Proceedings 2021 FAMU REU
supported by Cybertraining-DSC

Gregor von Laszewski,
Yohn Jairo Parra Bautista,
Carlos Theran, Geoffrey C. Fox
Richard Alo, Byron Greene
Editors

Contact: laszewski@gmail.com

https://cybertraining-dsc.github.io/pubs/reu2021.pdf

August 16, 2021 - 02:20 PM

Created by Cloudmesh & Cyberaide Bookmanager, https://github.com/cyberaide/bookmanager

PROCEEDINGS 2021 FAMU REU

Gregor von Laszewski, Yohn Jairo Parra Bautista, Carlos Theran, Geoffrey C. Fox Richard Alo, Byron Greene

PROCEEDINGS 2021 FAMU REU

1 PREFACE

Mon Aug 16 14:20:00 EDT 2021

1.1 DISCLAIMER

This book has been generated with Cyberaide Bookmanager.

Bookmanager is a tool to create a publication from a number of sources on the internet. It is especially useful to create customized books, lecture notes, or handouts. Content is best integrated in markdown format as it is very fast to produce the output.

Bookmanager has been developed based on our experience over the last 3 years with a more sophisticated approach. Bookmanager takes the lessons from this approach and distributes a tool that can easily be used by others.

The following shields provide some information about it. Feel free to click on them.

1.1.1 ACKNOWLEDGMENT

If you use bookmanager to produce a document you must include the following acknowledgement.

> *"This document was produced with Cyberaide Bookmanager developed by Gregor von Laszewski available at https://pypi.python.org/pypi/cyberaide-bookmanager. It is in the responsibility of the user to make sure an author acknowledgement section is included in your document. Copyright verification of content included in a book is responsibility of the book editor."*

The bibtex entry is

```
@Misc{www-cyberaide-bookmanager,
  author =    {Gregor von Laszewski},
  title =     {{Cyberaide Book Manager}},
  howpublished = {pypi},
  month =     apr,
  year =      2019,
  url={https://pypi.org/project/cyberaide-bookmanager/}
}
```

1.1.2 Extensions

We are happy to discuss with you bugs, issues and ideas for enhancements. Please use the convenient github issues at

- https://github.com/cyberaide/bookmanager/issues

Please do not file with us issues that relate to an editors book. They will provide you with their own mechanism on how to correct their content.

2 REFERENCES

2021 REU Course

This course introduces the REU students to various topics in Intelligent Systems Engineering. The course was taught in Summer 2021.

7 minute read

This course introduces the REU students to various topics in Intelligent Systems Engineering. The course was taught in Summer 2021.

Rstudio with Git and GitHub Slides

Rstudio with Git and GitHub Slides

Programming with Python

Python is a great languge for doing data science and AI, a comprehensive list of features is available in book form. Please note that when installing Python, you always want to use a venv as this is best practice.

Introduction to Python (ePub) (PDF)

Installation of Python

Installation of Python — June 7th, 2021 (AM)

Update to the Video:

Best practices in Python recommend to use a Python venv. This is pretty easy to do and creates a separate Python environment for you so you do not interfere with your system Python installation. Some IDEs may do this automatically, but it is still best practice to install one and bind the IDE against it. To do this:

1. Download Python version 3.9.5 just as shown in the first lecture.
2. After the download you do an additional step as follows:
    - on Mac:

      ```
      python3.9 –m venv ~/ENV3
      source ~/ENV/bin/activate
      ```

      you need to do the source every time you start a new window or on mac ass it to .zprofile

- on Windows you first install gitbash and do all yuour terminal work from gitbash as this is more Linux-like. In gitbash, run

  ```
  python –m venv ~/ENV3
  ~/ENV/Script/activate
  ```

  In case you like to add it to gitbash, you can add the source line to .bashrc and/or .bash_profile

3. In case you use VSCode, you can also do it individually in a directory where you have your code.
    - On Mac: `cd TO YOUR DIR; python3.9 –m venv .`
    - On Windows `cd TO YOUR DIR; python –m venv .`

    Then start VSCode in the directory and it will ask you to use this venv. However, the global ENV3 venv may be better and you cen set your interpreter to it.
4. On Pycharm we recommend you use the ENV3 and set the clobal interpreter

Jupyter Notebooks

Jupyter Notebooks — June 7th, 2021 (PM): This lecture provides an introduction to Jupyter Notebooks using Visual Studio as IDE.

Github

Video: Github

Video-Github 2 — June 8th, 2021 (PM): In this lecture the student can learn how to create a project on RStudio and link it with a repository on GitHub to commit, pull and push the code from RStudio.

Introduction to Python

Slides: This introduction to Python cover the different data type, how to convert type of variable, understand and create flow control usign conditional statements.

Video-Introduction to Python (1) — June 9th, 2021 (AM): This introduction to Python cover the different data type, how to convert type of variable, understand and create flow control usign conditional statements.

Video-Introduction to Python (2) — June 9th, 2021 (PM): This introduction to Python cover the different data type, how to convert type of variable, understand and create flow control usign conditional statements.

Video-Introduction to Python (3) — June 10th, 2021 (AM): This lecture introduces the use of Google Colab to code your python program using the resources provided by Google. Also, DataFrame is introduced and use to manipulate and analyze data.

Slides — June 10th, 2021 (PM): String, Numbers, Booleans Flow of control Using If statements

Slides: String, Numbers, Booleans Flow of control Using If statements (2)

Python Exercises - Lab 2

The first exercise will require a simple for loop, while the second is more complicated, requiring nested for loops and a break statement.

General Instructions: Create two different files with extension .ipnyb, one for each problem. The first file will be named factorial.ipnyb which is for the factorial problem, and the second prime_number.ipnyb for the prime number problem.

1. Write a program that can find the factorial of any given number. For example, find the factorial of the number 5 (often written as 5!) which is 1*2*3*4*5 and equals 120. Your program should take as input an integer from the user.

   Note: The factorial is not defined for negative numbers and the factorial of Zero is 1; that is 0! = 1.

   You should

   1. If the number is less than Zero return with an error message.
   2. Check to see if the number is Zero—if it is then the answer is 1—print this out.
   3. Otherwise use a loop to generate the result and print it out.
2. A Prime Number is a positive whole number, greater than 1, that has no other divisors except the number 1 and the number itself. That is, it can only be divided by itself and the number 1, for example the numbers 2, 3, 5 and 7 are prime numbers as they cannot be divided by any other whole number. However, the numbers 4 and 6 are not because they can both be divided by the number 2 in addition the number 6 can also be divided by the number 3.

   You should write a program to calculate prime number starting from 1 up to the value input by the user.

You should

1. If the user inputs a number below 2, print an error message.
2. For any number greater than 2 loop for each integer from 2 to that number and determine if it can be divided by another number (you will probably need two for loops for this; one nested inside the other).
3. For each number that cannot be divided by any other number (that is its a prime number) print it out.

Motivation for the REU

Video — June 11th, 2021 (AM): Motivation for the REU: Data is Driven Everything

Slides: Motivation for the REU: Data is Driven Everything

Slides: Descriptive Statistic

Slides: Probability

Video — June 28th, 2021 (AM): Working on GitHUb Template and Mendeley references management

Data Science Tools

Slides: Data Science Tools

Video — June 14th, 2021 (AM): Numpy

Video — June 14th, 2021 (PM): Pandas data frame

Video — June 15th, 2021 (AM): Web data mining

Video — June 15th, 2021 (PM): Pandas IO

Video — June 16th, 2021 (AM): Pandas

Video-Matrix computation — June 16th, 2021 (PM): Linear algebra is a main compornent in the field of Data Science. As a consequence, this lecture introduces the main matrix operations such as, addition, substraction, multiplication, and picewise multiplication.

Video: Pycharm Installation and Virtual Environment setup — June 18th, 2021 (AM)

Video: This lecture the student can learn the different applications of Matrix Operation using images on Python. — June 21st, 2021 (AM)

Video: Data wrangling and Descriptive Statistic Using Python — June 21st, 2021 (AM)

Video: Data wrangling and Descriptive Statistic Using Python — June 22nd, 2021 (PM)

Video: FURY Visualization and Microsoft Lecture — June 25th, 2021 (PM)

Video: Instroduction to Probability — June 25th, 2021 (PM)

Video: Digital Twins and Virtual Tissue ussing CompuCell3D Simulating Cancer Somatic Evolution in nanoHUB — July 2nd, 2021 (AM)

AI First Engineering

Video: AI First Engineering: Learning material — June 25th, 2021 (AM)

- Video: Adding content to your su21-reu repositories — June 17th, 2021 (PM)

- Slides: AI First Engineering

Datasets for Projects

- Video: Datasets for Projects: Data world and Kaggle — June 29th, 2021 (AM)

- Video: Datasets for Projects: Data world and Kaggle part 2 — June 29th, 2021 (PM)

Machine Learning Models

- Video: K-Means: Unsupervised model — June 30th, 2021 (AM)

- Video: Support Vector Machine: Supervised model — July 2nd, 2021 (PM)

- Slides: Support Vector Machine Supervised model.

- Video: Neural Networks: Deep Learning Supervised model — July 6th, 2021 (AM)

- Video: Neural Networks: Deep learning Model — July 6th, 2021 (AM)

- Video: Data Visualization: Visualizaton for Data Science — July 7th, 2021 (AM)

- Video: Convulotional Neural Networks: Deep learning Model — July 8th, 2021 (AM)

Students Report Help

Video: Student Report Help with Introduction and Datasets — July 7th, 2021 (AM)

Video: Student Report Help with Introduction and Datasets — July 13th, 2021 (AM)

COVID-19

Video: Chemo-Preventive Effect of Vegetables and Fruits Consumption on the COVID-19 Pandemic — July 1st, 2021 (AM)

- Yedjou CG, Alo RA, Liu J, et al. Chemo-Preventive Effect of Vegetables and Fruits Consumption on the COVID-19 Pandemic. J Nutr Food Sci. 2021;4(2):029
- Geoffrey C. Fox, Gregor von Laszewski, Fugang Wang, Saumyadipta Pyne, AICov: An Integrative Deep Learning Framework for COVID-19 Forecasting with Population Covariates, J. data sci. 19(2021), no. 2, 293-313, DOI 10.6339/21-JDS1007

Last modified July 14, 2021 : Update index.md (24abd640)

2021 REU Reports

Research experience for under graduate reports for the summer of 2021.

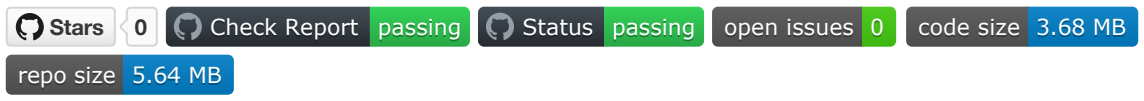
1 minute read

This page contains the list of the reports and projects done in the REU 2021.

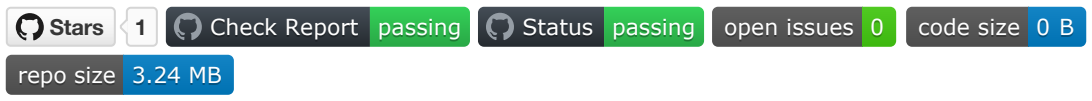
List for 2021

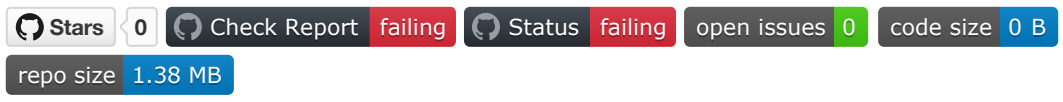
Reports and Projects

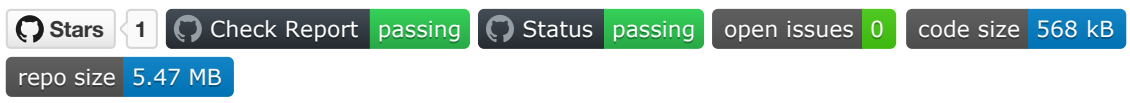
- su21-reu-361: Project: Time Series Analysis of Blockchain-Based Cryptocurrency Price Changes, Jacques Fleischer
  - Stars 1 | Check Report passing | Status passing | open issues 0 | code size 1.79 MB | repo size 8.03 MB
- su21-reu-362: Project: Breast Cancer and Genetics, Kehinde Ezekiel
  - Stars 1 | Check Report passing | Status failing | open issues 0 | code size 447 kB | repo size 1.23 MB
- su21-reu-363: Project: AI in Orthodontics, Whitney McNair
  - Stars 1 | Check Report passing | Status passing | open issues 0 | code size 0 B | repo size 2.74 MB
- su21-reu-364: Project: Object Recognition, David Umanzor
  - Stars 1 | Check Report passing | Status failing | open issues 0 | code size 144 kB | repo size 3.86 MB
- su21-reu-365: Project: Cyber Attacks Detection Using AI Algorithms, Victor Adankai
  - Stars 0 | Check Report passing | Status failing | open issues 0 | code size 6.69 kB | repo size 100 kB
- su21-reu-366: Project: Handwriting Recognition Using AI, Mikahla Reeves
  - Stars 0 | Check Report passing | Status passing | open issues 0 | code size 0 B | repo size 6.95 MB
- su21-reu-369: Project: Increasing Cervical Cancer Risk Analysis, Theresa Jeanbaptiste
  - Stars 0 | Check Report passing | Status failing | open issues 0 | code size 70.8 kB | repo size 4.36 MB
- su21-reu-370: Project: Aquatic Animals Classification Using AI, Timia Williams
  - Stars 0 | Check Report passing | Status passing | open issues 0 | code size 0 B | repo size 9.95 MB
- su21-reu-371: Project: Detecting Multiple Sclerosis Symptoms using AI, Raeven Hatcher
  - Stars 0 | Check Report failing | Status failing | open issues 0 | code size 0 B | repo size 187 kB
- su21-reu-372: Project: Analysing Hashimoto disease causes using AI, Sheimy Paz

  - 

- su21-reu-375: Project: Analysis of Covid-19 Vaccination Rates in Different Races, Ololade Latinwo
  - 

- su21-reu-376: Project: AI and Dentronics, Jamyla Young
  - 

- su21-reu-377: Project: Analyzing the Advantages and Disadvantages of Artificial Intelligence for Breast Cancer Detection in Women, RonDaisja Dunn
  - 

- su21-reu-378: Project: Analysis of Autism in three different cities using AI, Myra Saunders
  - 

Last modified August 14, 2021 : fix name and title (142c998d)

Time Series Analysis of Blockchain-Based Cryptocurrency Price Changes

This project applies neural networks and Artificial Intelligence (AI) to historical records of high-risk cryptocurrency coins to train a prediction model that guesses their price. The code in this project contains Jupyter notebooks, one of which outputs a timeseries graph of any cryptocurrency price once a csv file of the historical data is inputted into the program. Another Jupyter notebook trains an LSTM, or a long short-term memory model, to predict a cryptocurrency's closing price. The LSTM is fed the close price, which is the price that the currency has at the end of the day, so it can learn from those values. The notebook creates two sets: a training set and a test set to assess the accuracy of the results. The data is then normalized using manual min-max scaling so that the model does not experience any bias; this also enhances the performance of the model. Then, the model is trained using three layers— an LSTM, dropout, and dense layer—minimizing the loss through 50 epochs of training; from this training, a recurrent neural network (RNN) is produced and fitted to the training set. Additionally, a graph of the loss over each epoch is produced, with the loss minimizing over time. Finally, the notebook plots a line graph of the actual currency price in red and the predicted price in blue. The process is then repeated for several more cryptocurrencies to compare prediction models. The parameters for the LSTM, such as number of epochs and batch size, are tweaked to try and minimize the root mean square error.

Tags: project reu blockchain finance

10 minute read

Check Report passing Status passing Status: final, Type: Project

Jacques Fleischer, su21-reu-361, Edit

- Code:
  - Install documentation README.md[1]
  - yfinance-lstm.ipynb[2]

Abstract

This project applies neural networks and Artificial Intelligence (AI) to historical records of high-risk cryptocurrency coins to train a prediction model that guesses their price. The code in this project contains Jupyter notebooks, one of which outputs a timeseries graph of any cryptocurrency price once a csv file of the historical data is inputted into the program. Another Jupyter notebook trains an LSTM, or a long short-term memory model, to predict a cryptocurrency's closing price. The LSTM is fed the close price, which is the price that the currency has at the end of the day, so it can learn from those values. The notebook creates two sets: a training set and a test set to assess the accuracy of the results.

The data is then normalized using manual min-max scaling so that the model does not experience any bias; this also enhances the performance of the model. Then, the model is trained using three layers— an LSTM, dropout, and dense layer—minimizing the loss through 50 epochs of training; from this training, a recurrent neural network (RNN) is produced and fitted to the training set.

Additionally, a graph of the loss over each epoch is produced, with the loss minimizing over time. Finally, the notebook plots a line graph of the actual currency price in red and the predicted price in blue. The process is then repeated for several more cryptocurrencies to compare prediction models. The parameters for the LSTM, such as number of epochs and batch size, are tweaked to try and minimize the root mean square error.

Contents

Keywords: cryptocurrency, investing, business, blockchain.

1. Introduction

Blockchain is *an open, distributed ledger* which records transactions of cryptocurrency. Systems in blockchain are decentralized, which means that these transactions are shared and distributed among all participants on the blockchain for maximum accountability. Furthermore, this new blockchain technology is becoming an increasingly popular alternative to mainstream transactions through traditional banks[3]. These transactions utilize blockchain-based *cryptocurrency*, which is a popular investment of today's age, particularly in Bitcoin. However, the U.S. Securities and Exchange Commission warns that high-risk accompanies these investments[4].

Artificial Intelligence (AI) can be used to predict the prices' behavior to avoid cryptocurrency coins' severe volatility that can scare away possible investors[5]. AI and blockchain technology make an ideal partnership in data science; the insights generated from the former and the secure environment ensured by the latter create a goldmine for valuable information. For example, an up-and-coming innovation is the automatic trading of *digital investment assets* by AI, which will hugely outperform trading conducted by humans[6]. This innovation would not be possible without the construction of a program which can pinpoint the most ideal time to buy and sell. Similarly, AI is applied in this experiment to predict the future price of cryptocurrencies on a number of different blockchains, including the Electro-Optical System and Ethereum.

Long short-term memory (LSTM) is a neural network (form of AI) which ingests information and processes data using a *gradient-based learning algorithm*[7]. This creates an algorithm that improves with additional parameters; the algorithm *learns* as it ingests. LSTM neural networks will be employed to analyze pre-existing price data so that the model can attempt to generate the future price in varying timetables, such as ten days, several months, or a year from the last date. This project will provide as a boon for insights into investments with potentially great returns. These findings can contribute to a positive cycle of attracting investors to a coin, which results in a price increase, which repeats. The main objective is to provide insights for investors on an up-and-coming product: cryptocurrency.

