Virginia Plant Image Classification

Huilin Chang

Introduction

- Virginians enjoy more than 3,000 square miles of waterways across the state
- Aquatic plants play a major role in their environmental health

* Photos used in this presentation were either taken by one of the group members or taken from our dataset; all images used in the dataset were from sources that allow use for educational purposes

Introduction

- Invasive species of aquatic plants threaten the waterway’s health, suffocating native plants, harming fish and aquatic organism populations, changing the water’s chemistry – and making it harder for humans to enjoy swimming, boating, and fishing

- Hydrilla is one of the most widespread invasive aquatic plant species in VA

- Government and community entities have to expend monetary and human resources to identify and eradicate it

Motivation

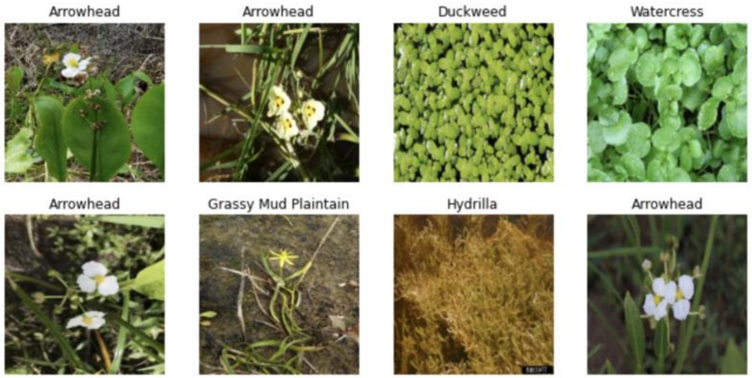
- Unfortunately, hydrilla can sometimes be hard to distinguish from other types of aquatic plants that are "healthy" native plant species

- We built and developed 3 transfer learning/CNN models for image classification of 5 different types of aquatic plant species (hydrilla, arrowhead, duckweed, grassy mud plantain, and watercress)

- Image classification can reduce costs and increase efficiencies when identifying bodies of water where intervention to reduce invasive species is needed

Data Collection & Data

- Dataset
  - Image sources: invasive.org, Google, gbif.org, Shutterstock
  - 450 Images - 5 Aquatic Plants
    - Invasive
      - Hydrilla (101)
    - Non-invasive
      - Duckweed (98), Watercress (100), Arrowhead (76), Grassy Mud Plantain (75)
- Data Split
  - Train: 0.8   Validation: 0.1   Test: 0.1



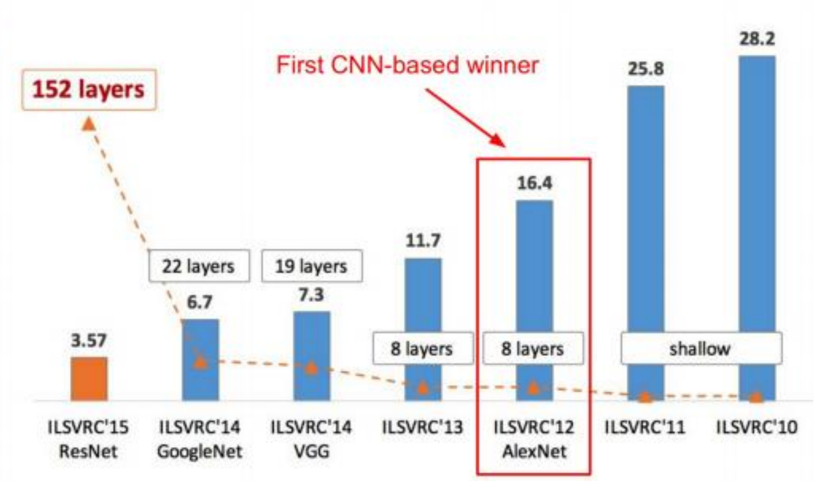
Data Preprocessing

- Uploaded dataset to Google Drive and mounted to Google Colab
- Data Processing
  - Rescaled all images (224, 224)
  - Random Image Augmentation
    - Flip
    - Rotate
    - Contrast
    - Zoom

Winners of the ImageNet Challenges

Looking at the evolution of the ImageNet winning entries is a good way to understand how CNNs work.

- AlexNet (2012 winner)
- GoogLeNet (2014 winner)
- ResNet (2015 winner)

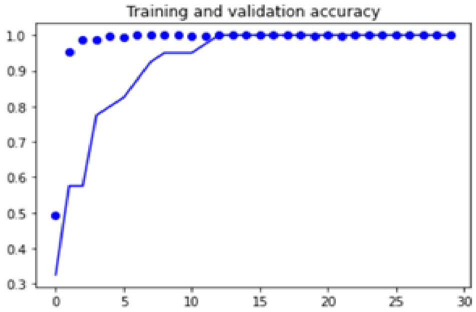


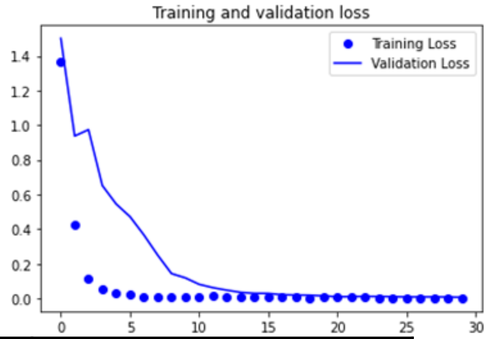
Initial Experiments

- Custom Neural Network
  - Custom CNN
- Transfer Learning
  - ***Xception***
  - ***DenseNet 121***
  - VGG19
  - ***EfficientNet - B6***

Xception

- Fewer parameters and computations than a regular convolution layer yet has better performance
- Added additional BatchNormalization layer
- **Optimizer**: SGD **LR**: 0.01
- 5 prediction class

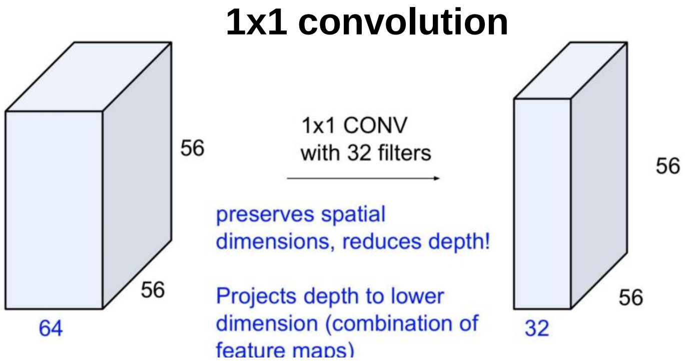

Training and validation accuracy


Training and validation loss

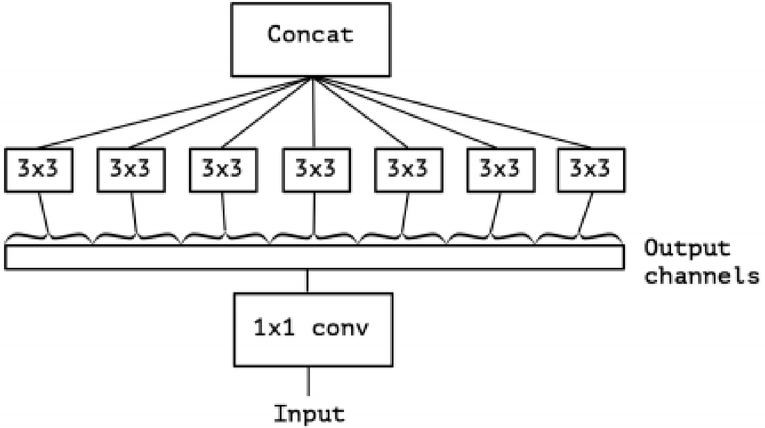
	Training	Validation	Testing
Accuracy	99.00%	99.00%	95.56%
Loss	0.0130	0.0119	0.0977

Xception Architecture

- Inception (network within network)
- Use 1x1 convolutions to reduce feature depth
- Apply parallel operations on the input from the previous layer

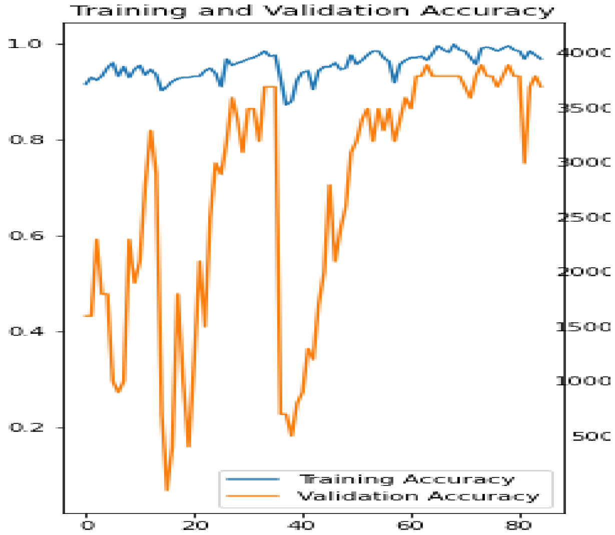


1x1 convolution



DenseNet 121

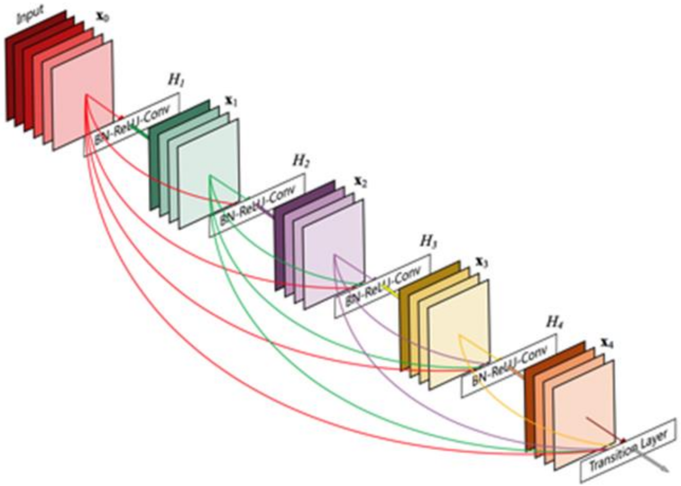
- 121 layers within four “dense blocks”
- Experimented with early stopping and drop-out; did not improve performance
- **Optimizer**: Adam **LR**: 0.001
- Epochs: 85



	Training	**Validation**	**Testing**
Accuracy	96.37%	90.91%	93.33%
Loss	0.2803	0.9246	0.7544

DenseNet

- The idea behind dense convolutional networks is simple: **it may be useful to reference feature maps from earlier in the network**. Thus, each layer's feature map is concatenated to the input of *every successive layer* within a dense block. This allows later layers within the network to *directly* leverage the features from earlier layers, encouraging feature reuse within the network. The authors state, "concatenating feature-maps learned by *different layers* increases variation in the input of subsequent layers and improves efficiency."



