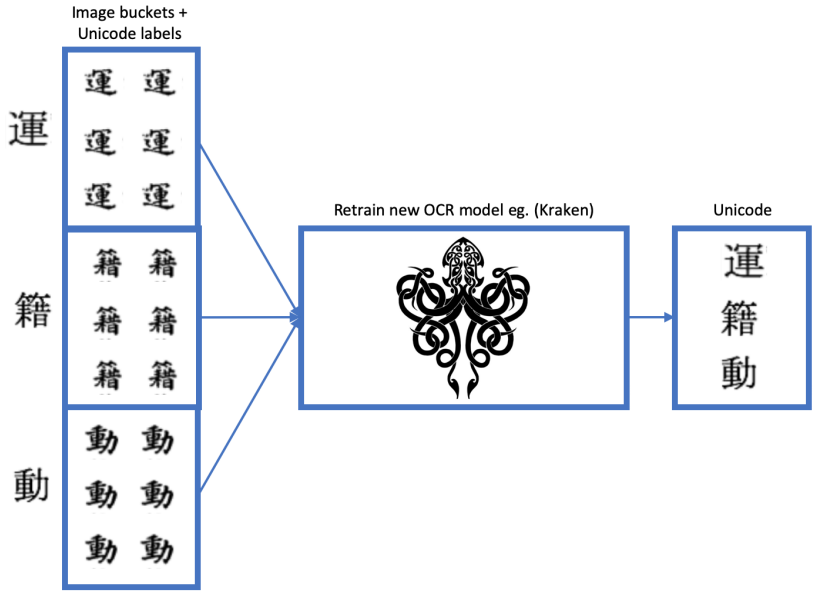


Fig. 8: Unicode Character Clustering Pipeline [6]

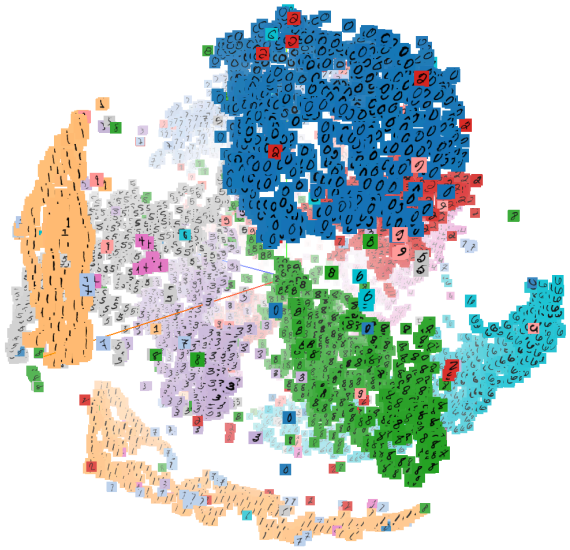
Fig. 9: Retraining a New Model [6]

(a) Step 1: Extracting initial training data and labels using Google Tesseract

(b) Step 2: Categorise all image buckets with corresponding Unicode label

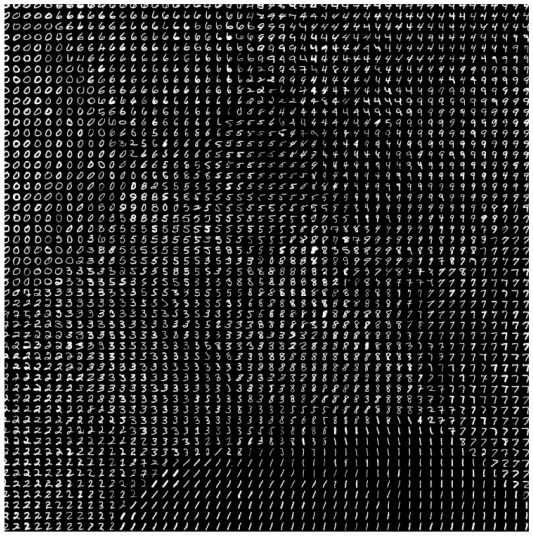


(c) Step 3: Use image buckets with Unicode label to retrain a new model



(d) Step 4: Use t-SNE to visually explore buckets to identify outliers

Fig. 10: Proposed Solution



(e) Step 5: Design UX using t-SNE grid with tagging schema to recluster data

Fig. 10: Proposed Solution

4 Proposed Evaluation

Evaluation in this project can be broken down into two parts: evaluating the quality of the classifier itself and evaluating the user experience for Ya-Wen. Metrics to consider in evaluating the classifier include:

- Confusion Matrix - Error matrix that is often used to describe the performance of a classifier.