2. Datasets

This project utilizes yfinance, a Python module which downloads the historical prices of a cryptocurrency from the first day of its inception to whichever day the program is executed. For example, the Yahoo Finance page for EOS-USD is the source for Figure 1[8]. Figure 1 shows the historical data on a line graph when the program receives "EOS-USD" as an input.

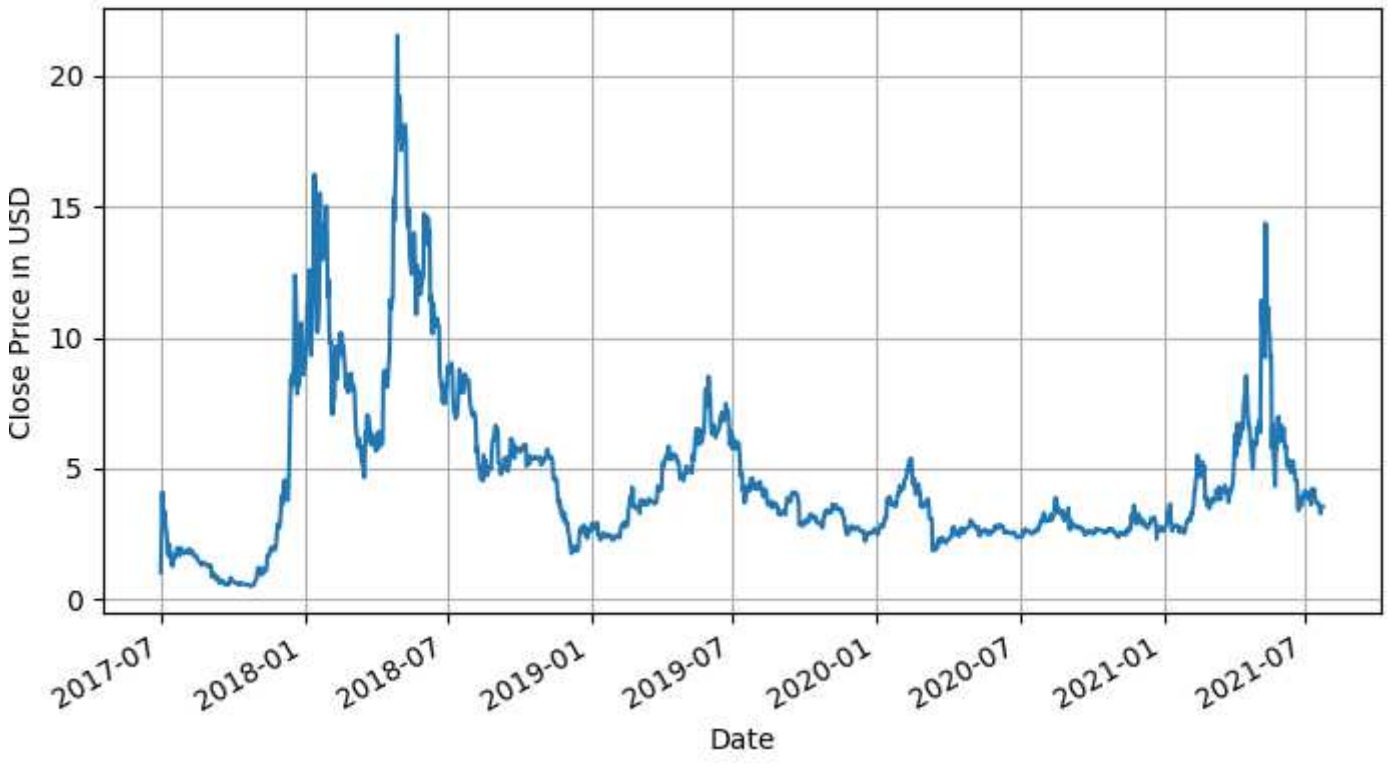


Figure 1: Line graph of EOS price from 1 July 2017 to 22 July 2021. Generated using yfinance-lstm.ipynb[2] located in project/code, utilizing price data from Yahoo Finance[8].

3. Architecture

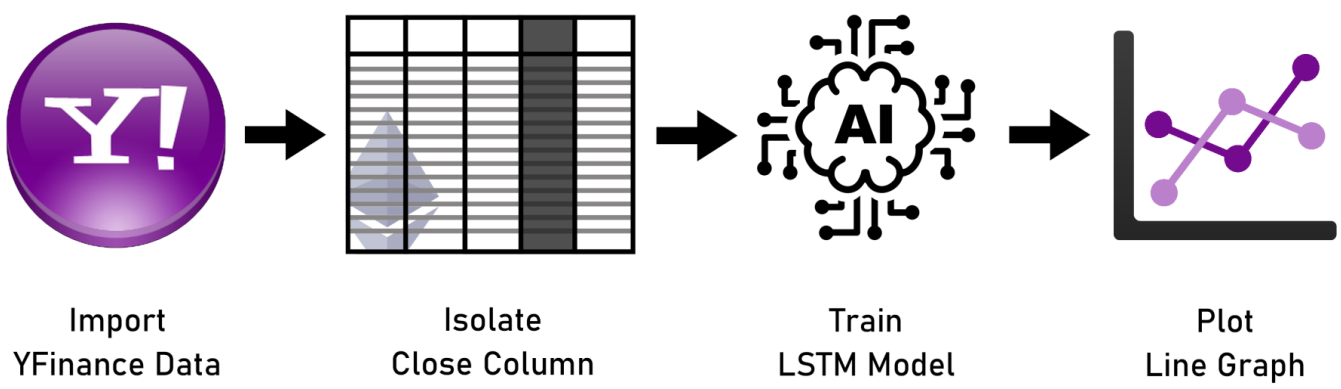


Figure 2: The process of producing LSTM timeseries based on cryptocurrency price.

This program undergoes the four main phases outlined in Figure 2: retrieving data from Yahoo Finance[8], isolating the Close prices (the price the cryptocurrency has at the end of each day), training the LSTM to predict Close prices, and plotting the prediction model, respectively.

4. Implementation

Initially, this project was meant to scrape prices using the BeautifulSoup Python module; however, slight changes in a financial page's website caused the code to break. Alternatively, Kaggle offered historical datasets of cryptocurrency, but they were not up to date. Thus, the final method of retrieving data is from Yahoo Finance through the yfinance Python module, which returns the coins' price from the day to its inception to the present day.

The code is inspired from Towards Data Science articles by Serafeim Loukas[9] and Viraf[10], who explore using LSTM to predict stock timeseries. This program contains adjustments and changes to their code so that cryptocurrency is analyzed instead. This project opts to use LSTM (long short-term memory) to predict the price because it has a memory capacity, which is ideal for a timeseries data set analysis such as cryptocurrency price over time. LSTM can remember historical patterns and use them to inform further predictions; it can also selectively choose which datapoints to use and which to disregard for the model[11]. For example, this experiment's code isolates only the close values to predict them and nothing else.

Firstly, the code asks the user for the ticker of the cryptocurrency that is to be predicted, such as EOS-USD or BTC-USD. A complete list of acceptable inputs is under the Symbol column at https://finance.yahoo.com/cryptocurrencies but theoretically, the program should be able to analyze traditional stocks as well as cryptocurrency.

Then, the program downloads the historical data for the corresponding coin through the yfinance Python module. The data must go through normalization for simplicity and optimization of the model. Next, the Close data (the price that the currency has at the end of the day, everyday since the coin's inception) is split into two sets: a training set and a test set, which are further split into their own respective x and y sets to guide the model through training.

The training model is run through a layer of long short-term memory, as well as a dropout layer to prevent overfitting and a dense layer to give the model a memory capacity. Figure 3 showcases the setup of the LSTM layer.

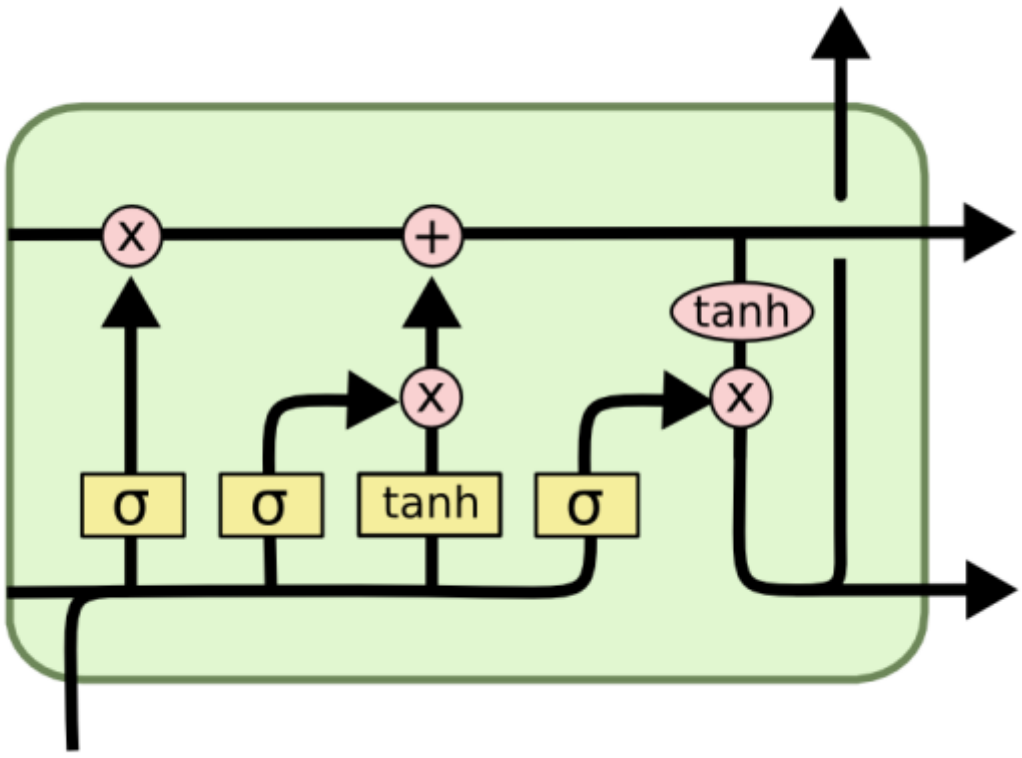


Figure 3: Visual depiction of one layer of long short-term memory[12]

After training through 50 epochs, the program generated Figure 4, a line graph of the prediction model. Unless otherwise specified, the following figures use the EOS-USD data set from July 1st, 2017 to July 26th, 2021. Note that only the last 200 days are predicted so that the model can analyze the preexisting data prior to the 200 days for training purposes.

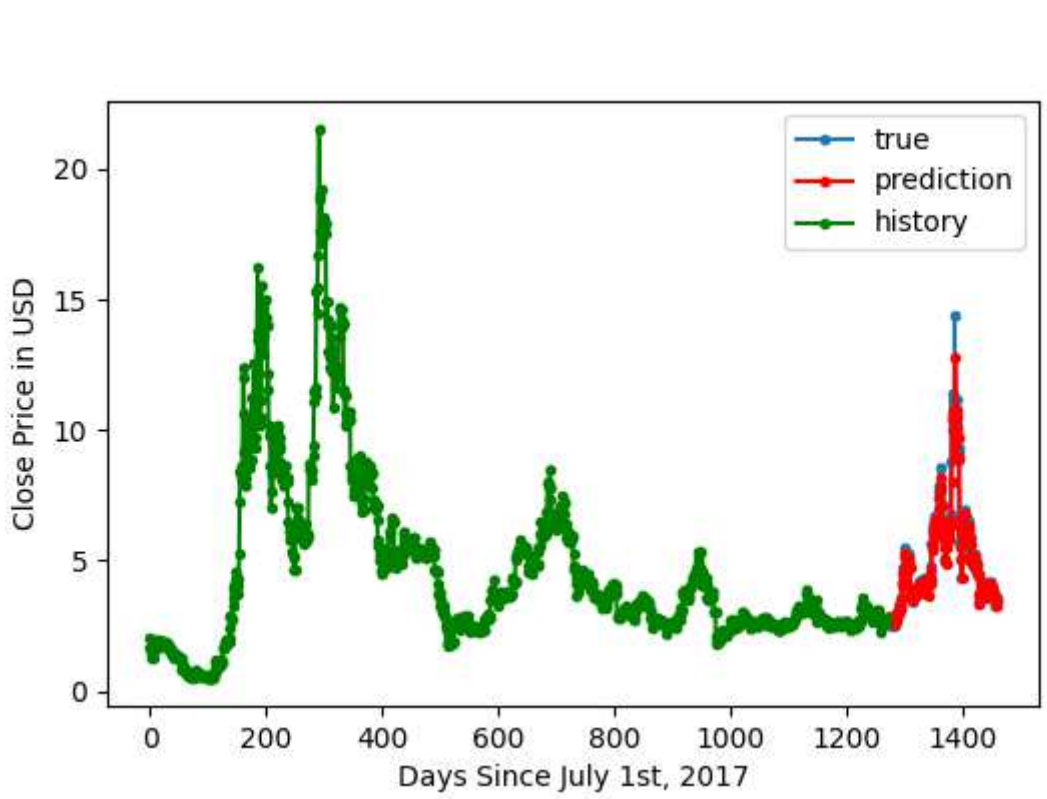


Figure 4: EOS-USD price overlayed with the latest 200 days predicted by LSTM

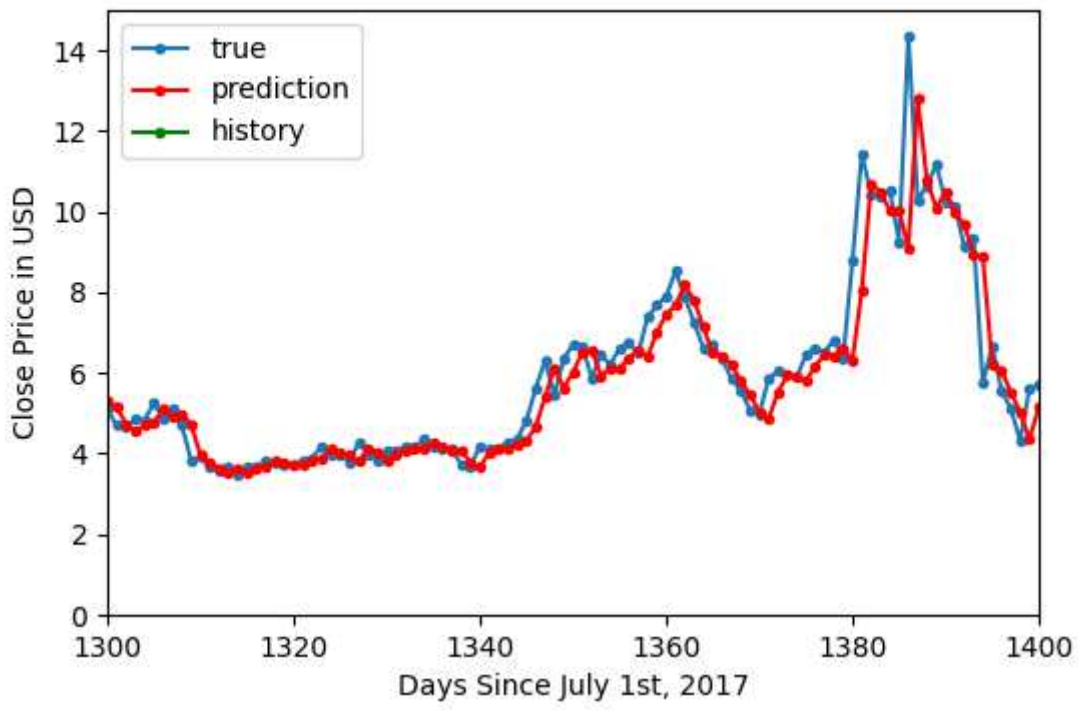


Figure 5: Zoomed-in graph (same as Figure 4 but scaled x and y-axis for readability)

During training, the number of epochs can affect the model loss. According to the following figures 6 and 7, the loss starts to minimize around the 30th epoch of training. The greater the number of epochs, the sharper and more accurate the prediction becomes, but it does not vastly improve after around the 30th epoch.

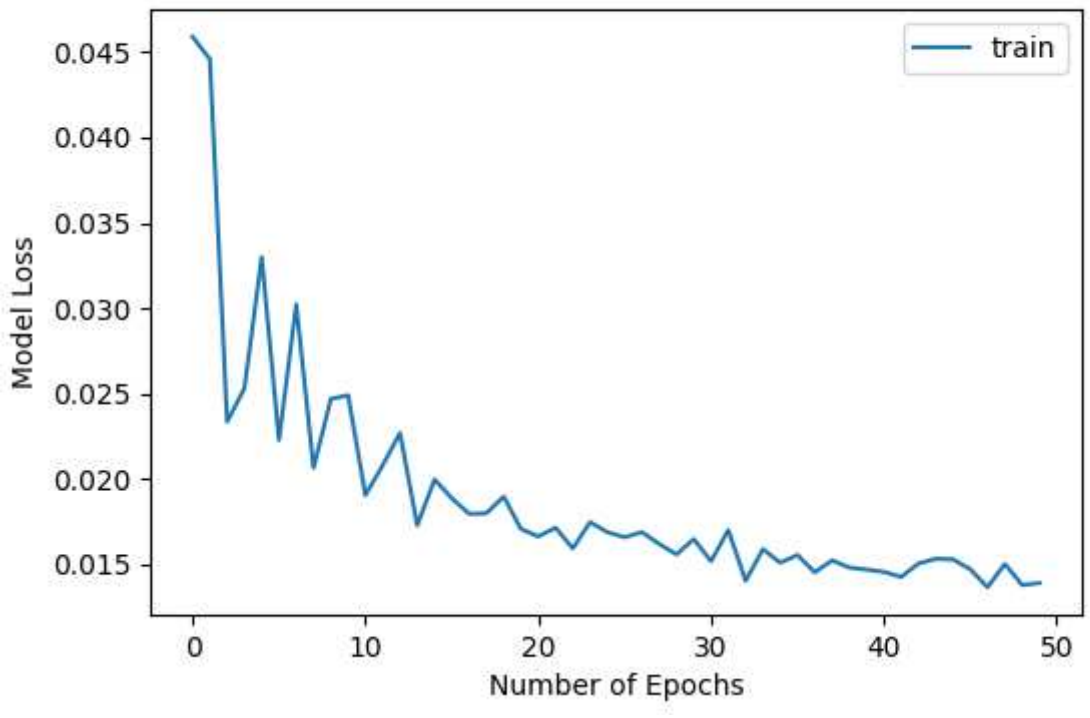


Figure 6: Line graph of model loss over the number of epochs the prediction model completed using EOS-USD data set

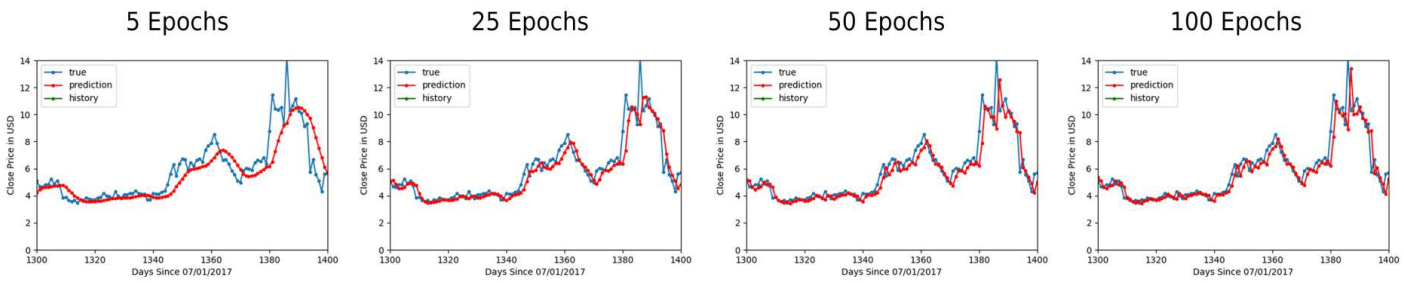


Figure 7: Effect of EOS-USD prediction model based on number of epochs completed

The epochs can also affect the Mean Squared Error, which details how close the prediction line is to the true Close values in United States Dollars (USD). As demonstrated in Table 1, more epochs lessens the Mean Squared Error (but the change becomes negligible after 25 epochs).