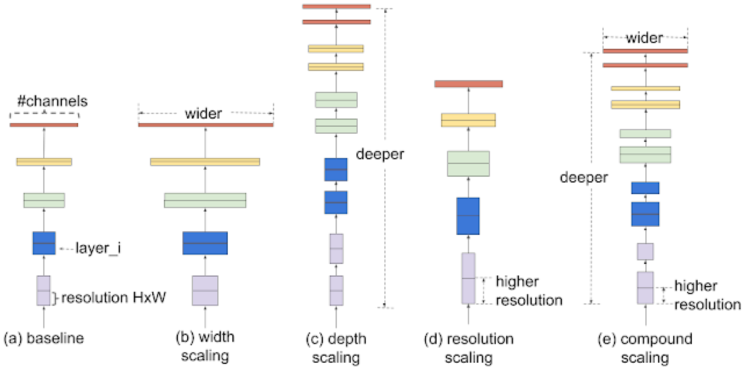
EfficientNet-B6

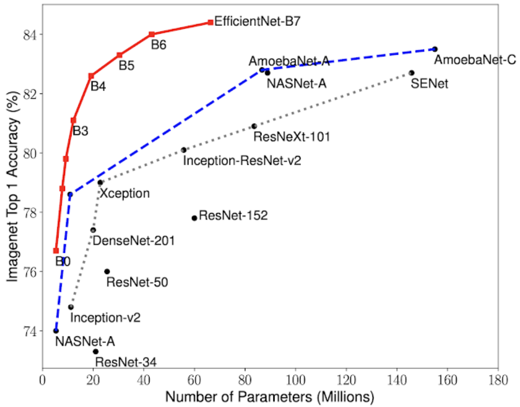
- EfficientNet is more accurate and efficient than past CNNs under resource constraints
- Added additional layers
  - GlobalAveragePooling2D
  - BatchNormalization
  - Dropout
- **Optimizer**: SGD **LR**: 0.01
- **Callbacks**: EarlyStopping, ReduceLROnPlateau

	Training	Validation	Testing
Accuracy	98.44%	100.00%	97.78%
Loss	0.0562	0.0067	0.1106

EfficientNet

- Model performance based on the study of the impact of scaling different dimensions of the model
- Performing a neural architecture search using the AutoML MNAS framework which optimizes both accuracy and efficiency



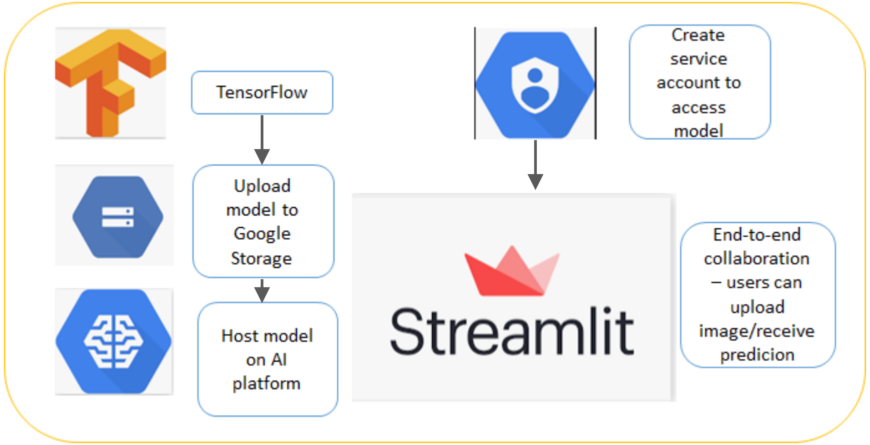


Comparison

Model	Testing Accuracy
Xception	95.56%
DenseNet 121	93.33%
EfficientNet - B6	97.78%

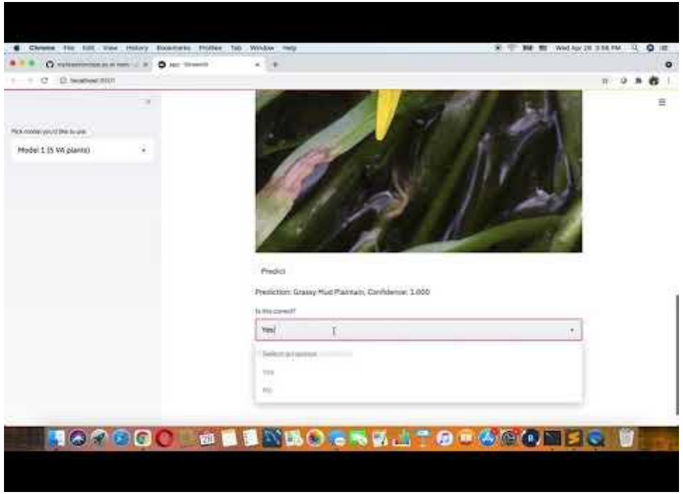
Model Deployment

- Deployed all models to Google Cloud Platform



TensorFlow
Upload model to Google Storage
Host model on AI platform
Create service account to access model
Streamlit
End-to-end collaboration – users can upload image/receive predicion

Xception Streamlit App Demo



Pick model you'd like to use

Model 1 (5 VA plants)

Predict

Prediction: Grassy Mud Plaintain, Confidence: 1.000

Is this correct?

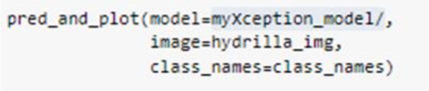
Yes

Select an option

Yes

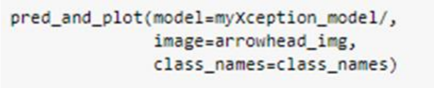
No

Model Prediction



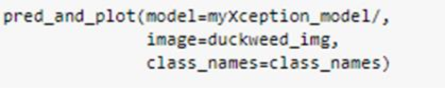
```
pred_and_plot(model=myXception_model/,
              image=hydrilla_img,
              class_names=class_names)
```

Hydrilla, prob: 0.79



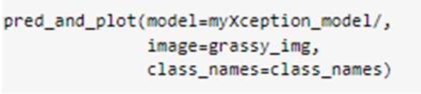
```
pred_and_plot(model=myXception_model/,
              image=arrowhead_img,
              class_names=class_names)
```

Arrowhead, prob: 1.00



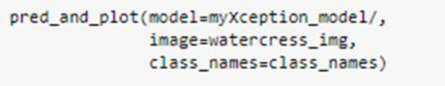
```
pred_and_plot(model=myXception_model/,
              image=duckweed_img,
              class_names=class_names)
```

Duckweed, prob: 1.00



```
pred_and_plot(model=myXception_model/,
              image=grassy_img,
              class_names=class_names)
```

Grassy Mud Plaintain, prob: 1.00



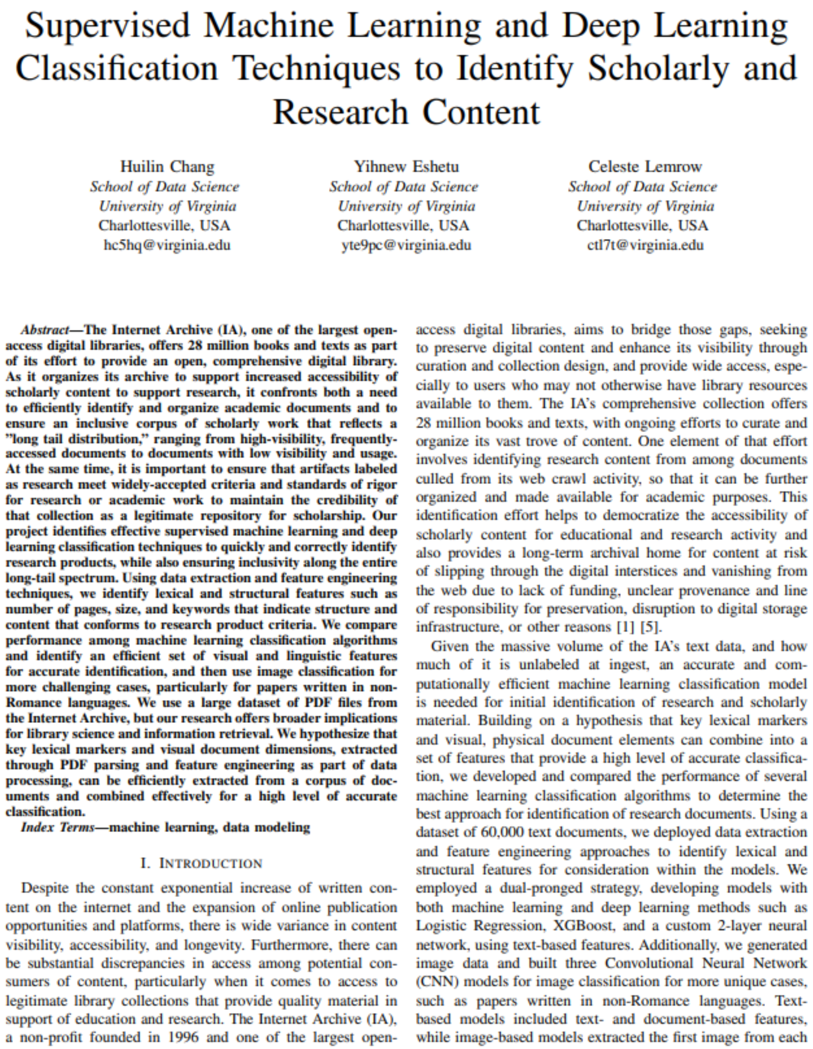
```
pred_and_plot(model=myXception_model/,
              image=watercress_img,
              class_names=class_names)
```

Watercress, prob: 1.00

Conclusions

- Our models are able to distinguish 5 classes of plants, including Hydrilla
- Model accuracy indicates sufficient reliability (more limited risk in incorrectly identifying a plant as invasive, resulting in a “healthy” plant being eradicated from a waterway)
- Models provide a platform to more efficiently identify where invasive species are located in waterways through crowdsourcing of photos, rather than limited in-person inspections
- Efficiencies on identification allows for more resources to be allocated directly to intervention
- Future work can expand the training set to improve accuracy and generalizability, include more types of aquatic plant species, and create an app for wider citizen participation

IEEE SIEDS 21’ Publication



Supervised Machine Learning and Deep Learning Classification Techniques to Identify Scholarly and Research Content

Huilin Chang
School of Data Science
University of Virginia
Charlottesville, USA
hc5hq@virginia.edu

Yihnew Eshetu
School of Data Science
University of Virginia
Charlottesville, USA
yte9pc@virginia.edu