	Negative (predicted)	Positive (predicted)
Negative (actual)	true negative	false positive
Positive (actual)	false negative	true positive

- Accuracy - how many correct out of all instances, however is not sufficient when there is significant class imbalance

$$Accuracy = \frac{truepositives + truenegatives}{totalinstances}$$

- Precision - How many of the instances found were correct

$$Precision = \frac{truepositives}{truepositives + falsepositives}$$

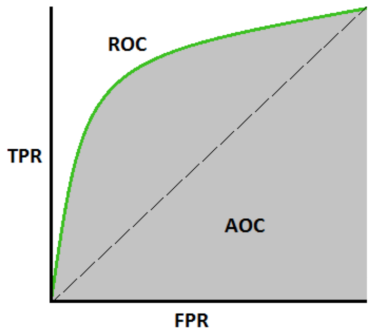
- Recall - How many correct instances were found out of all the correct instances

$$Recall = \frac{truepositives}{truepositives + falsenegatives}$$

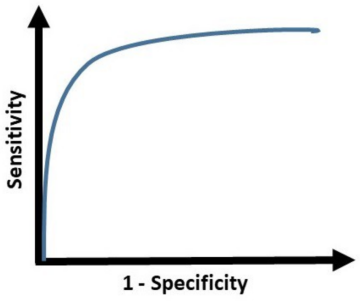
- F1 Score - Overall measure of a model's accuracy that combines precision and recall

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

- AUC-ROC Curve



(a) TP rate vs FP rate



(b) Sensitivity vs Specificity

Fig. 11: AUC ROC Curve

Metrics to consider in evaluating the user experience for Ya-Wen include:

- The time taken to evaluate a cluster
- The number of clicks required to evaluate a cluster
- The number of unclassified or misclassified instances in a cluster
- Combinations of the metrics identified above such as time taken to evaluate a cluster by number of unclassified or misclassified instances
- Qualitative survey after each user test to critique and improve the user experience and friendliness and intuition of the user interface design

Model metrics from Google Tesseract or Qilin would be an appropriate baseline. Appropriate benchmarks for user experience metrics have yet to be decided and will be further explored in the literature review.

Glossary

principal component analysis (PCA)

A dimensionality reduction technique used to emphasise variation and bring out strong patterns in a dataset by reducing a dataset down to its most principal components. It's often used to make data easy to explore and visualise. See [14] for visual explanation.

t-distributed stochastic neighbor embedding (t-SNE)

A dimensionality reduction technique that is particularly well suited for the visualisation of high-dimensional datasets. See [15] for visual explanation.

References

[1] VUW. (2019, mar) Wai-te-ata press. [Online]. Available: https://www.victoria.ac.nz/wtapress (Cited on page 1.)

[2] E. Ng. (2016, sep) The single object: A metric tonne of chinese-new zealand history. [Online]. Available: https://thespinoff.co.nz/partner/objectspace/16-09-2018/the-single-object-a-metric-tonne-of-chinese-new-zealand-history/ (Cited on page 1.)

[3] A. Council. (2019, mar) Nz chinese journals. [Online]. Available: https://nzchinesejournals.org.nz/ (Cited on page 1.)

[4] N. L. of NZ. (2019, mar) Paperspast. [Online]. Available: https://paperspast.natlib.govt.nz/ (Cited on page 1.)

[5] Gitlab. (2018, mar) Qilin project. [Online]. Available: https://gitlab.com/wai-te-ata-press/qilin (Cited on pages 1 and 2.)

[6] R. Owen. (2017, mar) Gitlab: Type restoration. [Online]. Available: https://gitlab.com/wai-te-ata-press/type-restoration/issues/19 (Cited on pages 2 and 4.)

[7] tmbdev. (2018, mar) Ocropus. [Online]. Available: https://github.com/tmbdev/ocropy (Cited on page 2.)

[8] Google. (2018, mar) Tesseract. [Online]. Available: https://github.com/tesseract-ocr/ (Cited on page 2.)

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842 (Cited on page 2.)

[10] D. S. S. e. Rhys Owen, Ya-Wen Ho. (2018, nov) Europeanatech - issue 10: Innovation agenda. [Online]. Available: https://pro.europeana.eu/page/issue-10-innovation-agenda#wrestling-with-qilin (Cited on pages 2 and 3.)

[11] U. Consortium. (2017, mar) The unicode standard, version 10.0: East asia. [Online]. Available: http://www.unicode.org/versions/Unicode10.0.0/ch18.pdf (Cited on page 4.)

[12] VUW. (2019, apr) Rapoi hpc cluster documentation. [Online]. Available: https://eresearchsandpit.github.io/vuwrc/ (Cited on page 4.)

[13] G. Cloud. (2019, apr) Google cloud platform. [Online]. Available: https://cloud.google.com/ (Cited on page 4.)

[14] V. Powell. (2019, mar) Setosa.io: Principal component analysis. [Online]. Available: http://setosa.io/ev/principal-component-analysis/ (Cited on page 8.)

[15] C. Rossant. (2015, mar) An illustrated introduction to the t-sne algorithm. [Online]. Available: https://www.oreilly.com/learning/an-illustrated-introduction-to-the-t-sne-algorithm (Cited on page 8.)