Table 1: Number of epochs compared with Mean Squared Error; all tests were run with EOS-USD as input. The Mean Squared Error is rounded to the nearest thousandth.

Epochs	Mean Squared Error
5	0.924 USD
15	0.558 USD
25	0.478 USD
50	0.485 USD
100	0.490 USD

Lastly, cryptocurrencies other than EOS such as Dogecoin, Ethereum, and Bitcoin can be analyzed as well. Figure 8 demonstrates the prediction models generated for these cryptocurrencies.

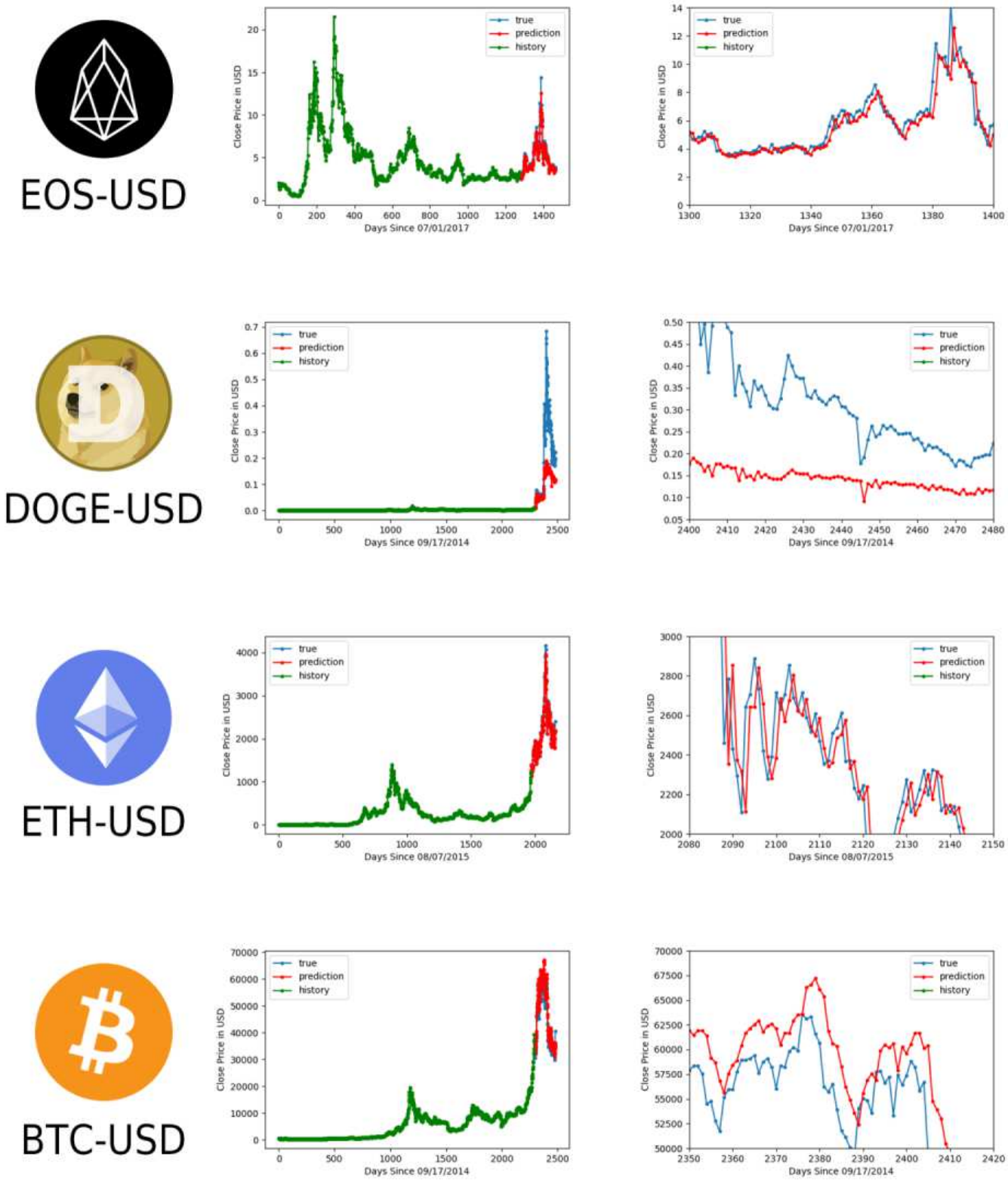


Figure 8: EOS, Dogecoin, Ethereum, and Bitcoin prediction models

Dogecoin presents a model that does not account well for the sharp rises, likely because the training period encompasses a period of relative inactivity (no high changes in price).

5. Benchmark

The benchmark is run within yfinance-lstm.ipynb located in project/code[2]. The program ran on a 64-bit Windows 10 Home Edition (21H1) computer with a Ryzen 5 3600 processor (3.6 GHz). It also has dual-channel 16 GB RAM clocked at 3200 MHz and a GTX 1660 Ventus XS OC graphics card. Table 2 lists these specifications as well as the allocated computer memory during runtime and module versions. Table 3 shows that the amount of time it takes to train the 50 epochs for the LSTM is around 15 seconds, while the entire program execution takes around 16 seconds. A StopWatch module was used from the package cloudmesh-common[13] to precisely measure the training time.

Table 2: First half of cloudmesh benchmark output, which details the specifications and status of the computer at the time of program execution

Attribute	Value
cpu_cores	6
cpu_count	12
cpu_threads	12
frequency	scpufreq(current=3600.0, min=0.0, max=3600.0)
mem.available	7.1 GiB
mem.free	7.1 GiB
mem.percent	55.3 %
mem.total	16.0 GiB
mem.used	8.8 GiB
platform.version	(‘10’, ‘10.0.19043’, ‘SP0’, ‘Multiprocessor Free’)
python	3.9.5 (tags/v3.9.5:0a7dcbd, May 3 2021, 17:27:52) [MSC v.1928 64 bit (AMD64)]
python.pip	21.1.3
python.version	3.9.5
sys.platform	win32
uname.machine	AMD64
uname.processor	AMD64 Family 23 Model 113 Stepping 0, AuthenticAMD
uname.release	10
uname.system	Windows
uname.version	10.0.19043

Table 3: Second half of cloudmesh benchmark output, which reports the execution time of training, overall program, and prediction

Name	Time	Sum	Start	OS	Version
Overall time	16.589 s	35.273 s	2021-07-26 18:39:57	Windows	(‘10’, ‘10.0.19043’, ‘SP0’, ‘Multiprocessor Free’)

Name	Time	Sum	Start	OS	Version
Training time	15.186 s	30.986 s	2021-07-26 18:39:58	Windows	(‘10’, ‘10.0.19043’, ‘SP0’, ‘Multiprocessor Free’)
Prediction time	0.227 s	0.474 s	2021-07-26 18:40:13	Windows	(‘10’, ‘10.0.19043’, ‘SP0’, ‘Multiprocessor Free’)

6. Conclusion

At first glance, the results look promising as the predictions have minimal deviation from the true values. However, upon closer look, the values lag by one day, which is a sign that they are only viewing the previous day and mimicking those values. Furthermore, the model cannot go several days or years into the future because there is no data to run on, such as opening price or volume. The experiment is further confounded by the nature of stock prices: they follow random walk theory, which means that the nature in which they move follows a random walk: the changes in price do not necessarily happen as a result of previous changes. Thus, this nature of stocks contradicts the very architecture of this experiment because long short-term memory assumes that the values have an effect on one another.

For future research, a program can scrape tweets from influencers' Twitter pages so that a model can guess whether public discussion of a cryptocurrency is favorable or unfavorable (and whether the price will increase as a result).

7. Acknowledgments

Thank you to Dr. Gregor von Laszewski, Dr. Yohn Jairo Parra Bautista, and Dr. Carlos Theran for their invaluable guidance. Furthermore, thank you to Florida A&M University for graciously funding this scientific excursion and Miami Dade College School of Science for this research opportunity.

8. References

1. Jacques Fleischer, README.md Install Documentation, [GitHub] https://github.com/cybertraining-dsc/su21-reu-361/blob/main/project/code/README.md ↩
2. Jacques Fleischer, yfinance-lstm.ipynb Jupyter Notebook, [GitHub] https://github.com/cybertraining-dsc/su21-reu-361/blob/main/project/code/yfinance-lstm.ipynb ↩
3. Marco Iansiti and Karim R. Lakhani, The Truth About Blockchain, [Online resource] https://hbr.org/2017/01/the-truth-about-blockchain ↩
4. Lori Schock, Thinking About Buying the Latest New Cryptocurrency or Token?, [Online resource] https://www.investor.gov/additional-resources/spotlight/directors-take/thinking-about-buying-latest-new-cryptocurrency-or ↩
5. Jeremy Swinfen Green, Understanding cryptocurrency market fluctuations, [Online resource] https://www.telegraph.co.uk/business/business-reporter/cryptocurrency-market-fluctuations/ ↩
6. Raj Shroff, When Blockchain Meets Artificial Intelligence. [Online resource] https://medium.com/swlh/when-blockchain-meets-artificial-intelligence-e448968d0482 ↩
7. Sepp Hochreiter and Jürgen Schmidhuber, Long Short-Term Memory, [Online resource] https://www.bioinf.jku.at/publications/older/2604.pdf ↩
8. Yahoo Finance, EOS USD (EOS-USD), [Online resource] https://finance.yahoo.com/quote/EOS-USD/history?p=EOS-USD ↩

9. Serafeim Loukas, Time-Series Forecasting: Predicting Stock Prices Using An LSTM Model, [Online resource] https://towardsdatascience.com/lstm-time-series-forecasting-predicting-stock-prices-using-an-lstm-model-6223e9644a2f ↩

10. Viraf, How (NOT) To Predict Stock Prices With LSTMs, [Online resource] https://towardsdatascience.com/how-not-to-predict-stock-prices-with-lstms-a51f564ccbca ↩

11. Derk Zomer, Using machine learning to predict future bitcoin prices, [Online resource] https://towardsdatascience.com/using-machine-learning-to-predict-future-bitcoin-prices-6637e7bfa58f ↩

12. Christopher Olah, Understanding LSTM Networks, [Online resource] https://colah.github.io/posts/2015-08-Understanding-LSTMs/ ↩

13. Gregor von Laszewski, Cloudmesh StopWatch and Benchmark from the Cloudmesh Common Library, [GitHub] https://github.com/cloudmesh/cloudmesh-common ↩

Investigating the Classification of Breast Cancer Subtypes using KMeans

This project provides an insight into an investigation of the classification of breast cancer sub-types using proetomic dataset through a machine learning approach.

Tags: project reu ai health

13 minute read

Check Report failing Status failing Status: draft, Type: Project

Kehinde Ezekiel, su21-reu-362, Edit

Abstract

Breast cancer is an heterogenous disease that is characterized by abnormal growth of the cells in the breast region[^1]. There are four major molecular subtypes of breast cancer. This classification was based on a 50-gene signature profiling test called PAM50. Each molecular subtype has a specific morphology and treatment plan. Early diagnosis and detection of possible cancerous cells usually increase survival chances and provide a better approach for treatment and management. Different tools like ultrasound, thermography, mammography utilize approaches like image processing and artificial intelligence to screen and detect breast cancer. Artificial Intelligence (AI) involves the simulation of human intelligence in machines and can be used for learning or to solve problems. A major subset of AI is Machine Learning which involves training a piece of software (called model) to makwe useful predictions using dataset.

In this project, a machine learning algorithm, KMeans, was implemented to design and analyze a proteomic dataset into clusters using its proteins identifiers. These protein identifiers were associated with the PAM50genes that was used to originally classify breast cancer into four molecular subtypes. The project revealed that further studies can be done to investigate the relationship between the data points in each cluster with the biological properties of the molecular subtypes which could lead to newer discoveries and developmeny of new therapies, effective treatment plan and management of the disease. It also suggests that several machine learning algorithms can be leveraged upon to address healthcare issues like breast cancer and other diseases which are characterized by subtypes.

Contents

Keywords: AI, cancer, breast, algorithms, machine learning, healthcare, subtypes, classification.

1. Introduction

Breast cancer is the most common cancer, and also the primary cause of mortality due to cancer in females around the World. It is an heterogenous disease that is characterized by the abnormal growth of cells in the breast region[1]. Early diagnosis and detection of possible cancerous cells in the breast usually increase survival chances and provide a better approach for treatment and management. Treatment and management often depend on the stage of cancer, the subtype, the tumor size, location and many other factors. During the last 20 years, four major intrinsic molecular subtypes for breast cancer- luminal A, luminal B, HER2-enriched and Basal-like have been identified, classified and intensively studied. Each subtype has its distinct morphologies and clinical treatment. The classification is based on gene expression profiling, specificaly defined by mRNA expression of 50 genes (also known as, PAM50 Genes). This test is known as the PAM50 test. The accurate grouping of breast cancer into its relevant subtypes can improve accurate treatment-decision making[2]. The PAM50 test is now known as the Prosigna Breast Cancer Prognostic Gene Signature Assay 50 (known as Prosigna) and it analyzes the activity of certain genes in early-stage, hormone-receptor-positive breast cancer[3]. This classification is based on the mRNA expression and the activity of 50 genes and it aims to estimate the risk of distant reccurrence of breast cancer. Since the assay was based on mRNA expression, this project suggested that a classification based on the final product of mRNA, that is protein, can be implemented to investigate its role in the classifictaion of molecular breast cancer subtypes. As a result, the project was focused on the use of a proteomic dataset which contained published iTRAQ proteome profiling of 77 breast cancer samples and expression values for the proteins of each sample.

Most times, breast cancer is diagnosed and detected through a combination of different approaches such as imaging (e.g. mammogram and ultrasound), physical examination by a radiologist and biopsy. Biopsy is used to confirm the breast cancer symptoms. However, research has shown that radiologists can miss up to 30% of breast cancer tissues during detection[4]. This gap has brought about the introduction of Computer aided Diagnosis (CAD) systems can help detect abnormalities in an efficient manner. CAD is a technology that includes utilizing the concept of artificial intelligence(AI) and medical image processing to find abnormal signs in the human body[5]. Machine Learning is a subset of AI and it has several algorithms that can be used to build a model to perform a specific task or to predict a pattern. KMeans is one of such algorithm.

Building a model using machine learning involves selecting and preparing the appropriate dataset, identifying the accurate machine learnning algorithm to use, training the algorithm on the data to build a model, validating the resulting model's performance on testing data and using the model on a new data[6]. In this project, KMeans was the algorithm used in this project, the datasets were prepared through several procedures like filtering, merging. KMeans clustering method was used to investigate the classification of the molecular subtypes. Its efficacy is often tested by a silhouette score. A silhouette score shows how similar an object is to its own cluster and it ranges from -1 to 1 where a high values indicates that an object is well matched to its own cluster. A homogeneity score determines if a cluster should only contain samples that belong to a particular class. It ranges from a value between 0 to 1 with low values indicating a low homogeneity.

The project investigated the efficient number of clusters that could be generated for the proteome dataset which would consequetly provide an optimal classification of the protein expression values for the breast cancer samples. The proteins that were used in the KMeans analysis were the proteins that were associated with the PAM50 genes. The result of the project could provide insights to medical scientists and researchers to identify any interelatedness between the original classification of breast cancer molecular subtypes.

2. Datasets

Datasets are eseential in drawing conclusion. In the diagnosis, detection and classification of breast cancecr, datasets have been essential to draw conclusion by identifying patterns. These datasets range from imaging datasets to clinical datasets, proteomic datasets etc. Large amounts of data have been collected due to new technological and computational advances like the use of websites like NCBI, equipments like Electroencephalogram (EEG) which record clinical

information. Medical researchers leverage these datasets to make useful health care decisions that affect a region, gender or the world. The need for accuracy and reproducibilty has led to the use of machine learning as an important tool for drawing conclusions.

Machine Learning involves training a piece of software, also known as model, to idnetify patterns from a dataset and make useful predictions. There are several factors to be considered when using datasets. One of such is data privacy. Recently, measures have been taken to ensure that the privacy of data. Some of these measures include, replacing codes for patients name, using documents and mobile applications that ask for permission from patients before using their data. Recently, the World Health Organization (WHO) made her report on AI and provided priniples that ensure that AI works for all. On of such is that the designer of AI technologies should satisfy regulatory requirements for safety, accuracy and efficacy for well-defined use cases or indications. Measures of quality control in practice and quality improvement in the use of AI must be available[7]. Building a model using machine learning involves selecting and preparing the appropriate dataset, identifying the accurate machine learnning algorithm to use, training the algorithm on the data to build a model, validating the resulting model's performance on testing data and using the model on a new data[6]. In this project, KMEans was the algorithm used in this project, the datasets were prepared through several procedures like filtering, merging.

3. The KMeans Approach

KMeans clustering is an unsupervised machine learning algorithm that makes inferences from datasets without referring to a known outcome. It aims to identify underlying patterns in a dataset by looking for a fixed number of clusters, (known as k). The required number of clusters is chosen by the person building the model. KMeans was used in this project to classify the protein IDs (or RefSeq_IDs) into clusters. Each cluster was designed to be associated with related protein IDs.

Three datasets were used for the algorithm. The first and main dataset was a proteomic dataset. It contained published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). Each sample contained expression values for ~12000 proteins, with missing values present when a given protein could not be quantified in a given sample. The variables in the dataset included the RefSeq_accession_number(also known as RefSeq protein ID), "the gene_symbol" (which was unique to each gene), "the gene_name" (which was the full name of the gene). The remaining columns were the log2 iTRAQ ratios for each of the 77 samples while the last three columns are from healthy individuals.