Celeste Lemrow
School of Data Science
University of Virginia
Charlottesville, USA
ctl7t@virginia.edu

***Abstract*—The Internet Archive (IA), one of the largest open-access digital libraries, offers 28 million books and texts as part of its effort to provide an open, comprehensive digital library. As it organizes its archive to support increased accessibility of scholarly content to support research, it confronts both a need to efficiently identify and organize academic documents and to ensure an inclusive corpus of scholarly work that reflects a ”long tail distribution,” ranging from high-visibility, frequently-accessed documents to documents with low visibility and usage. At the same time, it is important to ensure that artifacts labeled as research meet widely-accepted criteria and standards of rigor for research or academic work to maintain the credibility of that collection as a legitimate repository for scholarship. Our project identifies effective supervised machine learning and deep learning classification techniques to quickly and correctly identify research products, while also ensuring inclusivity along the entire long-tail spectrum. Using data extraction and feature engineering techniques, we identify lexical and structural features such as number of pages, size, and keywords that indicate structure and content that conforms to research product criteria. We compare performance among machine learning classification algorithms and identify an efficient set of visual and linguistic features for accurate identification, and then use image classification for more challenging cases, particularly for papers written in non-Romance languages. We use a large dataset of PDF files from the Internet Archive, but our research offers broader implications for library science and information retrieval. We hypothesize that key lexical markers and visual document dimensions, extracted through PDF parsing and feature engineering as part of data processing, can be efficiently extracted from a corpus of documents and combined effectively for a high level of accurate classification.**

***Index Terms*—machine learning, data modeling**

I. Introduction

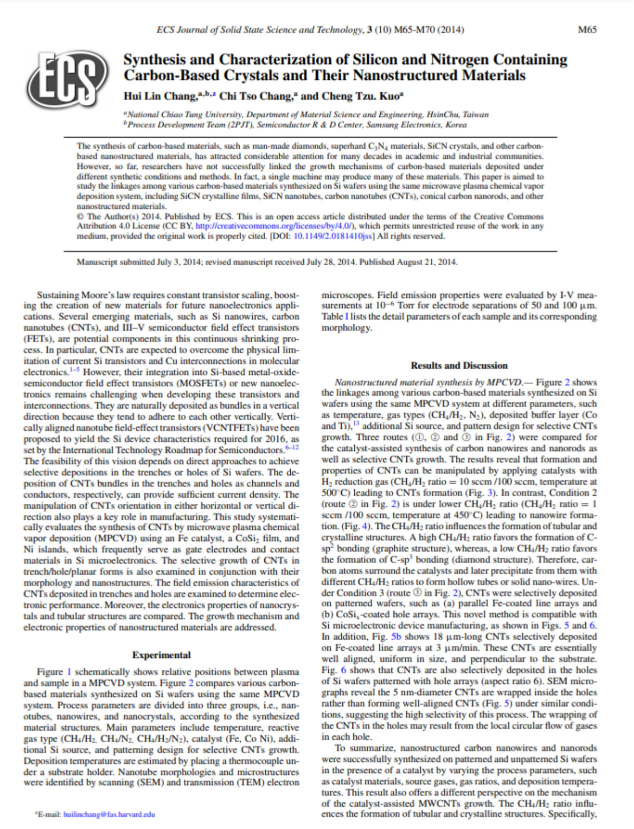
Despite the constant exponential increase of written content on the internet and the expansion of online publication opportunities and platforms, there is wide variance in content visibility, accessibility, and longevity. Furthermore, there can be substantial discrepancies in access among potential consumers of content, particularly when it comes to access to legitimate library collections that provide quality material in support of education and research. The Internet Archive (IA), a non-profit founded in 1996 and one of the largest open-access digital libraries, aims to bridge those gaps, seeking to preserve digital content and enhance its visibility through curation and collection design, and provide wide access, especially to users who may not otherwise have library resources available to them. The IA’s comprehensive collection offers 28 million books and texts, with ongoing efforts to curate and organize its vast trove of content. One element of that effort involves identifying research content from among documents culled from its web crawl activity, so that it can be further organized and made available for academic purposes. This identification effort helps to democratize the accessibility of scholarly content for educational and research activity and also provides a long-term archival home for content at risk of slipping through the digital interstices and vanishing from the web due to lack of funding, unclear provenance and line of responsibility for preservation, disruption to digital storage infrastructure, or other reasons [1] [5].

Given the massive volume of the IA’s text data, and how much of it is unlabeled at ingest, an accurate and computationally efficient machine learning classification model is needed for initial identification of research and scholarly material. Building on a hypothesis that key lexical markers and visual, physical document elements can combine into a set of features that provide a high level of accurate classification, we developed and compared the performance of several machine learning classification algorithms to determine the best approach for identification of research documents. Using a dataset of 60,000 text documents, we deployed data extraction and feature engineering approaches to identify lexical and structural features for consideration within the models. We employed a dual-pronged strategy, developing models with both machine learning and deep learning methods such as Logistic Regression, XGBoost, and a custom 2-layer neural network, using text-based features. Additionally, we generated image data and built three Convolutional Neural Network (CNN) models for image classification for more unique cases, such as papers written in non-Romance languages. Text-based models included text- and document-based features, while image-based models extracted the first image from each

Presentation Outline

- Problem Statement
- Data Overview
- Data Pipeline
- Data Feature Engineering
- Exploratory Data Analysis
- Models
- Conclusion

Research paper example



ECS Journal of Solid State Science and Technology, **3** (10) M65-M70 (2014) M65

Synthesis and Characterization of Silicon and Nitrogen Containing Carbon-Based Crystals and Their Nanostructured Materials

Hui Lin Chang,[a,b,z] Chi Tso Chang,[a] and Cheng Tzu. Kuo[a]

[a] *National Chiao Tung University, Department of Material Science and Engineering, HsinChu, Taiwan*
[b] *Process Development Team (2PJT), Semiconductor R & D Center, Samsung Electronics, Korea*

The synthesis of carbon-based materials, such as man-made diamonds, superhard C_3N_4 materials, SiCN crystals, and other carbon-based nanostructured materials, has attracted considerable attention for many decades in academic and industrial communities. However, so far, researchers have not successfully linked the growth mechanisms of carbon-based materials deposited under different synthetic conditions and methods. In fact, a single machine may produce many of these materials. This paper is aimed to study the linkages among various carbon-based materials synthesized on Si wafers using the same microwave plasma chemical vapor deposition system, including SiCN crystalline films, SiCN nanotubes, carbon nanotubes (CNTs), conical carbon nanorods, and other nanostructured materials.

Manuscript submitted July 3, 2014; revised manuscript received July 28, 2014. Published August 21, 2014.

Sustaining Moore's law requires constant transistor scaling, boosting the creation of new materials for future nanoelectronics applications. Several emerging materials, such as Si nanowires, carbon nanotubes (CNTs), and III–V semiconductor field effect transistors (FETs), are potential components in this continuous shrinking process. In particular, CNTs are expected to overcome the physical limitation of current Si transistors and Cu interconnections in molecular electronics.[1–5] However, their integration into Si-based metal-oxide-semiconductor field effect transistors (MOSFETs) or new nanoelectronics remains challenging when developing these transistors and interconnections. They are naturally deposited as bundles in a vertical direction because they tend to adhere to each other vertically. Vertically aligned nanotube field-effect transistors (VCNTFETs) have been proposed to yield the Si device characteristics required for 2016, as set by the International Technology Roadmap for Semiconductors.[6–12] The feasibility of this vision depends on direct approaches to achieve selective depositions in the trenches or holes of Si wafers. The deposition of CNTs bundles in the trenches and holes as channels and conductors, respectively, can provide sufficient current density. The manipulation of CNTs orientation in either horizontal or vertical direction also plays a key role in manufacturing. This study systematically evaluates the synthesis of CNTs by microwave plasma chemical vapor deposition (MPCVD) using an Fe catalyst, a $CoSi_2$ film, and Ni islands, which frequently serve as gate electrodes and contact materials in Si microelectronics. The selective growth of CNTs in trench/hole/planar forms is also examined in conjunction with their morphology and nanostructures. The field emission characteristics of CNTs deposited in trenches and holes are examined to determine electronic performance. Moreover, the electronics properties of nanocrystals and tubular structures are compared. The growth mechanism and electronic properties of nanostructured materials are addressed.

Experimental

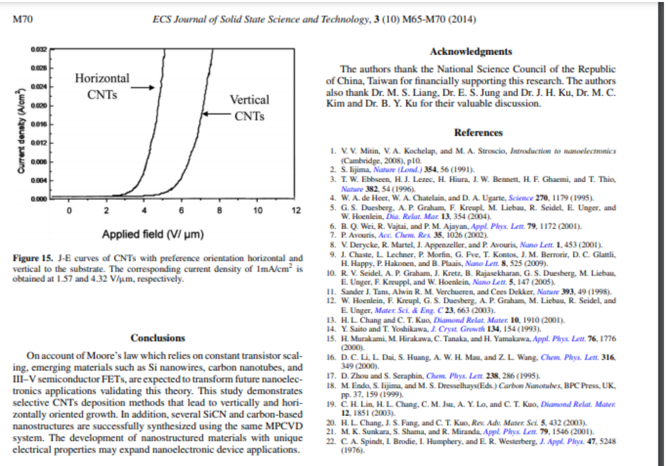
Figure 1 schematically shows relative positions between plasma and sample in a MPCVD system. Figure 2 compares various carbon-based materials synthesized on Si wafers using the same MPCVD system. Process parameters are divided into three groups, i.e., nanotubes, nanowires, and nanocrystals, according to the synthesized material structures. Main parameters include temperature, reactive gas type (CH_4/H_2, CH_4/N_2, $CH_4/H_2/N_2$), catalyst (Fe, Co Ni), additional Si source, and patterning design for selective CNTs growth. Deposition temperatures are estimated by placing a thermocouple under a substrate holder. Nanotube morphologies and microstructures were identified by scanning (SEM) and transmission (TEM) electron microscopes. Field emission properties were evaluated by I-V measurements at 10^{-6} Torr for electrode separations of 50 and 100 µm. Table I lists the detail parameters of each sample and its corresponding morphology.

Results and Discussion

Nanostructured material synthesis by MPCVD.— Figure 2 shows the linkages among various carbon-based materials synthesized on Si wafers using the same MPCVD system at different parameters, such as temperature, gas types (CH_4/H_2, N_2), deposited buffer layer (Co and Ti),[13] additional Si source, and pattern design for selective CNTs growth. Three routes (①, ② and ③ in Fig. 2) were compared for the catalyst-assisted synthesis of carbon nanowires and nanorods as well as selective CNTs growth. The results reveal that formation and properties of CNTs can be manipulated by applying catalysts with H_2 reduction gas (CH_4/H_2 ratio = 10 sccm /100 sccm, temperature at 500°C) leading to CNTs formation (Fig. 3). In contrast, Condition 2 (route ② in Fig. 2) is under lower CH_4/H_2 ratio (CH_4/H_2 ratio = 1 sccm /100 sccm, temperature at 450°C) leading to nanowire formation. (Fig. 4). The CH_4/H_2 ratio influences the formation of tubular and crystalline structures. A high CH_4/H_2 ratio favors the formation of C-sp^2 bonding (graphite structure), whereas, a low CH_4/H_2 ratio favors the formation of C-sp^3 bonding (diamond structure). Therefore, carbon atoms surround the catalysts and later precipitate from them with different CH_4/H_2 ratios to form hollow tubes or solid nano-wires. Under Condition 3 (route ③ in Fig. 2), CNTs were selectively deposited on patterned wafers, such as (a) parallel Fe-coated line arrays and (b) $CoSi_x$-coated hole arrays. This novel method is compatible with Si microelectronic device manufacturing, as shown in Figs. 5 and 6. In addition, Fig. 5b shows 18 µm-long CNTs selectively deposited on Fe-coated line arrays at 3 µm/min. These CNTs are essentially well aligned, uniform in size, and perpendicular to the substrate. Fig. 6 shows that CNTs are also selectively deposited in the holes of Si wafers patterned with hole arrays (aspect ratio 6). SEM micrographs reveal the 5 nm-diameter CNTs are wrapped inside the holes rather than forming well-aligned CNTs (Fig. 5) under similar conditions, suggesting the high selectivity of this process. The wrapping of the CNTs in the holes may result from the local circular flow of gases in each hole.

To summarize, nanostructured carbon nanowires and nanorods were successfully synthesized on patterned and unpatterned Si wafers in the presence of a catalyst by varying the process parameters, such as catalyst materials, source gases, gas ratios, and deposition temperatures. This result also offers a different perspective on the mechanism of the catalyst-assisted MWCNTs growth. The CH_4/H_2 ratio influences the formation of tubular and crystalline structures. Specifically,

[z] E-mail: huilinchang@fas.harvard.edu



M70 ECS Journal of Solid State Science and Technology, **3** (10) M65-M70 (2014)

Figure 15. J-E curves of CNTs with preference orientation horizontal and vertical to the substrate. The corresponding current density of $1mA/cm^2$ is obtained at 1.57 and 4.32 V/µm, respectively.

Conclusions

On account of Moore's law which relies on constant transistor scaling, emerging materials such as Si nanowires, carbon nanotubes, and III–V semiconductor FETs, are expected to transform future nanoelectronics applications validating this theory. This study demonstrates selective CNTs deposition methods that lead to vertically and horizontally oriented growth. In addition, several SiCN and carbon-based nanostructures are successfully synthesized using the same MPCVD system. The development of nanostructured materials with unique electrical properties may expand nanoelectronic device applications.

Acknowledgments

The authors thank the National Science Council of the Republic of China, Taiwan for financially supporting this research. The authors also thank Dr. M. S. Liang, Dr. E. S. Jung and Dr. J. H. Ku, Dr. M. C. Kim and Dr. B. Y. Ku for their valuable discussion.

References

1. V. V. Mitin, V. A. Kochelap, and M. A. Stroscio, *Introduction to nanoelectronics* (Cambridge, 2008), p10.
2. S. Iijima, *Nature (Lond.)* **354**, 56 (1991).
3. T. W. Ebbseen, H. J. Lezec, H. Hiura, J. W. Bennett, H. F. Ghaemi, and T. Thio, *Nature* **382**, 54 (1996).
4. W. A. de Heer, W. A. Chatelain, and D. A. Ugarte, *Science* **270**, 1179 (1995).
5. G. S. Duesberg, A. P. Graham, F. Kreupl, M. Liebau, R. Seidel, E. Unger, and W. Hoenlein, *Dia. Relat. Mar.* **13**, 354 (2004).
6. B. Q. Wei, R. Vajtai, and P. M. Ajayan, *Appl. Phys. Lett.* **79**, 1172 (2001).
7. P. Avouris, *Acc. Chem. Res.* **35**, 1026 (2002).
8. V. Derycke, R. Martel, J. Appenzeller, and P. Avouris, *Nano Lett.* **1**, 453 (2001).
9. J. Chaste, L. Lechner, P. Morfin, G. Fve, T. Kontos, J. M. Berrorir, D. C. Glattli, H. Happy, P. Hakonen, and B. Plaais, *Nano Lett.* **8**, 525 (2009).
10. R. V. Seidel, A. P. Graham, J. Kretz, B. Rajasekharan, G. S. Duesberg, M. Liebau, E. Unger, F. Kreuppl, and W. Hoenlein, *Nano Lett.* **5**, 147 (2005).
11. Sander J. Tans, Alwin R. M. Verchueren, and Cees Dekker, *Nature* **393**, 49 (1998).
12. W. Hoenlein, F. Kreupl, G. S. Duesberg, A. P. Graham, M. Liebau, R. Seidel, and E. Unger, *Mater. Sci. & Eng. C* **23**, 663 (2003).
13. H. L. Chang and C. T. Kuo, *Diamond Relat. Mater.* **10**, 1910 (2001).
14. Y. Saito and T. Yoshikawa, *J. Cryst. Growth* **134**, 154 (1993).
15. H. Murakami, M. Hirakawa, C. Tanaka, and H. Yamakawa, *Appl. Phys. Lett.* **76**, 1776 (2000).
16. D. C. Li, L. Dai, S. Huang, A. W. H. Mau, and Z. L. Wang, *Chem. Phys. Lett.* **316**, 349 (2000).
17. D. Zhou and S. Seraphin, *Chem. Phys. Lett.* **238**, 286 (1995).
18. M. Endo, S. Iijima, and M. S. Dresselhays(Eds.) *Carbon Nanotubes*, BPC Press, UK, pp. 37, 159 (1999).
19. C. H. Lin, H. L. Chang, C. M. Jsu, A. Y. Lo, and C. T. Kuo, *Diamond Relat. Mater.* **12**, 1851 (2003).
20. H. L. Chang, J. S. Fang, and C. T. Kuo, *Rev. Adv. Mater. Sci.* **5**, 432 (2003).
21. M. K. Sunkara, S. Shama, and R. Miranda, *Appl. Phys. Lett.* **79**, 1546 (2001).
22. C. A. Spindt, I. Brodie, I. Humphery, and E. R. Westerberg, *J. Appl. Phys.* **47**, 5248 (1976).

Research paper or not

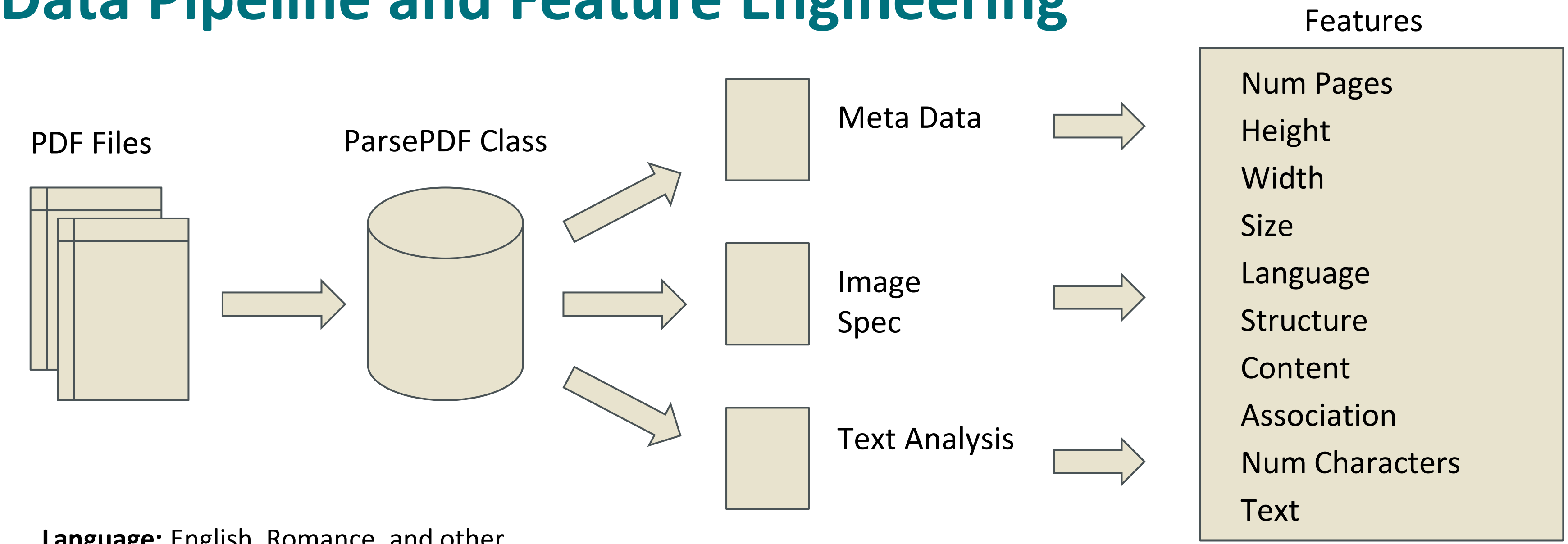
Problem Statement

- One of the Internet Archive’s mission areas is “Universal Access to All Knowledge”, which includes collecting and providing access to the “scholarly web” -- research publications and datasets
- Curation to accurately identify legitimate research publications is needed to help users find scholarly content
- An inclusive approach that accounts for diverse content, particularly from underrepresented geographic areas, groups, and content domains, is important to avoid excluding relevant content due to implicit bias and narrow criteria
- Our project aims to help this mission by implementing a fast PDF identification tool, which will score files on their likelihood of being a research publication

Data Overview

- 4 IA Training Datasets
  - Global Wayback Random - Random sampled PDFs from the Wayback Machine
  - Fatcat - A set of PDFs from the existing ‘Fatcat’ catalog of research papers
  - Fatcat Longtail Language - Papers from less-represented languages
  - Longtail - A set of PDFs created using heuristics (GROBID)
- Minor issues with data
  - Encrypted PDFs
  - Corrupted/Unparsable
- Plan to branch out further into IA content archives as well as other known sources of PDF scholarly documents

Data Pipeline and Feature Engineering



PDF Files

ParsePDF Class

Meta Data

Image Spec

Text Analysis

Features

Num Pages
Height
Width
Size
Language
Structure
Content
Association
Num Characters
Text