The second dataset was a PAM50 dataset. It contained the list of genes and proteins used in the PAM50 classification system. The variables include the RefSeqProteinID which matched the Protein IDs(or RefSeq_IDs) in the main proteome dataset.

The third dataset was a clinical data of about 105 clinical breast cancer samples. 77 of the breast cancer samples were the samples in the first dataset. The excluded samples were as a result of protein degradation[8]. The variables in the dataset are: 'Complete TCGA ID', 'Gender', 'Age at Initial Pathologic Diagnosis', 'ER Status', 'PR Status', 'HER2 Final Status', 'Tumor', 'Tumor–T1 Coded', 'Node', 'Node-Coded', 'Metastasis', 'Metastasis-Coded', 'AJCC Stage', 'Converted Stage', 'Survival Data Form', 'Vital Status', 'Days to Date of Last Contact', 'Days to date of Death', 'OS event', 'OS Time', 'PAM50 mRNA', 'SigClust Unsupervised mRNA', 'SigClust Intrinsic mRNA', 'miRNA Clusters', 'methylation Clusters', 'RPPA Clusters', 'CN Clusters', 'Integrated Clusters (with PAM50)', 'Integrated Clusters (no exp)', 'Integrated Clusters (unsup exp).'

During the preparation of the datasets for KMeans analysis, unused columns like "gene_name" and "gene_symbol" were removed in the first dataset. The first and third dataset were merged together. Prior to merging, the variable 'Complete TCGA ID' in the third dataset was found to be the same as the TCGAs in the first dataset. The Complete TCGA ID refered to a breast cancer patient, some patients were found in both datasets. The TCGA ID in the first dataset was renamed to match with the TCGA of the third dataset, thereby giving the same syntax. The first dataset was also transposed as a row and its gene expression as the columns. These processes were done in order to merge both dataset efficiently.

After merging, the “PAM5O RNA” variable from the second dataset was selected to join the merged dataset. This single dataset was named “pam50data”. It contained all the variables that were needed for KMeans Analysis which included the genes that were used for the PAM50 classification (only 43 were available in the dataset), the complete TCGA ID of each 80 patient, and their molecular tumor type. Missing values in the dataset were imputed using SimpleImputer. Then, KMeans clustering was performed. The metrics were tested with cluster numbers of 3, 4, 5, 20 and 79. The bigger numbers (20 and 79) were tested just for comparison. Further details on the codes written can be found in [9]. Also, [10] and [11] were kernels that provided insights for the written code.

5. Results and Images

Several codes were written to determine the best number of clusters for the model. The effectiveness of a cluster is often measured by scores such as silhouette score, homogeneity score and adjusted rand score.

The silhouette score for a cluster of 3, 4, and 5, 8, 20 and 79 were 0.143, 0.1393, 0.1193, 0.50968, 0.0872, 0.012 while the homogenenity scores were 0.4635, 0.4749, 0.1193, 0.5617, 0.6519 and 1.0 respectively. The homogeneity score for 79 is 1.0 since the algorithm can assign all the points into sepearate clusters. However, it is not efficient for the dataset we used. A cluster of 3 works best since the silhouette score is high and the homogeneity score jumps ~2-fold.

Figures 1 and 2 show the results of the visualization of the clusters of 3 and 4.

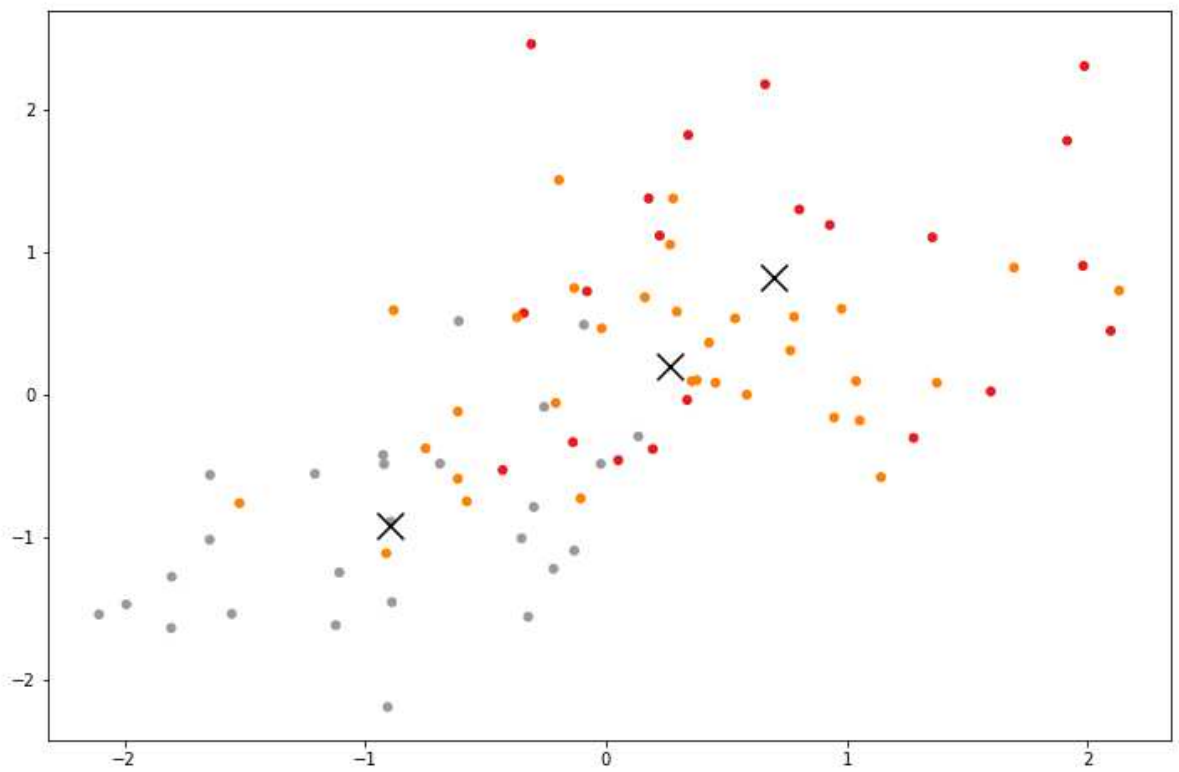


Figure 1: The classification of Breast Cancer Moleecular Subtypes using KMeans Clustering. (k=3). Each data point represnt the expression value for the genes that were used for clustering.

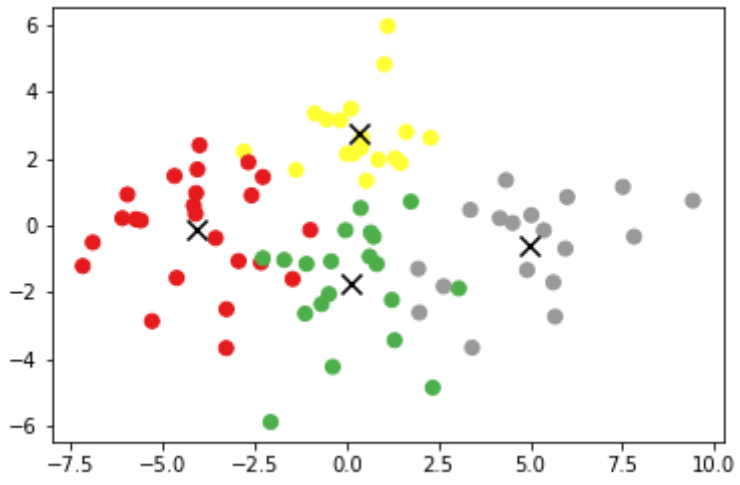


Figure 2: The classification of Breast Cancer Moleecular Subtypes using KMeans Clustering. (k=4). Each data point represnt the expression value for the genes that were used for clustering.

6. Benchmark

This program was executed on a Google Colab server and the entire runtime took 1.012 seconds Table 1 lists the amount of time taken to loop for n_components. The n_components is gotten from the code and it refers to the features of the dataset.

Name	Status	Time(s)
parallel 1	ok	0.647
parallel 3	ok	0.936
parallel 5	ok	0.952
parallel 7	ok	0.943
parallel 9	ok	1.002
parallel 11	ok	0.991
parallel 13	ok	0.958
parallel 15	ok	1.012

Benchmark: The table shows the parallel process time take the for loop for n_components.

7. Conclusion

The results of the KMeans analysis showed that three clusters provided an optimal result for the classification using a proteomic dataset. A cluster of 3 provided a balanced silhouette and homogeneity score. This predict that some interrelatedness could exist between the original PAM50 subtype classfication, since the result of classifying a protein dataset using a machine learning algorithm identified a cluster of 3 as one with the optimal result. Also, future research could be done by using other machine learning algorithms, possibly a supervised learning algotithm, to identify the correlation between the clusters and the four molecular subtypes. This model can be improved on and if proven to show that there truly exist a relationship between the four molecular subtypes, more research could be done to identify the factors that contribure to the interelatedness. This would lead medical scientists and researchers to work on better innovative methods that will aid the treatment and management of breast cancer.

8. Acknowledgments

This projected was immensely supported by Dr. Gregor von Laszewski. Also, a big appreciation to the REU Instructors (Carlos Theran, Yohn Jairo and Victor Adankai) for their contribution, support, teachings and advice. Also, gratitude to my colleagues who helped me out; Jacques Fleischer, David Umanzor and Sheimy Paz Serpa. gratitude to my colleagues. Lastly, appreciation to Dr. Byron Greene, the Florida A&M University, the Indiana University and Bethune Cookman University for providing a platform to be able to learn new things and embark on new projects.

9. References

1. Akram, Muhammad et al. “Awareness and current knowledge of breast cancer.” Biological research vol. 50,1 33. 2 Oct. 2017, doi:10.1186/s40659-017-0140-9 ↩

2. Wallden, Brett et al. “Development and verification of the PAM50-based Prosigna breast cancer gene signature assay.” BMC medical genomics vol. 8 54. 22 Aug. 2015, doi:10.1186/s12920-015-0129-6 ↩

3. Breast Cancer.org Prosigna Breast Cancer Prognostic Gene Signature Assay. https://www.breastcancer.org/symptoms/testing/types/prosigna

4. L. Hussain, W. Aziz, S. Saeed, S. Rathore and M. Rafique, “Automated Breast Cancer Detection Using Machine Learning Techniques by Extracting Different Feature Extracting Strategies,” 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 2018, pp. 327-331, doi: 10.1109/TrustCom/BigDataSE.2018.00057.

5. Halalli, Bhagirathi et al. “Computer Aided Diagnosis - Medical Image Analysis Techniques.” 20 Dec. 2017, doi: 10.5772/intechopen.69792

6. Salod, Zakia, and Yashik Singh. “Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol.” Journal of public health research vol. 8,3 1677. 4 Dec. 2019, doi:10.4081/jphr.2019.1677Articles

7. WHO, WHO issues first global report on Artificial Intelligence (AI) in health and six guiding principles for its design and use. https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use

8. Mertins, Philipp et al. “Proteogenomics connects somatic mutations to signalling in breast cancer.” Nature vol. 534,7605 (2016): 55-62. doi:10.1038/nature18003

9. Kehinde Ezekiel, Project Code, https://github.com/cybertraining-dsc/su21-reu-362/blob/main/project/code/final_breastcancerproject.ipynb

10. Kaggle_breast_cancer_proteomes «https://pastebin.com/A0Wj41DP>

11. Proteomes_clustering_analysis https://www.kaggle.com/shashwatwork/proteomes-clustering-analysis ↩

Report: AI in Orthodontics

In this effort we are analyzing X-ray images in AI and identifying cavitites

Tags: report reu ai health

4 minute read

Check Report passing Status passing Status: final, Type: Report

Whitney McNair, su21-reu-363, Edit

Abstract

In this effort we are analyzing X-ray images in AI and identifying cavitites

Contents

Keywords: ai, orthodontics, x-rays.

1. Introduction

Dental field technology capability has increased over the past 25 years, and has helped reduce time, cost, medical errors, and dependence on human expertise. Intelligence in orthodontics can learn, build, remember, understand and recognize designs from techniques used in correcting the teeth like retainers. Dental field can create alternatives, adapt to change and explore experiences with sub-groups of patients. AI has taken part of the dental field by accurately and efficiently processing the best data from treatments. For smart use of Health Data, machine learning and artificial intelligence are expected to promote further development of the digital revolution in (dental) medicine, like x-rays, using algorithms to simulate human cognition in the analysis of complex data. The performance is better, the higher the degree of repetitive pattern and the larger the amount of accessible data[1].

2. Data Sets

We found a dataset on a kaggle website that is about dental images. The data was collected by Mr. Parth Chokhra. The name of the dataset is Dental Images of kjbjl. The dataset did not have metadata and an explanation of how they collected the data. The data set supports how x-rays of teeth in dentistry becomes artificial intelligence. The Dental Images of kjbjl dataset was used in AI already using autoencoders. Autoencoders are an freely artificial neural network (located in the nervous system) that learns how to accurately encode data and reconstruct the data back from the reduced encoded depiction to a representation that is closes to the original. For some challenges with Orthodontics data sets with privacy, size, avalibility were surprisingly hard to find than we thought.

3. Figures

Below we observed actual dental x-rays. These dental x-rays images below came from Parth Chorkhra on kaggle.com [2]. The images are patient x-rays taken by Parth in his dental imagery data set. In the images we can see the caps and nerves of the teeth. Using these x-rays, we may can also find cavities if there are some. We can also identify other issues with patients teeth by taking and using x-rays.

Figure 1: First x-ray

Figure 2: Second x-ray

Figure 3: Third x-ray

4. Example of a AI algorighm in Orthodontics

On a separate kaggle website, we found a code for DENTAL PANORAMIC KNN [1]. The kaggle site shows dental codes taken place in Orthodontics.

5. Benchmark

Here is an algorithm/code from one of the researchers we found. are using to study the performance of their algorithms or code.

Lateral and frontal facial images of patients who visited the Orthodontic department (352 patients) were employed as the training and evaluation data. An experienced orthodontist examined all the facial images for each patient and identified as many clinically used facial traits during the orthodontic diagnosis process as possible (e.g., deviation of the lips, deviation of the mouth, asymmetry of the face, concave profile, upper lip retrusion, presence of scars). A sample patient's image, a list of sample assessments (i.e., labels), and the multi-label data used in the work by Murata et al. are shown in Figure 4.

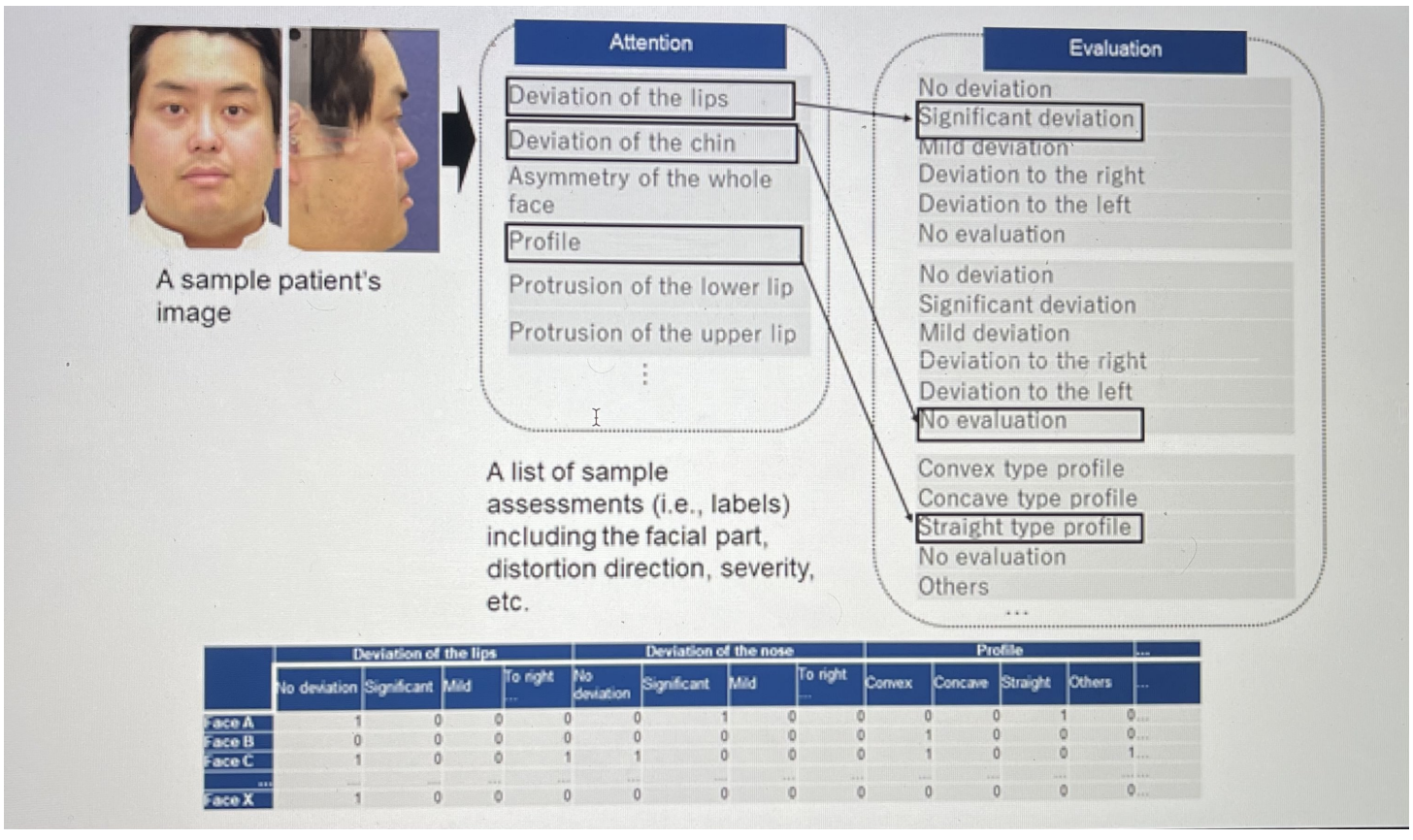


Figure 4: Sample patient's image and a list of sample assessments (i.e., labels) including the region of interest, evaluation, etc. In a previous study by Murata et al., they employed labels representing only the facial part (mouth, chin, and whole face), distorted direction (right and left), and its severity (severe, mild, no deviation).

6. Conclusion

Artificial intelligence is rapidly expanding into multiple facets of society. Orthodontics may be one of the fastest branches of dentistry to adapt AI for three reasons. First, patient encounters during treatment generate many types of data. Second, the standardization in the field of dentistry is low compared to other areas of healthcare. A range of valid treatment options exists for any given case. Using AI and large datasets (that include diagnostic results, treatments, and outcomes), one can now measure the effectiveness of different treatment modalities given very specific clinical findings and conditions. Third, orthodontics is largely practiced by independent dentists in their own clinics. Despite the promise of AI, the volume of orthodontic research in this field is relatively low. Further, the clinical accuracy of AI must be improved with an increased number and variety of cases. Before AI can take on a more important role in making diagnostic recommendations, the volume and quality of research data will need to increase[1].