Language: English, Romance, and other

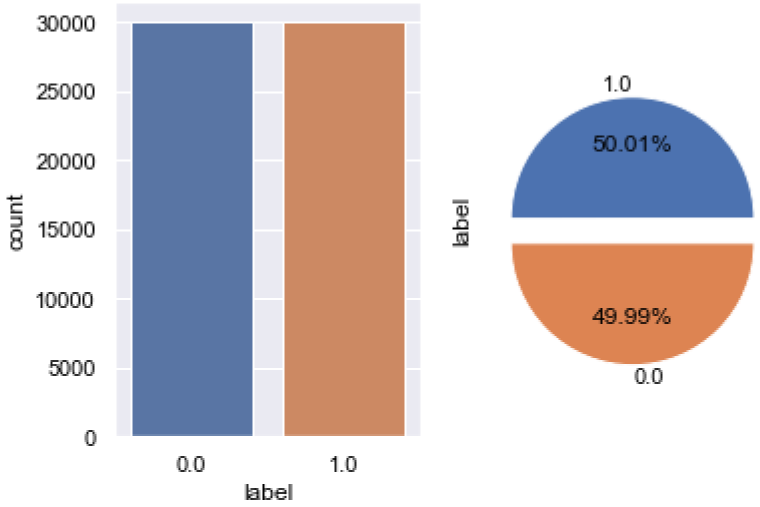
Structure: Words that represent the structure of a paper
{abstract, introduction, conclusion, reference, table of content}

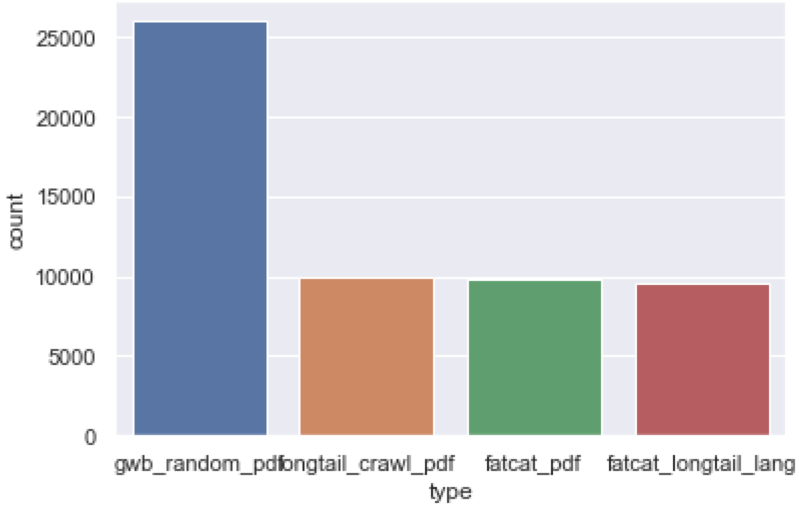
Content: Words that represent the content of a paper
{research, analyze, result, table, investigation, explain, theory, study, paper, data, perform}

Association: Words that represent association
{journal, association, organization, doi, university, school, board}

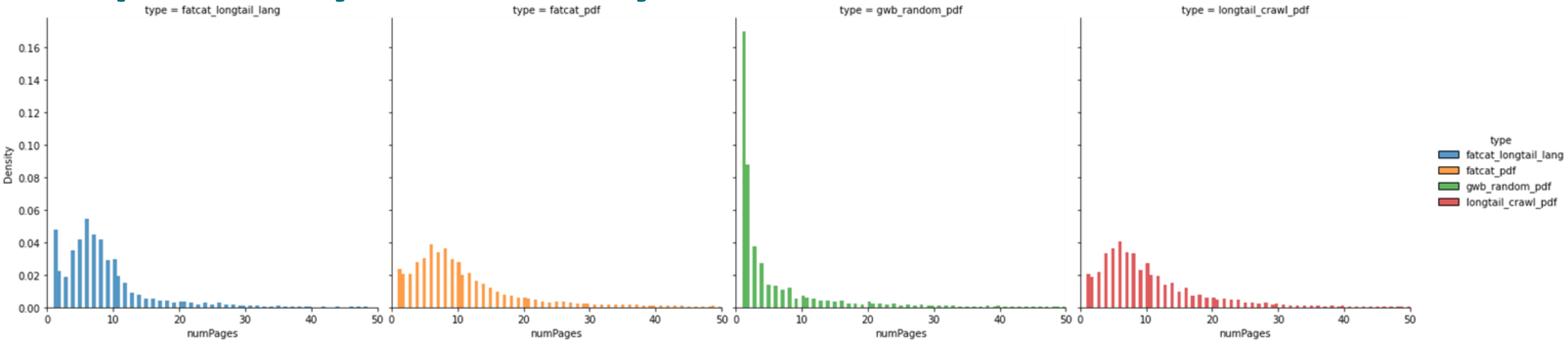
Balanced Data

- Balanced dataset

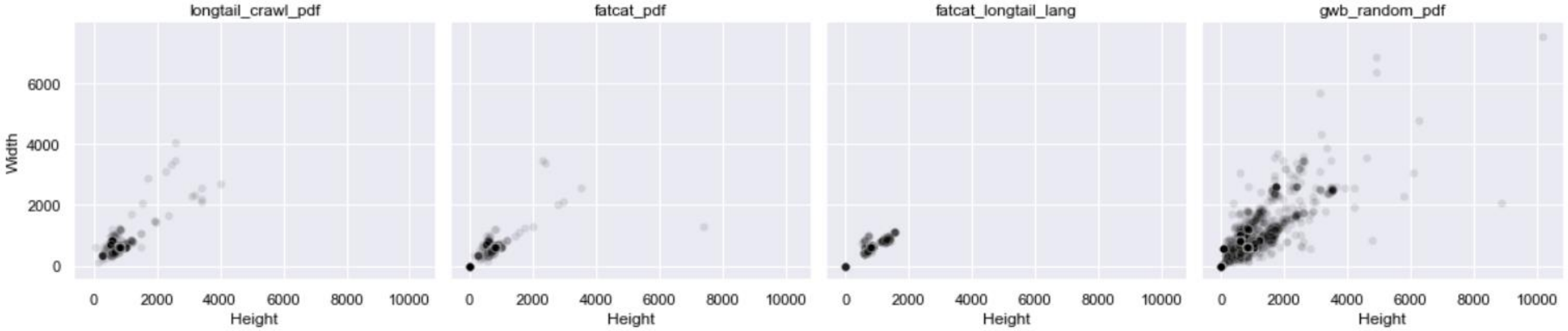




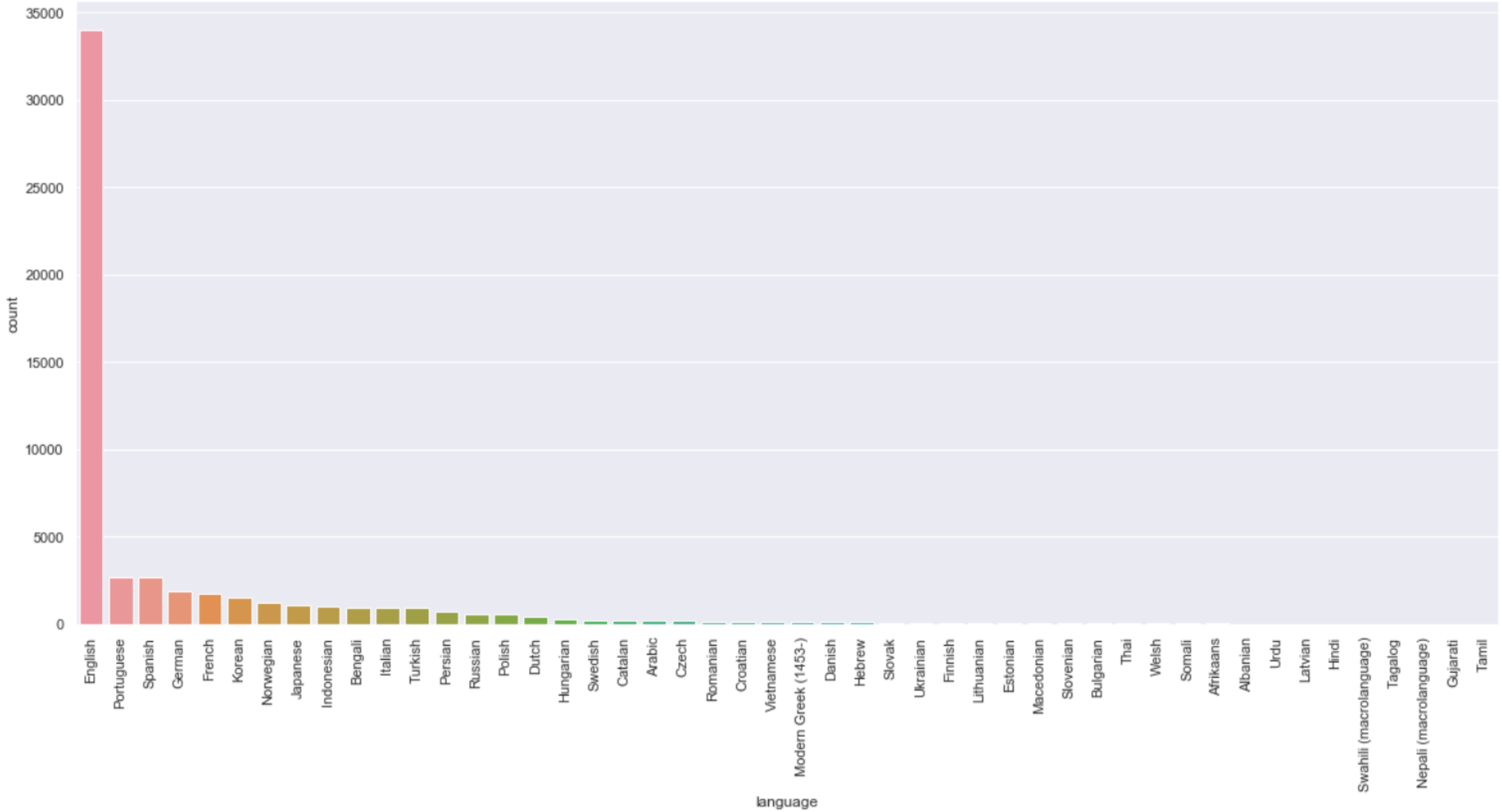
Exploratory Data Analysis



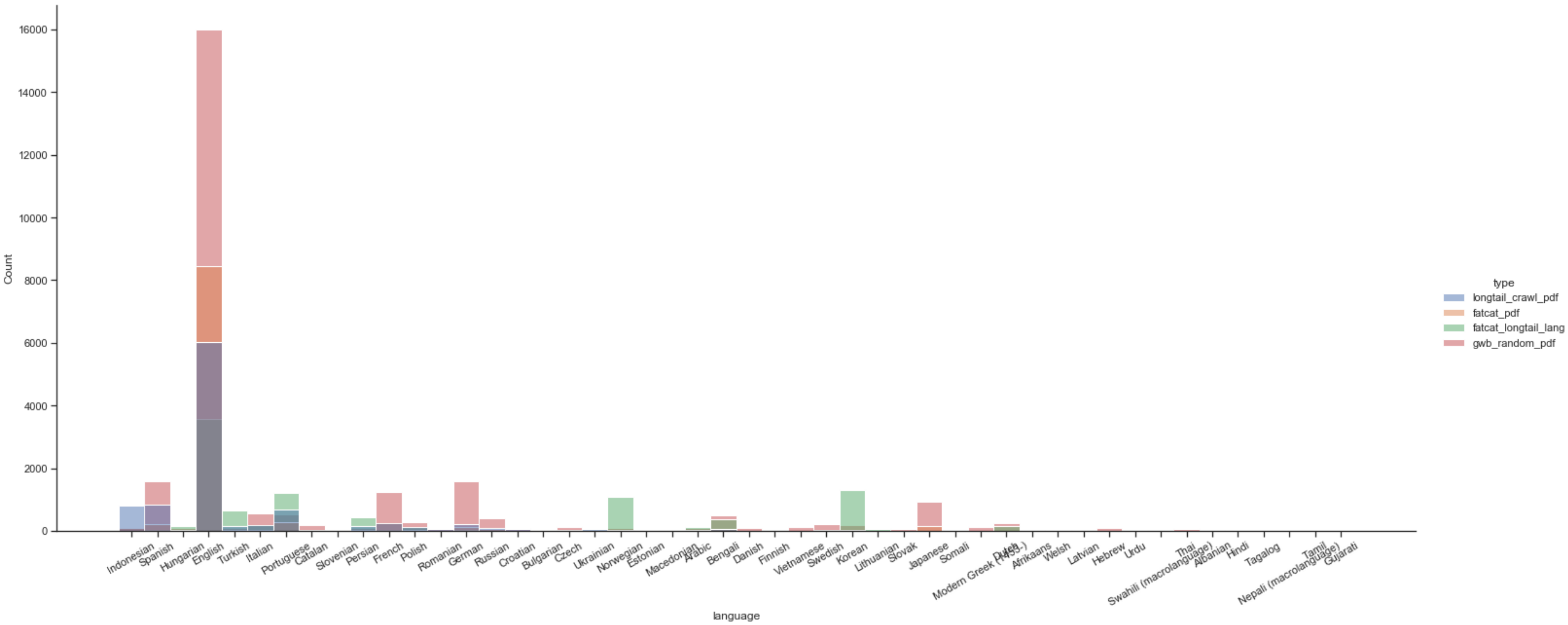
Evaluation of document dimension by document type



Exploratory Data Analysis



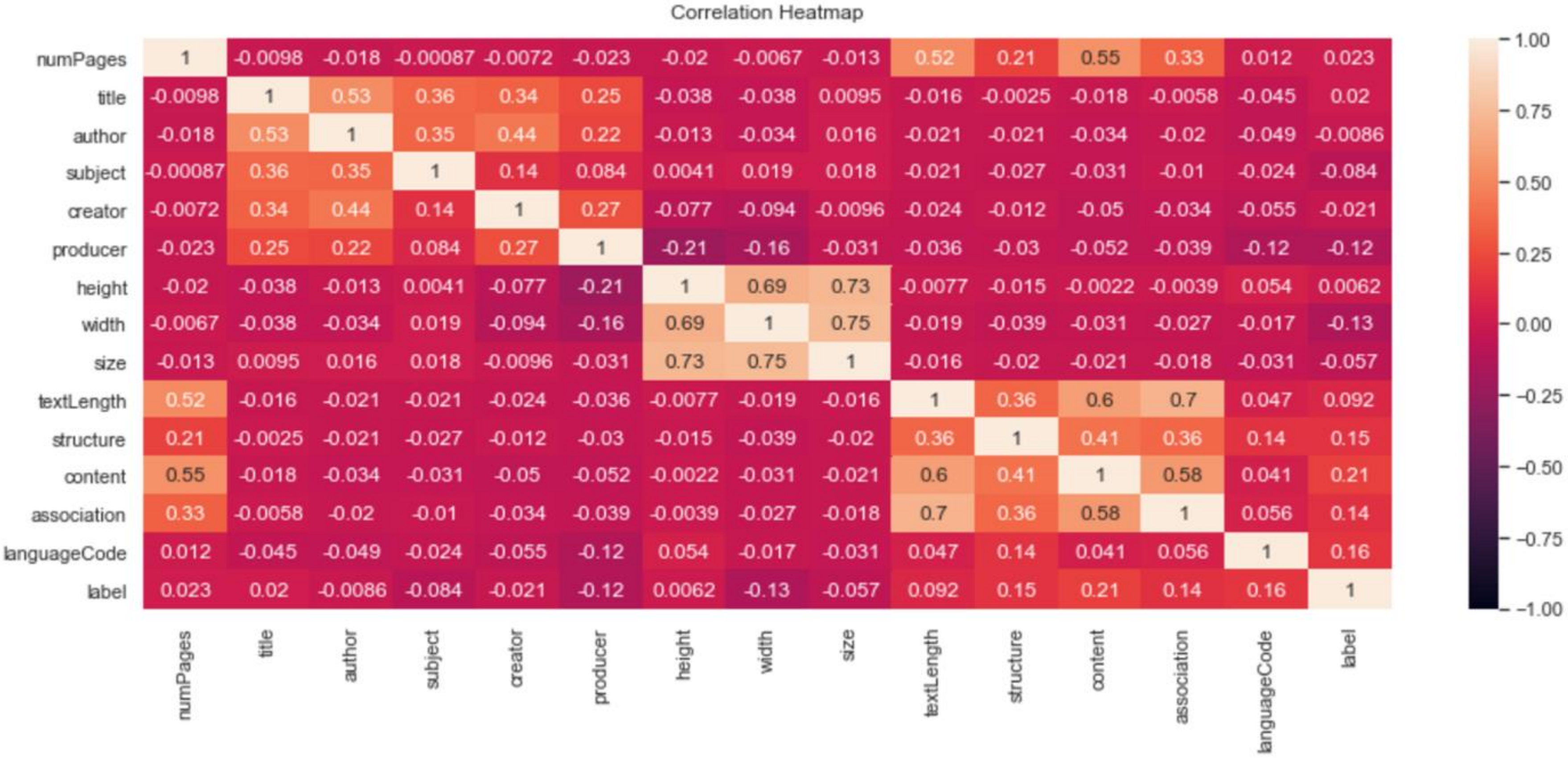
Exploratory Data Analysis



Data Feature Engineering

- The use of multiprocessing allows for further feature extraction
    - Ability to look for keywords in text
        - English
        - Non-english
            - Translate keywords to the language of the text
    - Process adds 14 minutes to the additionally extraction of meta and text data

Heatmap



Correlation Heatmap

	numPages	title	author	subject	creator	producer	height	width	size	textLength	structure	content	association	languageCode	label
numPages	1	-0.0098	-0.018	-0.00087	-0.0072	-0.023	-0.02	-0.0067	-0.013	0.52	0.21	0.55	0.33	0.012	0.023
title	-0.0098	1	0.53	0.36	0.34	0.25	-0.038	-0.038	0.0095	-0.016	-0.0025	-0.018	-0.0058	-0.045	0.02
author	-0.018	0.53	1	0.35	0.44	0.22	-0.013	-0.034	0.016	-0.021	-0.021	-0.034	-0.02	-0.049	-0.0086
subject	-0.00087	0.36	0.35	1	0.14	0.084	0.0041	0.019	0.018	-0.021	-0.027	-0.031	-0.01	-0.024	-0.084
creator	-0.0072	0.34	0.44	0.14	1	0.27	-0.077	-0.094	-0.0096	-0.024	-0.012	-0.05	-0.034	-0.055	-0.021
producer	-0.023	0.25	0.22	0.084	0.27	1	-0.21	-0.16	-0.031	-0.036	-0.03	-0.052	-0.039	-0.12	-0.12
height	-0.02	-0.038	-0.013	0.0041	-0.077	-0.21	1	0.69	0.73	-0.0077	-0.015	-0.0022	-0.0039	0.054	0.0062
width	-0.0067	-0.038	-0.034	0.019	-0.094	-0.16	0.69	1	0.75	-0.019	-0.039	-0.031	-0.027	-0.017	-0.13
size	-0.013	0.0095	0.016	0.018	-0.0096	-0.031	0.73	0.75	1	-0.016	-0.02	-0.021	-0.018	-0.031	-0.057
textLength	0.52	-0.016	-0.021	-0.021	-0.024	-0.036	-0.0077	-0.019	-0.016	1	0.36	0.6	0.7	0.047	0.092
structure	0.21	-0.0025	-0.021	-0.027	-0.012	-0.03	-0.015	-0.039	-0.02	0.36	1	0.41	0.36	0.14	0.15
content	0.55	-0.018	-0.034	-0.031	-0.05	-0.052	-0.0022	-0.031	-0.021	0.6	0.41	1	0.58	0.041	0.21
association	0.33	-0.0058	-0.02	-0.01	-0.034	-0.039	-0.0039	-0.027	-0.018	0.7	0.36	0.58	1	0.056	0.14
languageCode	0.012	-0.045	-0.049	-0.024	-0.055	-0.12	0.054	-0.017	-0.031	0.047	0.14	0.041	0.056	1	0.16
label	0.023	0.02	-0.0086	-0.084	-0.021	-0.12	0.0062	-0.13	-0.057	0.092	0.15	0.21	0.14	0.16	1

Models

- Balanced Data
- Data Feature Engineering
- Models
  - Text Based Models
    - XGBoost
    - Keras
    - SVM
  - Image Based Model
    - Keras (VGG16)
  - Bayesian statistic
    - Logistic models
  - Topic modeling (LDA)