7. Acknowledgments

Dr. Gregor von Laszewski, Carlos and Yohn guided me throughout this process.

8. References

1. Hasnitadita. (2021, July 10). DENTAL panoramic knn. Kaggle. https://www.kaggle.com/hasnitadita/dental-panoramic-knn ↩

2. Chokhra, P. (2020, June 29). Medical image dataset. Kaggle. https://www.kaggle.com/parthplc/medical-image-dataset ↩

Project: Hand Tracking with AI

In this project we study the ability of an AI to recognize letters from the American Sign Language (ASL) alphabet. We use a Convolutional Neural Network and apply it to a dataset of hands in different positionings showing the letters ‘a’, ‘b’, and ‘c’ in ASL. With this we build a model to recognize the letter and output the letter it predicts.

Tags: project reu ai communication

8 minute read

Check Report passing Status failing Status: draft, Type: Project

David Umanzor, su21-reu-364, Edit

Abstract

In this project, we study the ability of an AI to recognize letters from the American Sign Language (ASL) alphabet. We use a Convolutional Neural Network and apply it to a dataset of hands in different positionings showing the letters ‘a’, ‘b’, and ‘c’ in ASL. The proposed CNN model receives an ASL image and recognizes the feature of the image, generating the predicted letter.

Contents

Keywords: ai, object recognition, image processing, computer vision, american sign language.

1. Introduction

Object detection and feature selection are essential tasks in computer vision and have been approached from various perspectives over the past few decades [1]. The brain uses object recognition to solve an inverse problem: one where (surface properties, shapes, and arrangements of objects) need to be inferred from the perceived outcome of the image formation process [2]. Visual object recognition as a neural substrate in humans was revealed by neuropsychological studies. There are specific brain regions that cause object recognition, yet we still do not understand how the brain achieves this remarkable behavior [3]. Human beings rely and rapidly recognize objects despite considerable retinal image transformations arising from changes in lighting, image size, position, and viewing angle [3].

A gesture is a form of nonverbal communication done with positions and movements of the hand, arms, body parts, hand shapes, movements of the lips or face [4]. One of the key differences of hand gestures is that they allow communication over a long distance [5]. American Sign Language (ASL) is a formal language that has the same lingual properties as oral languages commonly used by deaf people to communicate [6]. ASL typically is formed by the finger, hand,

and arm positioning and can contain static and dynamic movement or a combination of both to communicate words and meanings to another [6]. Communication with other people can be challenging because people are not typically willing to learn sign language [6].

In this paper, we consider the problem of detecting and understanding American Sign Language. We test CNN's ability to recognize the ASL alphabet. As advancements in technology increase, there are more improvements to 2D methods of hand detection. Commonly these methods are visual-based, using color, shape, and edge to detect and recognize the hand [7]. There are issues to these technologies like inconsistent lighting conditions, non-hand color similarity, and varying viewpoints that can decrease the model's ability to recognize the hand and its positioning. We use a Convolutional Neural Network to create the model and detect different letters of American Sign Language.

2. Data Sets

In this research we use two sources of datasets, the first is from kaggle which it was already prepared but we needed more. The second is self made dataset by take images in good lighting against a white wall, it was then cropped to 400x400 pixels focused on the hand. The program then sets the images to grayscale as the color is not needed for this research. Finally, the images are reduced to 50x50 resolution for the AI to use for training.

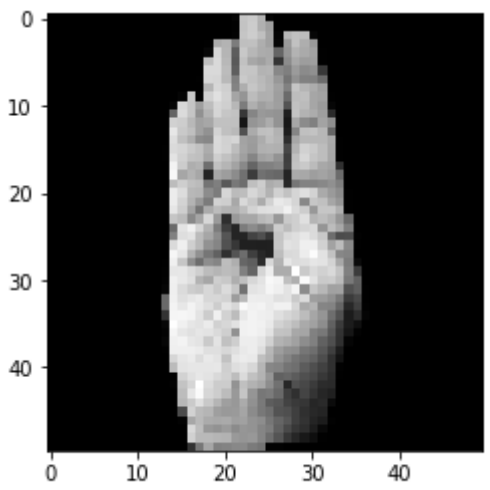


Figure 1: Dataset of hands doing different alphabet letters in ASL [5].

3. Documentation

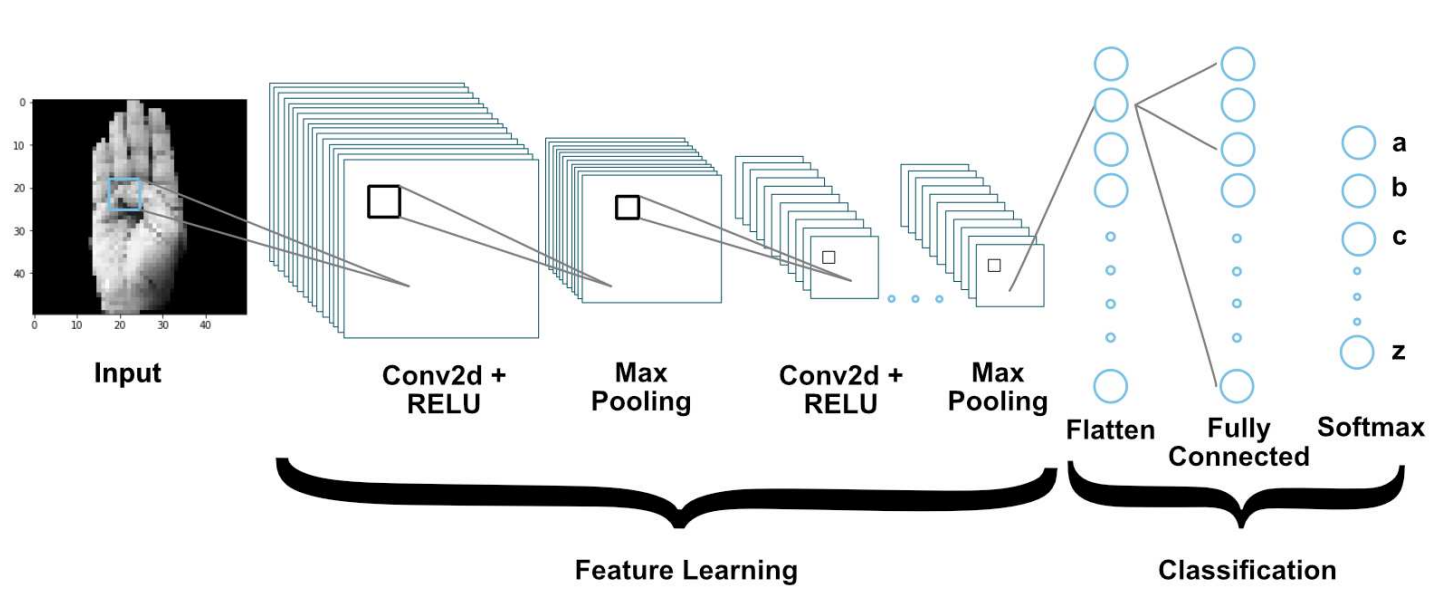


Figure 2: The Convolutional Neural Network (CNN) model.

This model shows the CNN model that we used to train the AI. The CNN takes pictures and breaks them down into smaller segments called features. It is trained to find patterns and features over the images allowing the CNN to predict an 'a', 'b', or 'c' upon the given ASL image with high accuracy. A CNN uses a convolution operation that filters every possible position the

feature it collected can be matched to and attempts to find where it fits in [8]. This process is repeated and becomes the convolution layer or in the image depicted as Conv2d + Relu. The ReLU stands for the rectified linear unit and is used as an activation function for the CNN [9].

- [] What is a Relu operation? ReLU operation is a rectified linear unit and is used as an activation function for the CNN, we use a Leaky ReLU in our model because it is easy to use to train the model quickly and it has a small tolerance for negatives values unlike the normal ReLU fuction. paperswithcode.com/method/leaky-relu add figure of a leaky ReLU
- [] What is a Conv2d? Conv2d is a 2D Convolution layer meant for a images as it uses height and width. They build a filter across the image by recognizing the similarities of the image
- [] What is a BashNormalization operation? Batch Normalization is a process that standardizes the updates as the Convolutional process sets weights and as the neural network goes through each layer the procedure keeps adjusting to a target that never stays the same, requiring more epochs and and reduces the time it takes to train a deep learning neural network. :Reference 12:
- [] what is Maxpooling operation? Maximum pooling is an operation the gathers the biggest number in each collection of each feature map. This provides a way to avoid over-fitting
- [] what is Fully Connected?
- [] what is Softmax operation?

4. Methodology

In this research, we built the model using a convolution neural network (CNN) to create an AI that can recognize ASL letters (‘a’, ‘b’, and ‘c’), using a collection of 282 images. The Dataset contains 94 images for each letter to train the AI’s CNN. This can be expanded to allow an AI to recognize letters, words, and any expression that can be made using a still image of the hands. A CNN fits this perfectly as we can use its ability to assign importance to segments of an image and tell the difference from one another using weights and biases. With the proper training, it is able to learn and identify these characteristics [9].

5. Benchmark

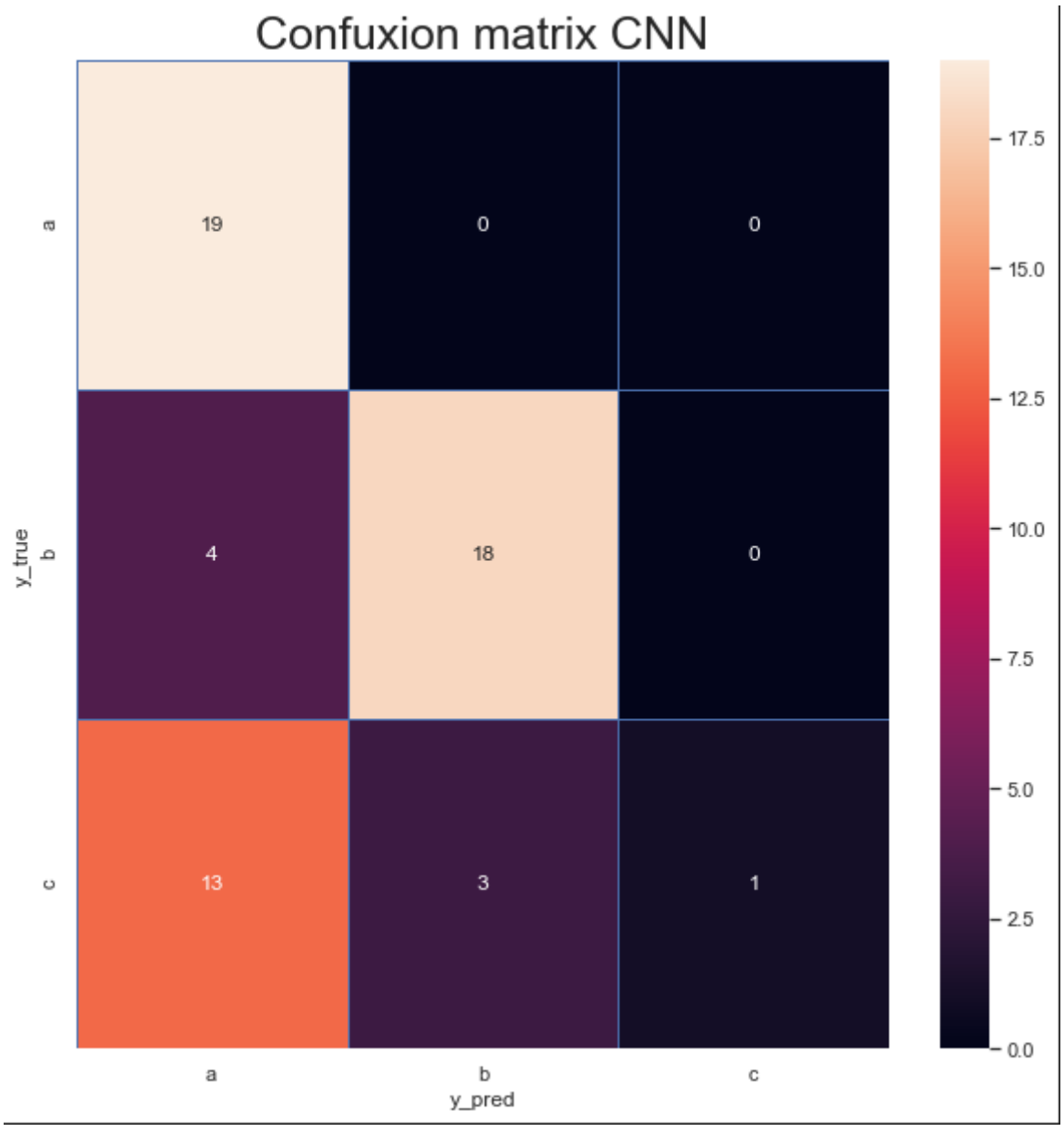


Figure 3: The Confusion Matrix of the finished CNN model.

The Confusion Matrix shows the results of the model after being tested on its ability to recognize each letter, in the image it shows that the AI had a difficult time recognizing the difference between an ‘a’ and ‘c’ only getting 6% of the images labelled as ‘c’ correct.

6. Conclusion

We build a model to recognize an ASL given an image and predict the corresponding letter using a convolutional neural network. The model provides a means of 66% accuracy in classifying the ASL among the three classes ‘a’, ‘b’, and ‘c’. From the given results, the letters ‘a’ and ‘c’ became the most difficult for the CNN to differentiate from each other, as shown in the confusion matrix in figure 3. We suggest that the low accuracy rate is based on similar appearing grayscale of the letters ‘a’ and ‘c’ and the lack of a larger dataset for the AI to learn from. We determine that using a larger dataset of the entire alphabet and increasing the number of examples of each letter to train the AI could improve the results.

We found that the low accuracy can be increased by improving the resolution of the image giving the program more features to go off of in its computing to recognize the image, going from model 1 at 50 x 50 pixels to model 2 at 80 x 80 pixels there was an increase from 66% to 76% in accuracy, this in theory should improve as the resolution of the image increases from 100x 100 to 200x 200 and at the best the image’s resolution would be left at the original size off 400 x 400. This is accuracy increase would be because as the resolution drops the program has less information and some of the important landmarks of the hand are lost due to the resolution of the image.

- ☐ Correct formatting and grammar

Future studies using a larger dataset can be applied to more complex methods than just singular letters but words from the ASL language to recreate a text to speech software based around ASL hand positioning.

7. Acknowledgments

We thank Carlos Theran (Florida A & M University) for advising, guidance, and resources used in the research; We thank Yohn Jairo (Florida A & M University) for guidance and aid on the research report; We thank Gregor von Laszewki (Florida A & M University) for advice and commenting on the code and report; We thank the Polk State LSAMP Program for aid in obtaining this opportunity. We thank Florida A & M University for funding this research.

8. References

1. Pan, T.-Y., Zhang, C., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., & Chao, W.-L. (2021, July 5). On Model Calibration for Long-Tailed Object Detection and Instance Segmentation. arXiv.org. https://arxiv.org/abs/2107.02170. ↩

2. Wardle, S. G., & Baker, C. (2020). Recent advances in understanding object recognition in the human brain: Deep neural networks, temporal dynamics, and context. F1000Research. F1000 Research Ltd. https://doi.org/10.12688/f1000research.22296.1

3. Wardle, S. G., & Baker, C. (2020). Recent advances in understanding object recognition in the human brain: Deep neural networks, temporal dynamics, and context. F1000Research. F1000 Research Ltd. https://doi.org/10.12688/f1000research.22296.1

4. Dabre, K., & Dholay, S. (2014). Machine learning model for sign language interpretation using webcam images. 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), 317-321. https://ieeexplore.ieee.org/document/6839279

5. tecperson, Sign Language MNIST Drop-In Replacement for MNIST for Hand Gesture Recognition Tasks, [Kaggle] https://www.kaggle.com/datamunge/sign-language-mnist

6. A. Rahagiyanto, A. Basuki, R. Sigit, A. Anwar and M. Zikky, “Hand Gesture Classification for Sign Language Using Artificial Neural Network,” 2017 21st International Computer Science and Engineering Conference (ICSEC), 2017, pp. 1-5, <doi: 10.1109/ICSEC.2017.8443898>

7. Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. 2020. RGB2Hands: real-time tracking of 3D hand interactions from monocular RGB video. ACM Trans. Graph. 39, 6, Article 218 (December 2020), 16 pages. https://doi.org/10.1145/3414685.3417852

8. Rohrer, B. (2016, August 18). How do Convolutional Neural Networks work? Library for end-to-end machine learning. https://e2eml.school/how_convolutional_neural_networks_work.html

9. Patel, K. (2020, October 18). Convolution neural networks - a beginner’s Guide. Towards Data Science. https://towardsdatascience.com/convolution-neural-networks-a-beginners-guide-implementing-a-mnist-hand-written-digit-8aa60330d022 ↩

Review: Handwriting Recognition Using AI

This study reviews two approaches and/or machine learning tools used by researchers/developers to convert handwritten information into digital forms using Artificial Intelligence.

Tags: report reu comunication

6 minute read

Status: final, Type: Report

Mikahla Reeves, su21-reu-366, Edit

Abstract

The first thing that comes to numerous minds when they hear *Handwriting Recognition* is simply computers identifying handwriting, and that is correct. Handwriting Recognition is the ability of a computer to interpret handwritten input received from different sources. In the artificial intelligence world, handwriting recognition has become a very established area. Over the years, there have been many developments and applications made in this field. In this new age, Handwriting Recognition technologies can be used for the conversion of handwritten and/or printed text to speech for the blind, language translation, and for any field that requires handwritten reports to be converted to digital forms instantly.

This study investigates two of the approaches taken by researchers/developers to convert handwritten information to digital forms using (AI) Artificial Intelligence. These two deep learning approaches are the (CNN) Convolutional Neural Network and (LSTM) Long Short Term Memory. The CNN takes advantage of the spatial correlation in data, while LSTM makes predictions based on sequences of data.

Contents

Keywords: handwriting recognition, optical character recognition, deep learning.

1. Introduction

Perhaps one of the most monumental things in this modern-day is how our devices can behave like brains. Our various devices can call mom, play our favorite song, and answer our questions by just a simple utterance of Siri or Alexa. These things are all possible because of what we call artificial intelligence. Artificial intelligence is a part of computer science that involves learning, problem-solving, and replication of human intelligence. When we hear of artificial intelligence, we often hear of machine learning as well. The reason for this is because machine learning also

involves the use of human intelligence. Machine learning is the process of a program or system getting more capable over time [1]. One example of machine learning at work is Netflix. Netflix is a streaming service that allows users to watch a variety of tv shows and movies, and it also falls under the category of a recommendation engine. Recommendation engines/applications like Netflix do not need to be explicitly programmed. However, their algorithms mine the data, identify patterns, and then the applications can make recommendations.