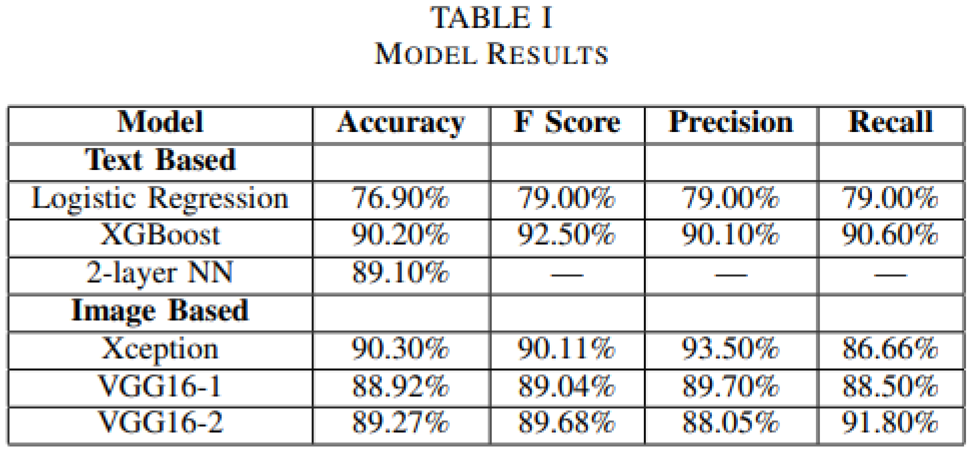
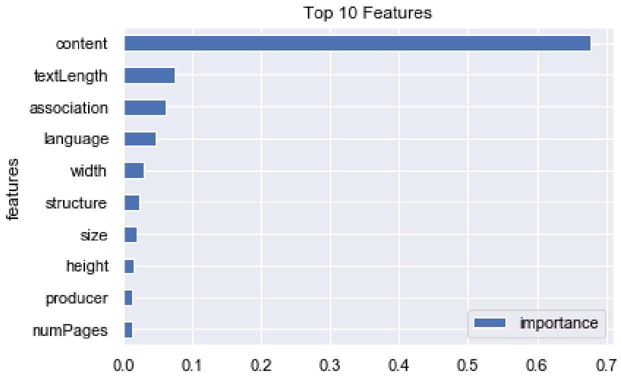


TABLE I
MODEL RESULTS

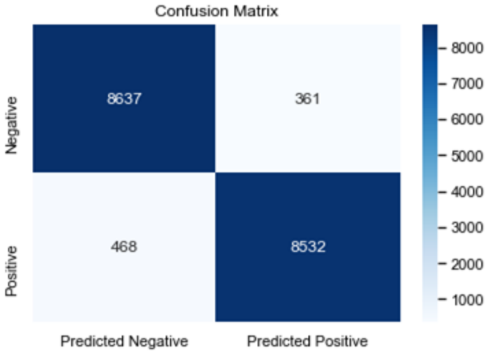
Model	**Accuracy**	**F Score**	**Precision**	**Recall**
Text Based				
Logistic Regression	76.90%	79.00%	79.00%	79.00%
XGBoost	90.20%	92.50%	90.10%	90.60%
2-layer NN	89.10%	—	—	—
Image Based				
Xception	90.30%	90.11%	93.50%	86.66%
VGG16-1	88.92%	89.04%	89.70%	88.50%
VGG16-2	89.27%	89.68%	88.05%	91.80%

XGBoost

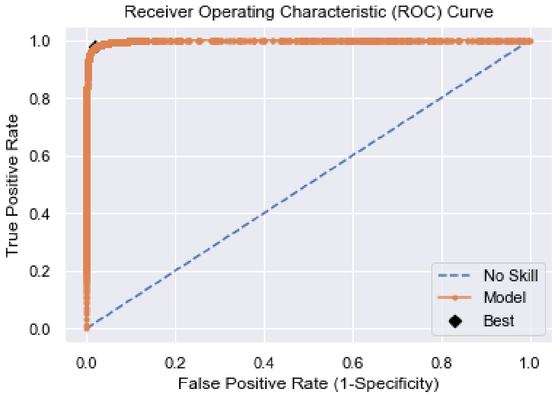
- Grid Search
  - N Estimators, Learning Rate, Depth
  - 3 k-fold
- F-score : 97.90%
- Accuracy: 95.39%



```
Confusion Matrix for Threshold=0.485
```



```
Best Threshold=0.485, F-Score=0.979
```



BAYESIAN STATISTICS

- Algorithm: Bayesian logistical regression, using PYMC3

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Mathematical connection : the likelihood is the product of n Bernoulli trials,

$$\coprod_{i=1}^{n} p_i^{y}(1-p_i)\ (1-p_i)^{1-y_i}\text{, where } p_i = \frac{1}{1+e^{-z_i}}$$

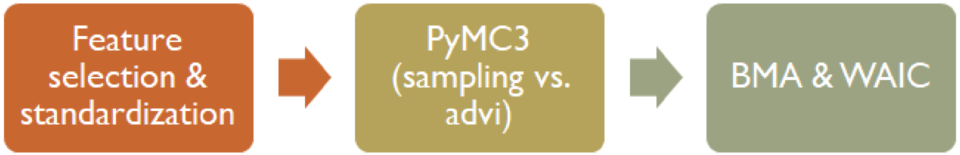
$$y_i = \beta_0 + \beta_1(\text{numPages})_\text{i} + \beta_2(\text{height})_{\text{i}\,+}\ \beta_3(\text{width})_\text{i} + \beta_4(\text{dim})_\text{i} + \beta_5(\text{structure})_\text{i} + \beta_6(\text{content})_\text{i}$$
$$+ \beta_7(\text{association})_\text{i} + \beta_8(\text{language})_\text{i} + \beta_9(\text{numChar})_\text{i}$$

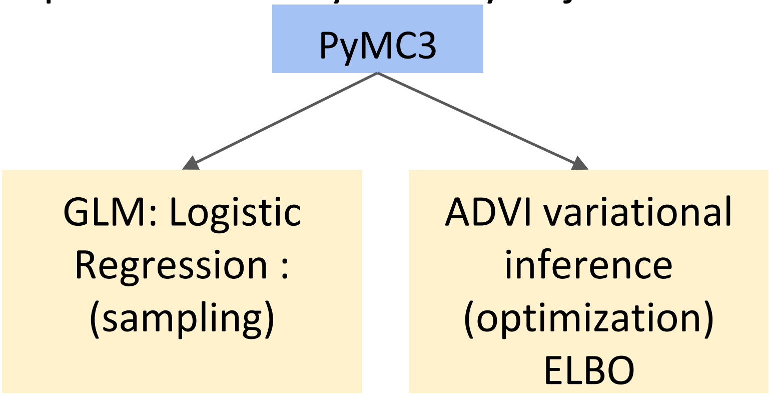
Where y_i = 1 if researchPublication and y_i = 0 otherwise

- Priors : default $p(\theta)$ = N(0, 10^{12}I)

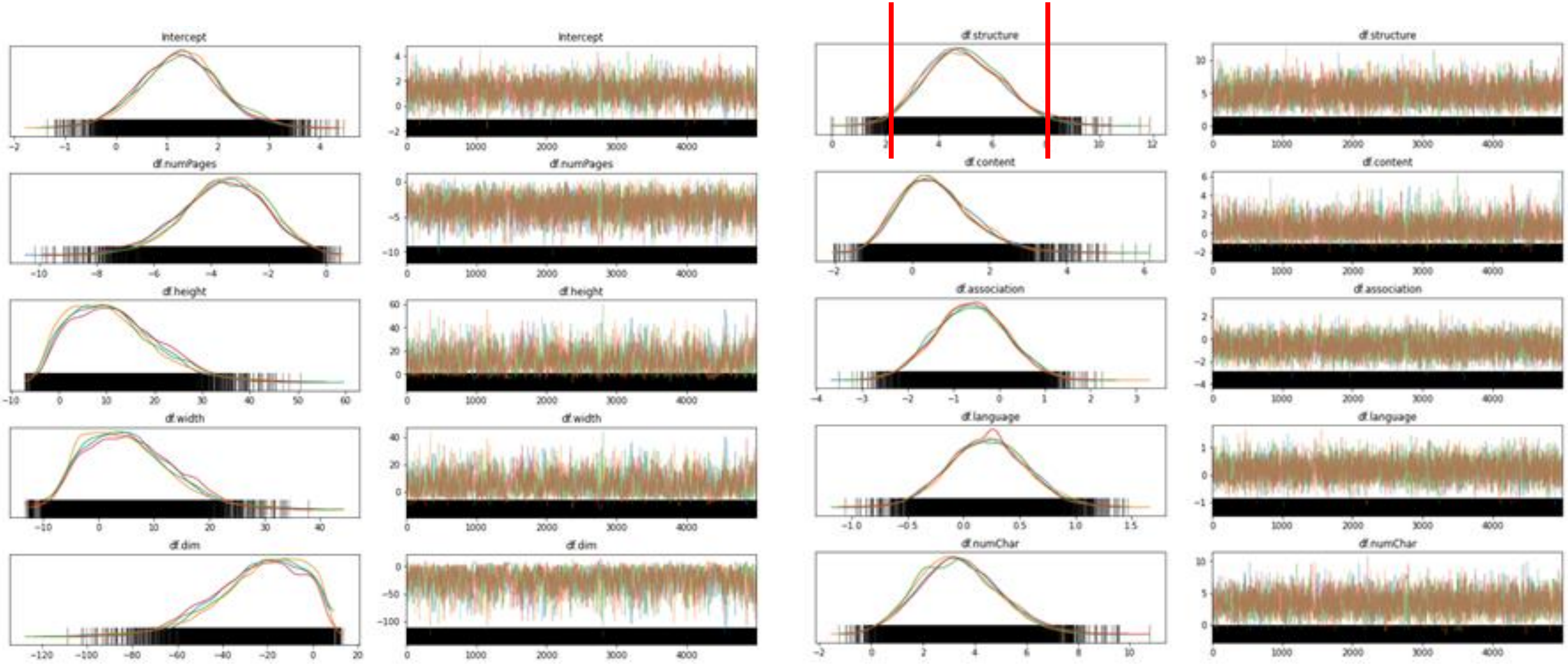
Methods

- Total set of features considered: number of pages, height, width, dimensions of page, structure, content, association, language, number of characters
- How likely is it a **research publication** based on the selective features?
- Model comparison approach – compared different sets of features and accompanying accuracy. Given parameters for the capstone project, including speed, prioritizing a balance of the smallest number of features with acceptable accuracy is a key objective


Feature selection & standardization
PyMC3 (sampling vs. advi)
BMA & WAIC

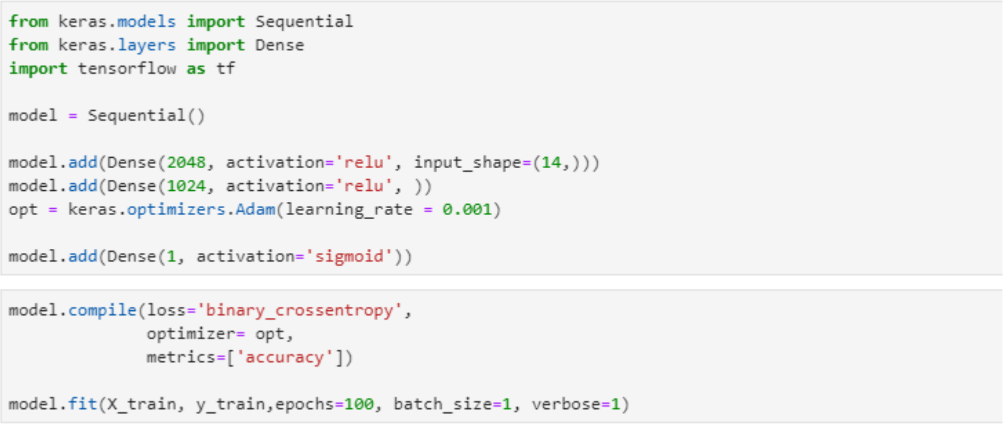

PyMC3
GLM: Logistic Regression : (sampling)
ADVI variational inference (optimization) ELBO