Now, what is handwriting recognition? Handwriting Recognition is a branch of (OCR) Optical Character Recognition. It is a technology that receives handwritten information from paper, images, and other items and interprets them into digital text in real-time [2]. Handwriting recognition is a well-established area in the field of image processing. Over the last few years, developers have created handwriting recognition technology to convert written postal codes, addresses, math questions, essays, and many more types of written information into digital forms, thus making life easier for businesses and individuals. However, the development of handwriting recognition technology has been quite challenging.

One of the main challenges of handwriting recognition is accuracy, or in other words, the variability in data. There is a wide variety of handwriting styles, both good and bad, thus making it harder for developers to provide enough samples of what a specific character/integer looks like [3]. In handwriting recognition, the computer has to translate the handwriting into a format that it understands, and this is where Optical Character Recognition becomes useful. In OCR, the computer focuses on a character, compares it to characters in its database, then identifies what the letters are and fundamentally what the words are. Also, this is why deep learning algorithms like Convolutional Neural Networks and Long Short Term Memory exist. This study will highlight the impact each algorithm has on the development of handwriting recognition.

2. Convolutional Neural Network Model

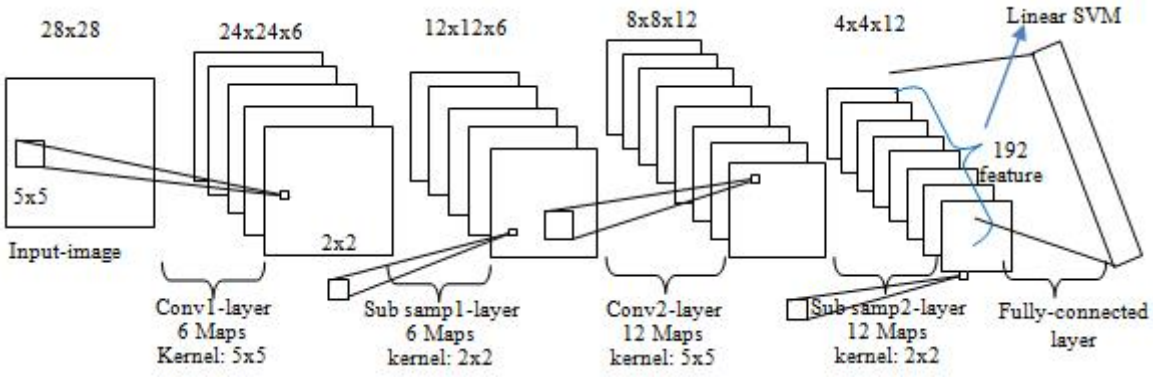


Figure 1: Architecture of CNN for feature extraction [4]

In this model, the *input image* passes through two convolutional layers, two sub-sample layers, and a linear SVM (Support Vector Machine) that allows for the output which is a *class prediction*. This class prediction leads to the editable text file.

Class prediction is a supervised learning method where the algorithm learns from samples with known class membership (training set) and establishes a prediction rule to classify new samples (test set). This method can be used, for instance, to predict cancer types using genomic expression profiling [5].

3. Long Short Term Memory Model

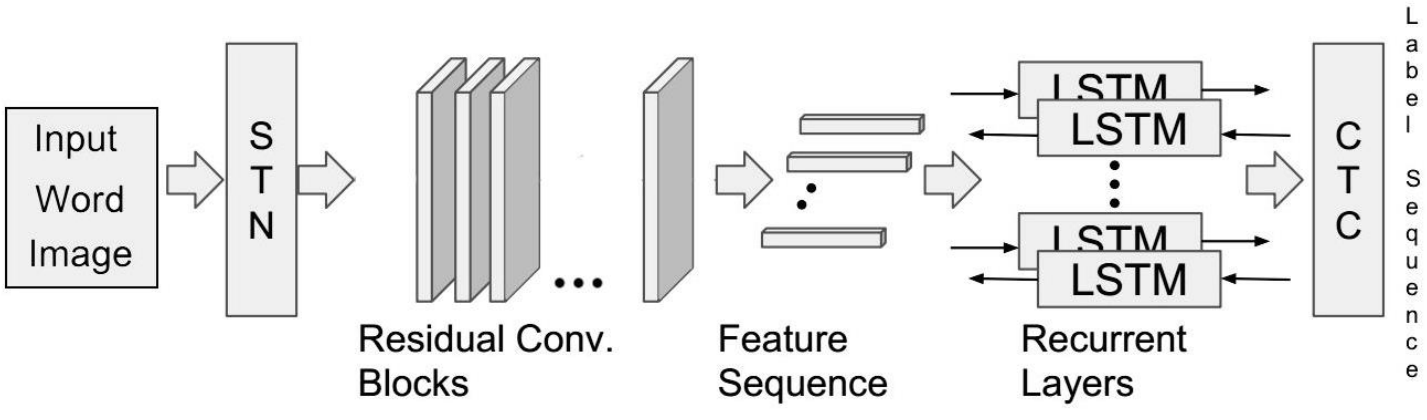


Figure 2: Overview of the CNN-RNN hybrid network architecture [6]

This model has a spacial transformer network, residual convolutional blocks, bidirectional LSTMs and the CTC loss (Connectionist Temporal Classification loss) which are all the processes the *input worded image* has to pass through before the output which is a *label sequence*.

Sequence labeling is a typical NLP task which assigns a class or label to each token in a given input sequence [7].

4. Handwriting Recognition using CNN

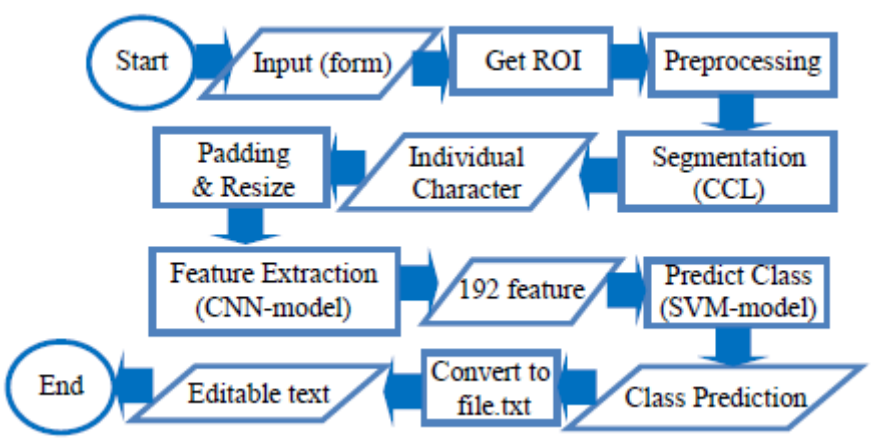


Figure 3: Flowchart of handwriting character recognition on form document using CNN [4]

In the study, former developers created a system to recognize the handwriting characters on form document automatically and convert it into editable text. The system consists of four stages: get ROI (Region of Interest), pre-processing, segmentation and classification. In the getting ROI stage, according to the specified coordinates, the ROI is cropped. Next, each ROI goes through pre-processing. The pre-processing consists of bounding box removal using the eccentricity criteria, median filter, and bare open. The output image of the pre-processing stage will be segmented using the Connected Component Labeling (CCL) method. It aims to get an individual character[4].

5. Conclusion

In this study, we learned how to use synthetic data, domain-specific image normalization, and augmentation - to train an LSTM architecture[6]. Additionally, we learned how a CNN is a powerful feature extraction method when applied to extract the feature of the handwritten characters and linear SVM using L1 loss function and L2 regularization used as end classifier[4].

For future research, we can focus on improving the CNN model to be able to better process information from images to create digital text.

6. Acknowledgments

This paper would not have been possible without the exceptional support of Gregor von Laszewski, Carlos Theran, Yohn Jairo. Their constant guidance, enthusiasm, knowledge and encouragement have been a huge motivation to keep going and to complete this work. Thank you to Jacques Fleicher, for always making himself available to answer questions. Finally, thank you to Byron Greene and the Florida A&M University for providing this great opportunity for undergraduate students to do research.

7. References

1. Brown, S., 2021. Machine learning, explained | MIT Sloan. [online] MIT Sloan. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. ↩

2. Handwriting Recognition in 2021: In-depth Guide. (n.d.). https://research.aimultiple.com/handwriting-recognition ↩

3. ThinkAutomation. 2021. Why is handwriting recognition so difficult for AI? - ThinkAutomation. [online] Available at: https://www.thinkautomation.com/bots-and-ai/why-is-handwriting-recognition-so-difficult-for-ai/. ↩

4. Darmatasia, and Mohamad Ivan Fanany. 2017. “Handwriting Recognition on Form Document Using Convolutional Neural Network and Support Vector Machines (CNN-SVM).” In 2017 5th International Conference on Information and Communication Technology (ICoIC7), 1–6. ↩

5. “Class Prediction (Predict Parameter Value).” NEBC: NERC Environmental Bioinformatics Centre. Silicon Genetics, 2002. http://nebc.nerc.ac.uk/courses/GeneSpring/GS_Mar2006/Class%20Prediction.pdf ↩

6. K. Dutta, P. Krishnan, M. Mathew and C. V. Jawahar, “Improving CNN-RNN Hybrid Networks for Handwriting Recognition,” 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 80-85, doi: 10.1109/ICFHR-2018.2018.00023. ↩

7. Jacob. “Deep Text Representation for Sequence Labeling.” Medium. Mosaix, August 15, 2019. https://medium.com/mosaix/deep-text-representation-for-sequence-labeling-2f2e605ed9d ↩

Increasing Cervical Cancer Risk Analysis

Cervical Cancer is an increasing matter that is affecting various women across the nation, in this project we will be analyzing risk factors that are producing higher chances of this cancer. In order to analyize these risk factors a machine learning technique is implemented to help us understand the leading factors of cervical cancer.

Tags: project reu health

5 minute read

Check Report passing Status failing Status: draft, Type: Project

Theresa Jean-Baptistee, su21-reu-369, Edit

Abstract

Cervical Cancer is an increasing matter that is affecting various women across the nation, in this project we will be analyzing risk factors that are producing higher chances of this cancer. In order to analyize these risk factors a machine learning technique is implemented to help us understand the leading factors of cervical cancer.

Contents

Keywords: Cervical, Cancer, Diseases, Data, conditions

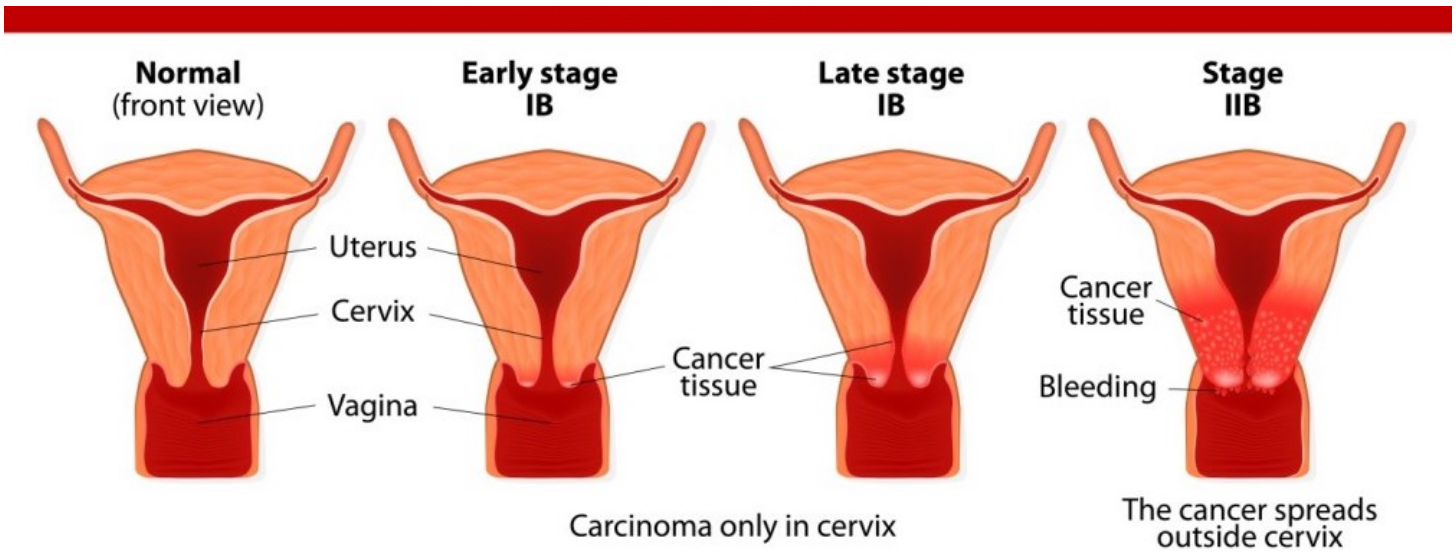
1. Introduction

Cervical cancer is a disease that is increasing in various women nationwide. It occurs within the cells of the cervix (can be seen in stage 1 of the image below). This cancer is the fourth leading cancer, where there are about 52,800 cases found each year, predominantly being in lower developed countries. Cervical cancer occurs most commonly in women who are within their 50's and who has symptoms such as watery and bloody discharge, bleeding, and painful intercourse. Two other common causes can be an early start on sexual activity and multiple partners. The most common way to determine if one may be affected by this disease is through a pap smear. When witnessed early it, can allow a better chance of results and treatment.

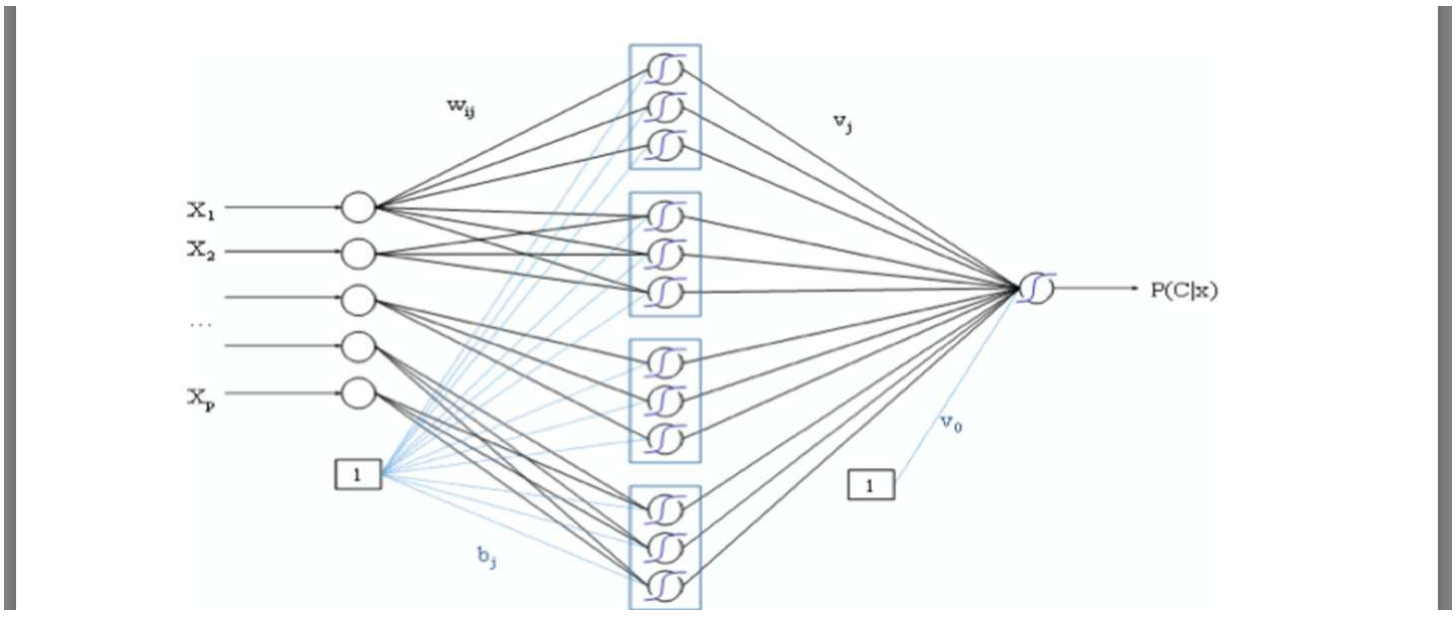
Cervical cancer is so important for the future of reproduction, being the cause of a successful or unsuccessful birth with completions like premature a child. The cervix help keeps the fetus stable within the uterus during this cycle, towards the end of development, it softens and dilates for the birth of a child. If diagnosed with this cancer, a miracle would be needed to conceive a child after

having treatment. Most treatments begin with a biopsy removing affected areas of cervical tissue. As it continues, to spread radiotherapy might be recommended to treat the cancer where may affect the womb. lastly, one may need to have a hysterectomy which is the removal of the womb.

In this paper, we will study the exact cause and risk factors that may place someone in this position. If spotted early it wouldn't affect someone's dream chance of conceiving or affect their reproductive parts. Using various data sets we will study the way everything may alignes in causes and machine leaning would be the primary technique to used interpretate the relation between variables and risk factor on cervical cancer.



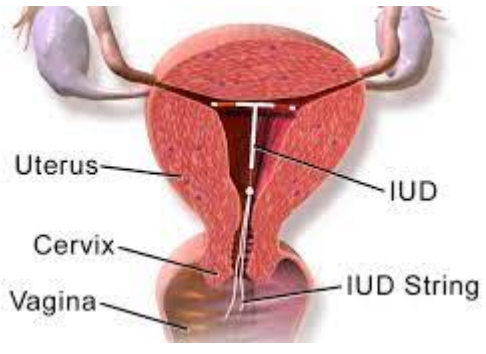
Model



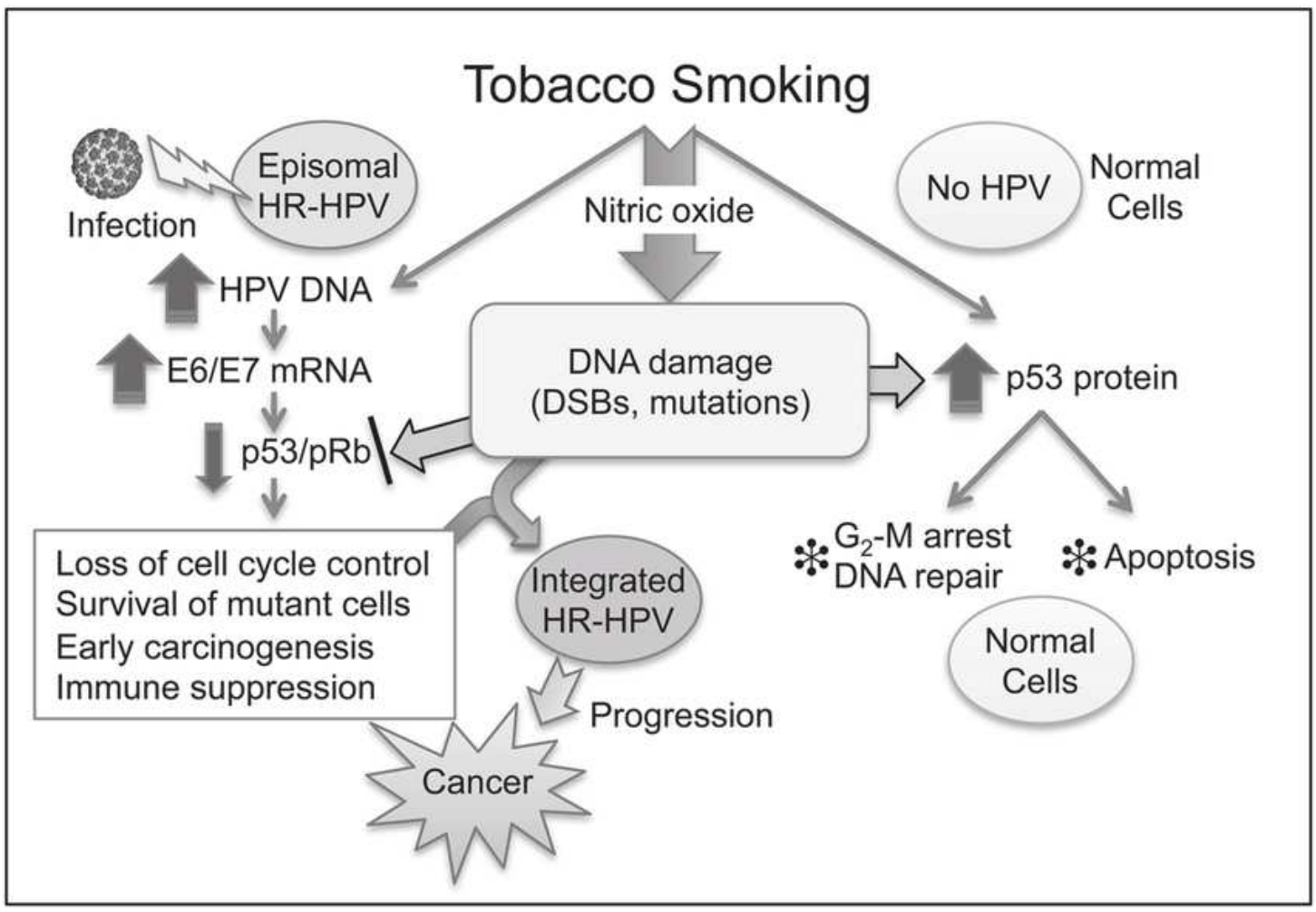
2. DataSets

The Data sets obatained shows the primary risk factors that affect women ages 15 and above. The few factors that sticked out the most were age, start of sexual activity, tabacoo intake, and IUD. The age and start of sexual activity maybe primary factor because a person is more liable to catch an STD and get this diease from mutiple parnters never really knowing what the other person may be doing outside of the encounterment. Tabcoo intake causes an affect making a person by weaking the immune system and making somone more septable to the disease. The IUD has the highest number on the data set being a primary factor that may put a person at risk, this device aids the prevention of pregency by thickneing the mucos of the cervix that could later cause infection or make your more spetiable to them.

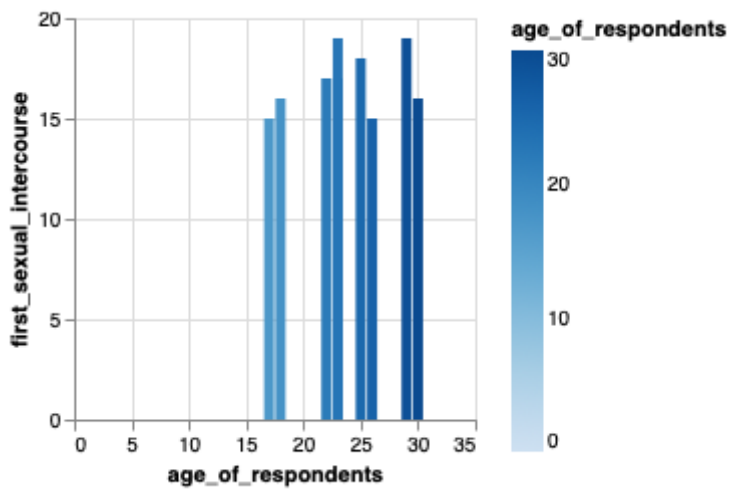
IUD Visulaization



Tabacoo Visulization Affect On Cervixs



Correlation of Age and Start Of sexual activity



3. Other People Works

The research of others work has made a huge imapact to this project starting from data to important knowledge needed to conduct the project. With the various research sites, we were able to witness what the affects various day to day activtie affect women long term. The Cervical Cancer Diagnosis Using a Chicken Swarm Optimization Based Machine Learning Method, was a big aid throught the project explaing the stages of cervical cancer, ways it can be treated, and the affects it may cause. With the data that was used from UCI Machine Learning, we were able to find efficent correlation into the data, helping the implented machine learning algorithm for the classification task.

4. Explantion of Confusion Matrix

The confusion matrix generated by multilayer perceptron can be explained as the perdicted summary results from the data obtained. Zero is when no cervical cancer is witnessed, one is when cervical cancer is seen. A hundrend and sixty-two is the highest number of this disease seen on the chart and the lowest number being winessed is two and eight being quiet of a jump.

5. Benchmark

```
+------------------+-------------------------------------------------------------------------------+
| Attribute        | Value                                                                         |
|------------------+-------------------------------------------------------------------------------|
| cpu              |                                                                               |
| cpu_cores        | 2                                                                             |
| cpu_count        | 4                                                                             |
| cpu_threads      | 4                                                                             |
| frequency        | scpufreq(current=2200.0, min=0.0, max=2201.0)                                 |
| mem.available    | 7.5 GiB                                                                       |
| mem.free         | 7.5 GiB                                                                       |
| mem.percent      | 52.9 %                                                                        |
| mem.total        | 15.9 GiB                                                                      |
| mem.used         | 8.4 GiB                                                                       |
| platform.version | ('10', '10.0.19041', 'SP0', 'Multiprocessor Free')                            |
| python           | 3.9.2 (tags/v3.9.2:1a79785, Feb 19 2021, 13:44:55) [MSC v.1928 64 bit (AMD64)] |
| python.pip       | 21.0.1                                                                        |
| python.version   | 3.9.2                                                                         |
| sys.platform     | win32                                                                         |
| uname.machine    | AMD64                                                                         |
| uname.node       | DESKTOP-B192DPK                                                               |
| uname.processor  | Intel64 Family 6 Model 61 Stepping 4, GenuineIntel                            |
| uname.release    | 10                                                                            |
| uname.system     | Windows                                                                       |
| uname.version    | 10.0.19041                                                                    |
| user             | carlo                                                                         |
+------------------+-------------------------------------------------------------------------------+
------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-61-3d18975177ee> in <module>
----> 1 StopWatch.benchmark()

~\AppData\Local\Programs\Python\Python39\lib\site-packages\cloudmesh\common\StopWatch.py in benchmark(cls, sysinfo, csv, prefi
x, tag, sum, node, user)
    253                     'start': time.strftime("%Y-%m-%d %H:%M:%S",
    254                                            time.gmtime(
--> 255                                                StopWatch.timer_start[timer])),
    256                     'time': StopWatch.get(timer, digits=3),
    257                     'sum': StopWatch.sum(timer, digits=3),

KeyError: 'Overall time'
```

6. Conclusion

In conclusion it can be found as women partake in their first sexual activity and continue they are more at risk. 162 is a dramatic number not necessarily being affected by age and 0 is only seen when a person does not partake in it. In the future I hope to keep furthering my Knowledge on Cervical Cancer, hopefully coming up with a realistic method to cure this disease where one can continue to live their life with as a human being.

7. Acknowledgments

The author would like to thank Yohn, Carlos, Gregor, Victor, and Jacques for all of their Help. Thank you!

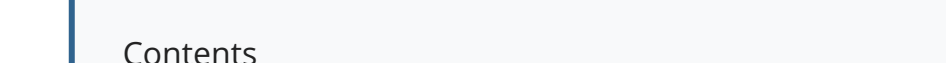
8. References

Project: Aquatic Animals Classification Using AI

Here comes the abstract

Tags: project reu ai biology

4 minute read



Status: draft, Type: Project

Timia Williams, su21-reu-370, Edit

Abstract

Marine animals play an important role in the ecosystem. “Aquatic animals play an important role in nutrient cycles because they store a large proportion of ecosystem nutrients in their tissues, transport nutrients farther than other aquatic animals and excrete nutrients in dissolved forms that are readily available to primary producers” (Vanni MJ 1) Fish images are captured by scuba divers, tourist, or underwater submarines. different angles of fishes image can be very difficult to get because of the constant movement of the fish. In addition to getting the right angles, the images of marine animals are usually low-quality because of the water. Underwater cameras that is required for a good quality image can be expensive. Using AI could potentially increase the marine population by the help of classification by testing the usage of machine learning using the images obtained from the aquarium combined with advanced technology. We collect 164 fish images data from Georgia acquarium to look at the different movements.

Contents

Keywords: tensorflow, example.

1. Introduction

It can be challenging to obtain a large number of different complex species in a single aquatic environment. Traditionally, it would take marine biologists years to collect the data and successfully classify the type of species obtained [1]. Scientist says that more than 90 percent of the ocean’s species are still undiscovered, with some estimating that there are anywhere between a few hundred thousand and a few million more to be discovered" (National Geographic Society). Currently, scientists know of around 226,000 ocean species. Now and days, Artificial intelligence and machine learning has been used for detection and classification in images. In this project, We will propose to use machine learning techniques to analyze the images obtained from the Georgia Aquarium to identify legal and illegal fishing.

2. Machine learning in fish species.

Aquatic ecologists often count animals to keep up the population count of providing critical conservation and management. Since the creation of underwater cameras and other recording equipment, underwater devices have allowed scientists to safely and efficiently classify fishes images without the disadvantages of manually entering data, ultimately saving lots of time, labor, and money. The use of machine learning to automate image processing has its benefits but has rarely been adopted in aquatic studies. With using efforts to use deep learning methods, the classification of specific species could potentially increase. In fact, there is a study done in Australia's ocean waters that classification of fish through deep learning was more efficient that manual human classification. In the study to test the abundance of different species, "The computer's performance in determining abundance was 7.1% better than human marine experts and 13.4% better than citizen scientists in single image test datasets, and 1.5 and 7.8% higher in video datasets, respectively" (Campbell, M. D.). This remarkably explain that using machiene learning in marine animals is a better method than a manually classifying Aquatic animals Not only is it good for classification, it will be used to answer broader questions such as population count, the location of species, its abundance, and how it appears to be thriving. Since Machine learning and deep learning are often defined as one, both learning methods will be used to analyze the images and find patterns on my data.

3. Datasets

We used two datasets in my project. The first dataset includes the pictures that I took at the Georgia Acquarium. That dataset was used for testing. The second dataset used was a fish dataset from kaggle which contains 9 different seafood types (Black Sea Sprat, Gilt-Head Bream, Hourse Mackerel, Red Mullet, Red Sea Bream, Sea Bass, Shrimp, Striped Red Mullet, Trout). For each type, there are 1000 augmented images and their pair-wise augmented ground truths.

The link to access the dataset I used from kaggle is https://www.kaggle.com/crowww/a-large-scale-fish-dataset

3.1. Sample of Images of Personal Dataset

Left to right: Banded Archerfish, Lionfish, and Red Piranha

Figure 1: These images are samples of my personal data which is made up of images of fishes taken at the Georgia Acquarium.

3.2. Sample of Images from Large Scale Fish Dataset

4. Conclusion

Deep learning methods provide a faster, cheaper, and more accurate alternative to manual data analysis methods currently used to monitor and assess animal abundance and have much to offer the field of aquatic ecology. We was able to create a model to prove that we can use AI to efficiently detect and classify marine animals.

5. Acknowledgments

Special thanks to these people that helped me with this paper: Gregor von Laszewski Yohn Jairo Carlos Theran Jacques Fleischer Victor Adankai

6. References

Project: Detecting Multiple Sclerosis Symptoms using AI

This work implements machine learning algorithim apply in Multiple Sclerosis symptoms and provides treatment options available

Tags: project reu health ai

7 minute read

Check Report failing Status failing Status: draft, Type: Project

Raeven Hatcher, su21-reu-371, Edit

Abstract

Multiple sclerosis (M.S.) is a chronic central nervous system disease that potentially affects the brain, spinal cord, and optic nerves in the eyes. People that suffer from M.S had their immune system attacks the myelin (protective sheath) that covers nerve fibers, resulting in communication problems between the brain and the body. The cause of M.S. is unknown; however, researchers believe that genetic and environmental factors play a role in those affected. Symptoms differ significantly from person to person due to varying nerves involved. The most common symptoms include tremors, numbness or weakness in limbs, vision loss, blurry vision, double vision, slurred speech, fatigue, dizziness, involuntary movement, and muscle paralysis. There is currently no cure for Multiple sclerosis and treatment focuses on slowing the progression of the disease and managing symptoms.

There is no proven way to predict how an individual with M.S will progress certainly. However, researchers established four phenotypes that will assist in identifying those who are more inclined to have disease progression and help aid in more effective treatment targeting. In this experiment, Artificial Intelligence (AI) will be applied by ascertaining what causes these different phenotypes and which phenotype is at most risk for disease progression using a Magnetic Resonance Scan.

Contents

Keywords: tensorflow, example.

1. Introduction

MS or Multiple sclerosis is a potentially disabling autoimmune disease that can damage the brain, spinal cord, and optic nerves located in the eyes. It is the most common progressive neurological disability that affects adolescents. The immune system attacks the central nervous system in this disease, specifically myelin (sheth that covers and protects nerve fibers),

oligodendrocytes (myelin-producing cells), and the nerve fibers located under myelin. Myelin enables nerves to send and receive electrical signals swiftly and effectively. The myelin sheath becomes scarred from being attacked. These attacks make the myelin sheath inflamed in little patches, observable on an MRI scan. These little inflamed patches potentially disrupt messages moving along the nerves, which ultimately lead to the symptoms of Multiple sclerosis. If these attacks happen frequently, permanent damage can occur to the involved nerve. Because Multiple sclerosis affects the central nervous system, which control all of the actions carried out in the body, symptoms can affect any part of the body and vary. The most common symptoms of this progressive disease include muscle weakness, pins and needle sensation, electrical shock sensation, loss of bladder control, muscle spasms, tremors, double or blurred vision, partial or total vision loss, to name a few. Researchers are not sure what causes Multiple sclerosis but believe those between the ages of 20 and 40, women, smoke, are exposed to certain infections, have a vitamin D and B12 deficiency, and related to someone affected by this disease are more susceptible.

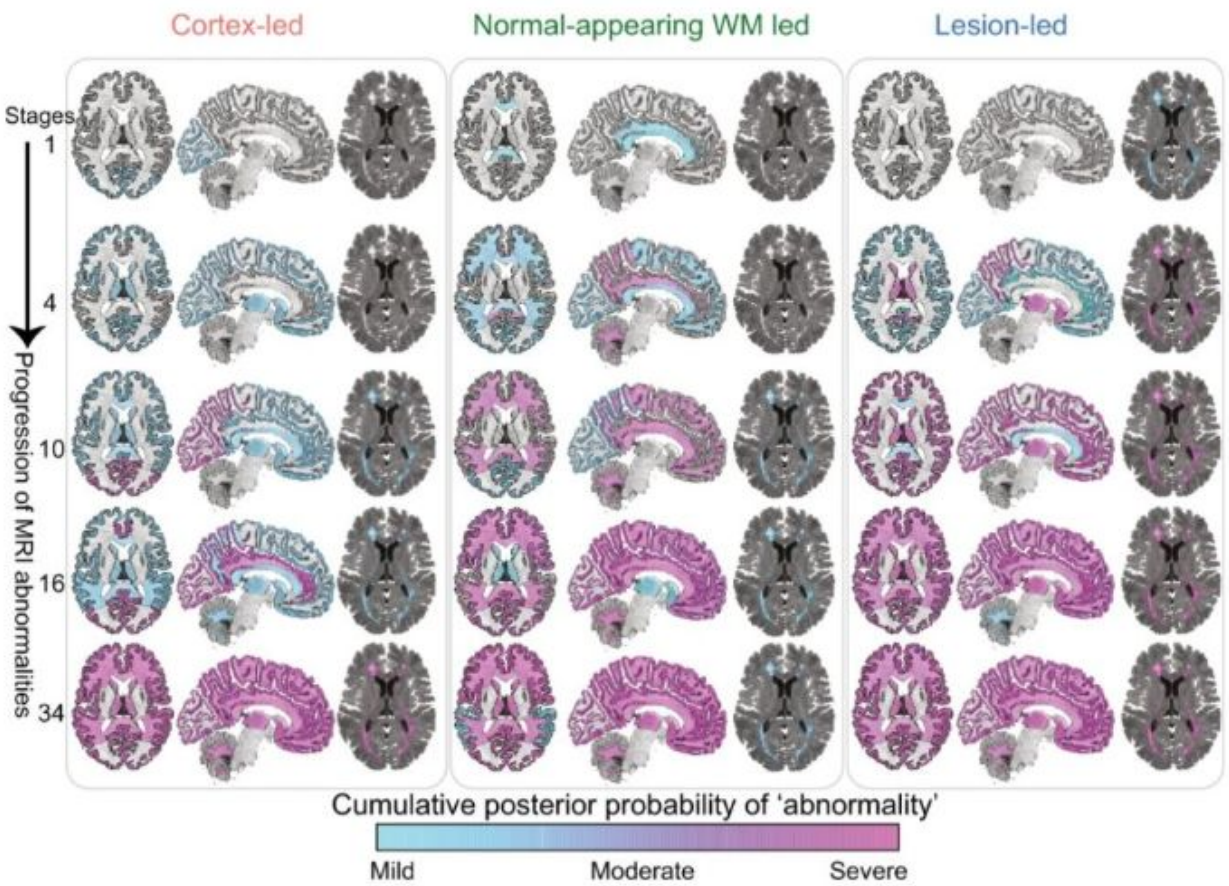
It can be difficult to diagnose MS due to the symptoms usually being vague or very similar to other conditions. There is no single test to diagnose it positively. However, doctors can choose a neurological examination, MRI scan, evoked potential test, lumbar puncture, or a blood test to diagnose a patient properly. Currently, clinical practices divide MS into four phenotypes: clinically isolated syndrome (CIS), relapsing-remitting MS (RRMS), primary-progressive MS (PPMS), and secondary progressive MS (SPMS). Two factors define these phenotypes; disease activity (evidenced by relapses or new activity on MRI scan) and progression of disability. Phenotypes are routinely used in clinical trials to choose patients and conduct treatment plans.