RESULTS - Full Model (Sampling)



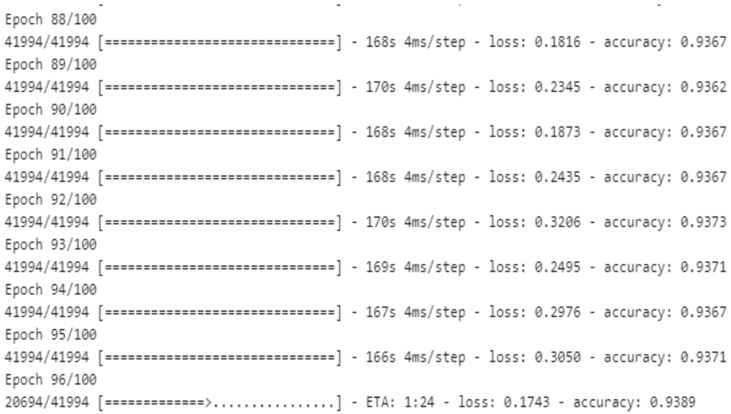
Keras-Tensorflow

- Model Structure
  - Input dimension of 14
  - Two hidden layers
  - Adam optimizer
  - Epochs 100
- Accuracy: 93.89%



```
from keras.models import Sequential
from keras.layers import Dense
import tensorflow as tf

model = Sequential()

model.add(Dense(2048, activation='relu', input_shape=(14,)))
model.add(Dense(1024, activation='relu', ))
opt = keras.optimizers.Adam(learning_rate = 0.001)

model.add(Dense(1, activation='sigmoid'))
```

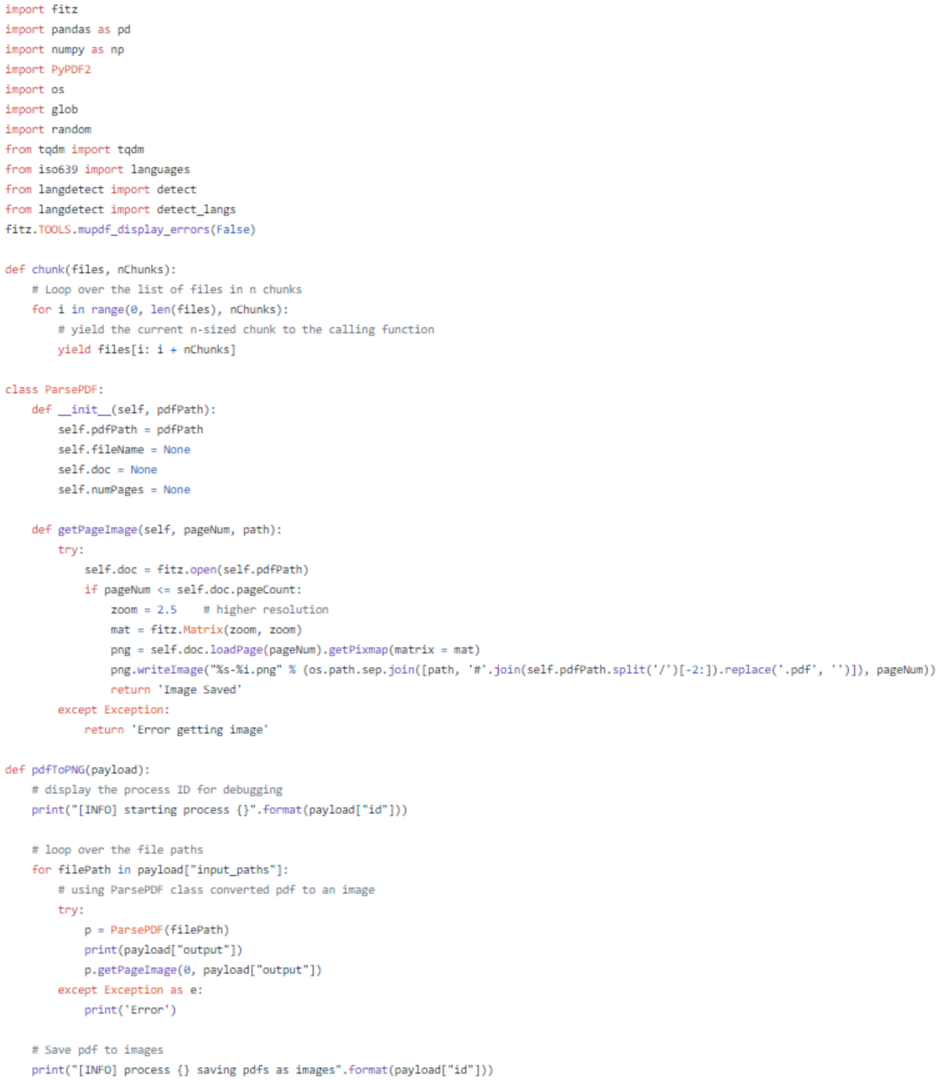
```
model.compile(loss='binary_crossentropy',
              optimizer= opt,
              metrics=['accuracy'])

model.fit(X_train, y_train,epochs=100, batch_size=1, verbose=1)
```



```
Epoch 88/100
41994/41994 [==============================] - 168s 4ms/step - loss: 0.1816 - accuracy: 0.9367
Epoch 89/100
41994/41994 [==============================] - 170s 4ms/step - loss: 0.2345 - accuracy: 0.9362
Epoch 90/100
41994/41994 [==============================] - 168s 4ms/step - loss: 0.1873 - accuracy: 0.9367
Epoch 91/100
41994/41994 [==============================] - 168s 4ms/step - loss: 0.2435 - accuracy: 0.9367
Epoch 92/100
41994/41994 [==============================] - 170s 4ms/step - loss: 0.3206 - accuracy: 0.9373
Epoch 93/100
41994/41994 [==============================] - 169s 4ms/step - loss: 0.2495 - accuracy: 0.9371
Epoch 94/100
41994/41994 [==============================] - 167s 4ms/step - loss: 0.2976 - accuracy: 0.9367
Epoch 95/100
41994/41994 [==============================] - 166s 4ms/step - loss: 0.3050 - accuracy: 0.9371
Epoch 96/100
20694/41994 [=============>................] - ETA: 1:24 - loss: 0.1743 - accuracy: 0.9389
```

Image Based Model

- Adopt first page of PDF file and convert to the image
- Using pretrain modeling to perform image classification

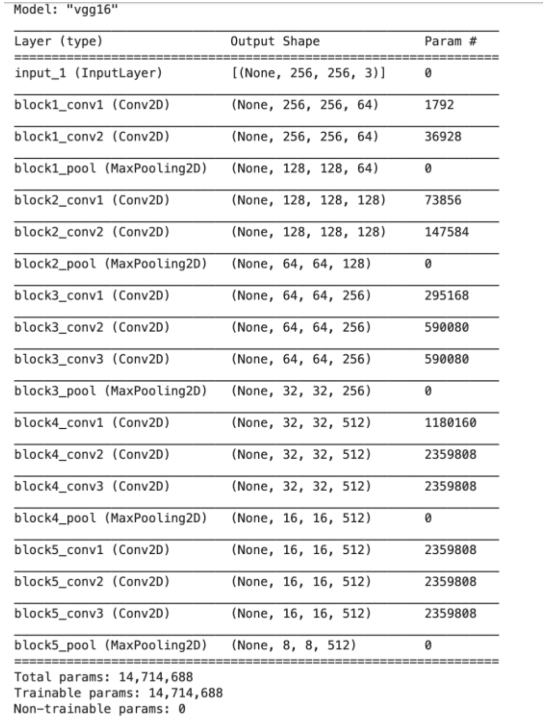


```
import fitz
import pandas as pd
import numpy as np
import PyPDF2
import os
import glob
import random
from tqdm import tqdm
from iso639 import languages
from langdetect import detect
from langdetect import detect_langs
fitz.TOOLS.mupdf_display_errors(False)

def chunk(files, nChunks):
    # Loop over the list of files in n chunks
    for i in range(0, len(files), nChunks):
        # yield the current n-sized chunk to the calling function
        yield files[i: i + nChunks]

class ParsePDF:
    def __init__(self, pdfPath):
        self.pdfPath = pdfPath
        self.fileName = None
        self.doc = None
        self.numPages = None

    def getPageImage(self, pageNum, path):
        try:
            self.doc = fitz.open(self.pdfPath)
            if pageNum <= self.doc.pageCount:
                zoom = 2.5    # higher resolution
                mat = fitz.Matrix(zoom, zoom)
                png = self.doc.loadPage(pageNum).getPixmap(matrix = mat)
                png.writeImage("%s-%i.png" % (os.path.sep.join([path, '#'.join(self.pdfPath.split('/')[-2:]).replace('.pdf', '')]), pageNum))
                return 'Image Saved'
        except Exception:
            return 'Error getting image'

def pdfToPNG(payload):
    # display the process ID for debugging
    print("[INFO] starting process {}".format(payload["id"]))

    # loop over the file paths
    for filePath in payload["input_paths"]:
        # using ParsePDF class converted pdf to an image
        try:
            p = ParsePDF(filePath)
            print(payload["output"])
            p.getPageImage(0, payload["output"])
        except Exception as e:
            print('Error')

    # Save pdf to images
    print("[INFO] process {} saving pdfs as images".format(payload["id"]))
```

Image Based Model

- Leveraged an existing Keras application, VGG16, for large scale image classification

Model Type	Accuracy
Keras (VGG16)	90.01%



```
Model: "vgg16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 256, 256, 3)]     0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 256, 256, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 256, 256, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 128, 128, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 128, 128, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 128, 128, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 64, 64, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 64, 64, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 32, 32, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 32, 32, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 16, 16, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 8, 8, 512)         0
=================================================================
Total params: 14,714,688
Trainable params: 14,714,688
Non-trainable params: 0
```

Conclusions

- Top feature - structure
- Additional key features - number of pages, width
- Format matters, in addition to content
- Simple model may be reasonable, given comparable accuracy to others, to prioritize speed
- Compare the pros and cons of text-based vs image-based models