New technologies, such as artificial intelligence and machine learning, help assess multidimensional data to recognize groups with similar features. When implemented in apparent abnormalities on MRI scans, these new technologies have assured promising results in classifying patients who share similar pathobiological mechanisms rather than the typical clinical features.

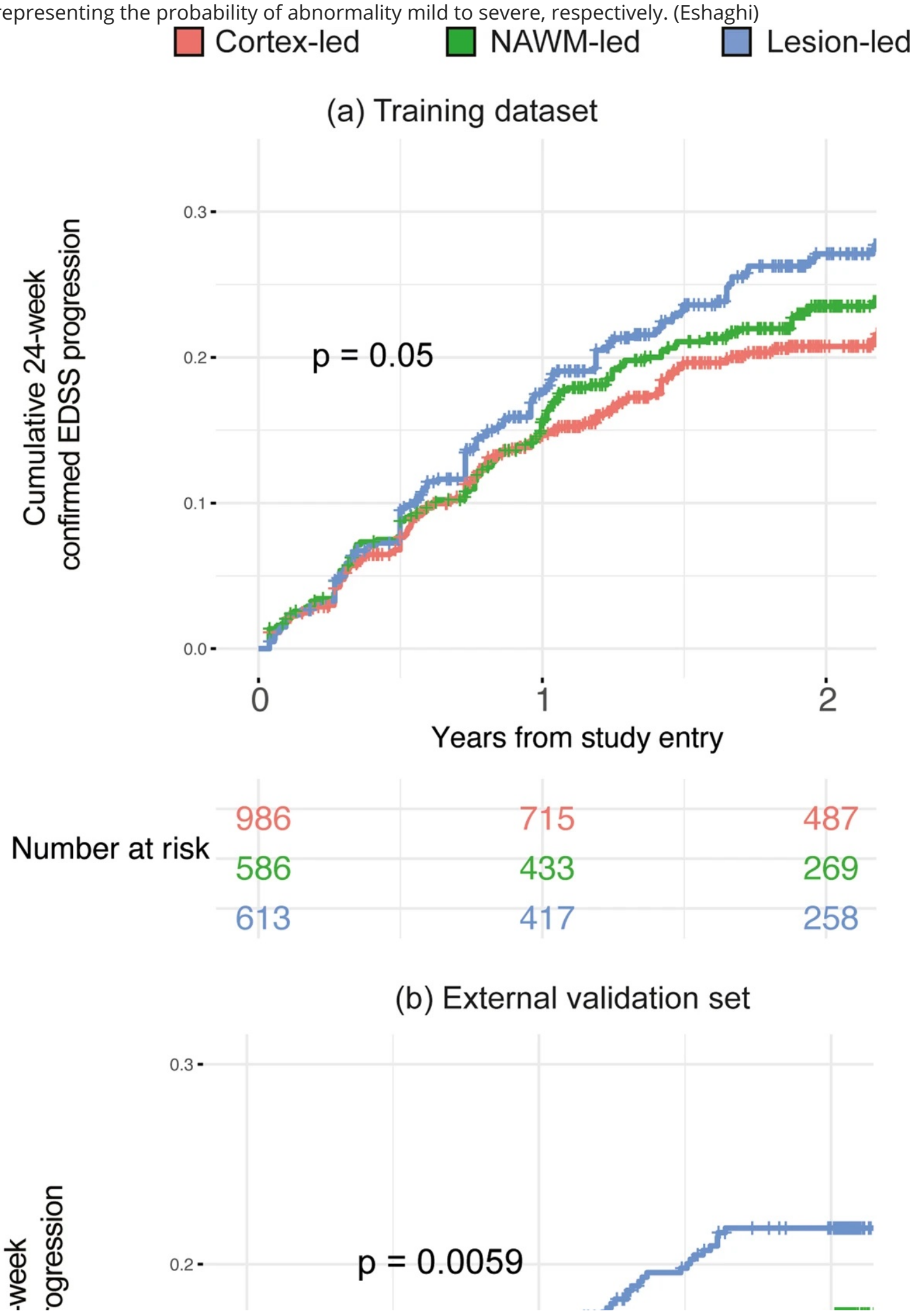
Researchers at UCL work with the Artificial intelligence (AI) tool SuStain (Subtype and Stage Inference) to ask whether AI can find Multiple sclerosis subtypes that follow a particular pattern on brain images? The results uncovered three data-driven MS subtypes defined by pathological abnormalities seen on brain images (Skylar). The three data-driven MS subtypes are cortex-led, normal-appearing WM-led, Lesion-led. Cortex-led MS is characterized by early tissue shrinkage (atrophy) in the outer layer of the brain. Normal-appearing WM-led is identified by irregular diffused tissue located in the middle of the brain. Lastly, a lesion-led subtype is characterized by early extension accumulation of brain damage areas that lead to severe atrophy in numerous brain regions. All three of these subtypes correlate to the earliest abnormalities observed on an MRI scan within each pattern.

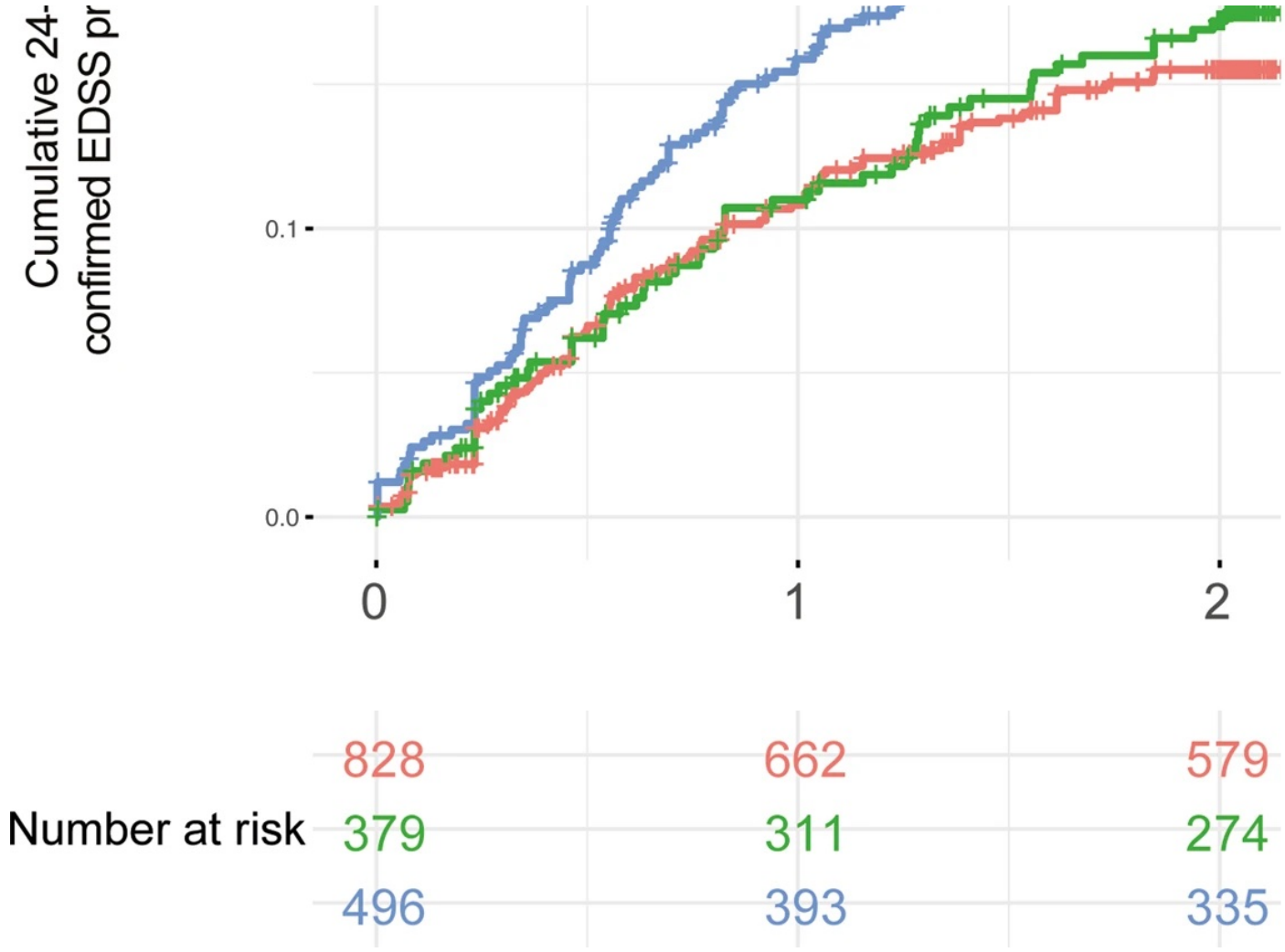
In this experiment, researchers utilized the SuStain tool to capture MRI scans of 6,332 patients. The unsupervised SuStain taught itself and identified those three patterns that were previously undiscovered.

3. Using Images



The above image shows the MRI-based subtypes. The color shades range from blue to pink, representing the probability of abnormality mild to severe, respectively. (Eshaghi)





4. Datasets

MRI brain scans of 6,322 MS patients. look if you can find figures descrbing the data.

5. Benchmark

Your project must include a benchmark. The easiest is to use cloudmesh-common.

6. Conclusion

A vital barrier in distinguishing subtypes in Multiple sclerosis is to stitch observations together from cross-sectional or longitudinal studies. Grouping individuals based wholly on their MRI scan is ineffective because patients belonging to the same subgroup could show ranging abnormalities as the disease progresses and would appear different. SuStaIn, Subtype and Staging Inference, a newly developed unsupervised machine learning algorithm aids in uncovering data-driven disease subtypes that have distinct temporal progression patterns. “The ability to disentangle temporal and phenotypic heterogeneity makes SuStain different from other unsupervised learning or clustering algorithms” (Eshaghi). SuStaIn identifies subtypes given the data, defined by a particular pattern of variation in a set of features, such as MRI abnormalities. Once the SuStain subtypes and their MRI trajectories are adequately identified, the disease model can conclude how approximately a patient, whose MRI is unseen, belongs to each of the three subtypes and stages.

A total of 9,390 patients participated in this research study. Six thousand three hundred twenty-two patients were utilized in training, and 3,068 patients were used for the validation dataset. Patient characteristics such as sex, age, disease duration, and expanded disability status scale (EDSS) were similar between the training and validation dataset. There were 18 MRI features measured, 13 of those differed dramatically from those between the MS training dataset and control group and were maintained in the SustaIn model. Three subtypes, with very distinct patterns, were identified in the training dataset and validated in the validation dataset. The early abnormalities noticed by SuStain helped define the three subtypes: cortex-led, normal-appearing white matter-led, and lesion-led.

There was a statistically significant difference in the rate of the disease progression between the subtypes in the training dataset and validation datasets. The lesion-led subtype held a 30% higher risk of developing 24-week confirmed disability progression (CDP) than the cortex-led subtype in the training dataset. The lesion-led validation dataset had a 32% higher risk of

confirmed disability progression than the cortex-led subtype. No other differences in the advancement of disability between subtypes were noted. When SuStaIn was applied to the training and validation dataset, it was pointed out that there were differences in the risk of disability progression between SuStaIn stages.

Each MRI-based subtype had a different response to treatment, comparing those on treatment and those on placebo. The lesion-led subtype showed a remarkable response to the treatment. Patients on the lesion-led active treatment subtype showed a significantly slower worsening of EDSS than those on the placebo. No differences in the rate of EDSS were observed in those on the placebo compared to active treatment in the NAWM-led and cortex-led subtypes.

When SuStain was applied to a large set of Multiple sclerosis scans, it identified three subtypes. Researchers found out the patient's baseline subtype and stage were associated with an increased risk of disease progression. Combining clinical information with the MRI-based three subtypes increased the predictive accuracy of just using the MRI scan information alone. The patterns of MRI abnormality in these subtypes provide perspicacity into disease mechanisms, and, alongside clinical phenotypes, they may aid the stratification of patients for future studies.

7. Acknowledgments

The author likes to thank Gregor von Laszewski, Yohn Jairo, and Carlos Theran.

8. References

[^6] What Is MS? National Multiple Sclerosis Society. (n.d.). https://www.nationalmssociety.org/What-is-MS.

Project: Analyzing Hashimoto disease causes, symptoms and cases improvements using Topic Modeling

Analyzing factors as immune systems, genetics and diets than can lead to Hashimoto disease

Tags: project reu ai health

18 minute read

`Status: final`, Type: Project

Sheimy Paz, su21-reu-372, Edit

- Code:
  - Install documentation requirements.txt
  - Tyroidhitis_Project.ipynb

Abstract

This project proposes a new view of Hashimoto's disorder, its association with other pathologies, possible causes, symptoms, diets, and recommendations. The intention is to explore the association of Hashimoto disorder with disease like h pylori bacteria, inappropriate diet, environmental factors, and genetic factors. To achieve this, we are going to utilize AI in particular topic modeling which is a technic used to process large collection of data to identifying topics. Topic modeling is a text-mining tool that help to correlate words with topics making the research process easy and organized with the purpose to get a better understanding of the disorder and the relationship that this has with other health issues hoping to find clear information about the causes and effect that can have on the human body. The dataset was collected from silo breaker software, which contains information about news, reports, tweets, and blogs. The program will organize our findings highlighting key words related to symptoms, causes, cures, anything that can apport clarification to the disorder.

Contents

Keywords: Thyroid disease, Hashimoto, H Pylori, Implants, Food Sensitivity, Diary sensitivity, Healthy Diets, Exercise, topic modeling, text mining, BERT model.

1. Introduction

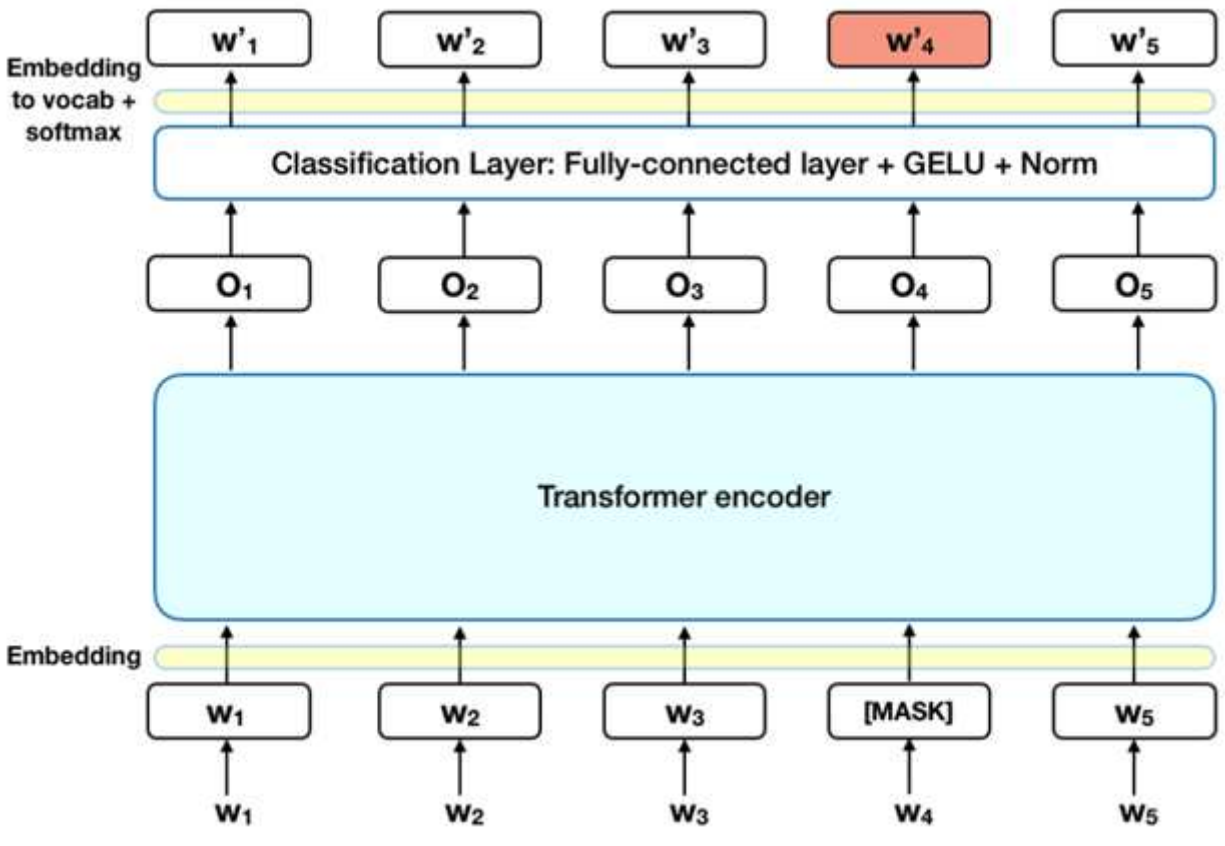
Hashimoto thyroiditis is an organ-specific autoimmune disorder. its symptoms were first described 1912 but the disease was not recognized until 1957. Hashimoto is an autoimmune disorder that destroys thyroid cells and is antibody-mediated [1]. In a female-to-men radio at least 10:4 women are more often affected than men. The diagnostic is often called between the ages of 30 to 50 years [2]. Pathologically speaking, Hashimoto stimulates the formation of antithyroid antibodies that attack the thyroid tissue, causing progressive fibrosis. Hashimoto is believe to be the concequence of a combination of mutated genes and eviromental factors[3]. The disorder is difficult to diagnose since in the early course of the disease the patients may or may not exhibit symptoms or laboratory findings of hyperthyroidism, it may show normal values because the destruction of the gland cells may be intermittent [1]. Clinical and epidemiological studies suggest worldwide that the most common cause of hypothyroidism is an inadequate dietary intake of iodine.

Due to the arduous labor to identify this disorder a Machine Learning algorithm based on prediction would help to identify Hashimoto in early stages as well as any other health issues related to it [4]. This will be helpful for patients that would be able to get the correct treatment in an early stage of the illness avoiding future complications. This research algorithm was mainly intended to find patient testimonies of improvements, completed healed cases, early symptoms, trigger factors or any useful information about the disorder.

Hashimoto autoimmune diseases have been linked to the infection caused by H pylori bacteria. H pylori is until the date the most common chronic bacterial infection, affecting half of the world's population and is known for the presence of Caga antigens which are virulent strains that have been found in organ and non-organ specific autoimmune diseases [23]. Another important trigger of Hashimoto disorder is the inadequate modern diet patterns and the environmental factors that are closely related to it [4]. For instance, western diet consumption is an essential factor that trigs the disorder since this food is highly preserved and predominate the consumption of artificial flavors and sugars which have dramatically increase in the past years, adding to it the use of chemicals and insecticides in the fruits and vegetables and the massive introduction of hormones for meat production, all this can be the cause of the rise of autoimmune diseases [5].

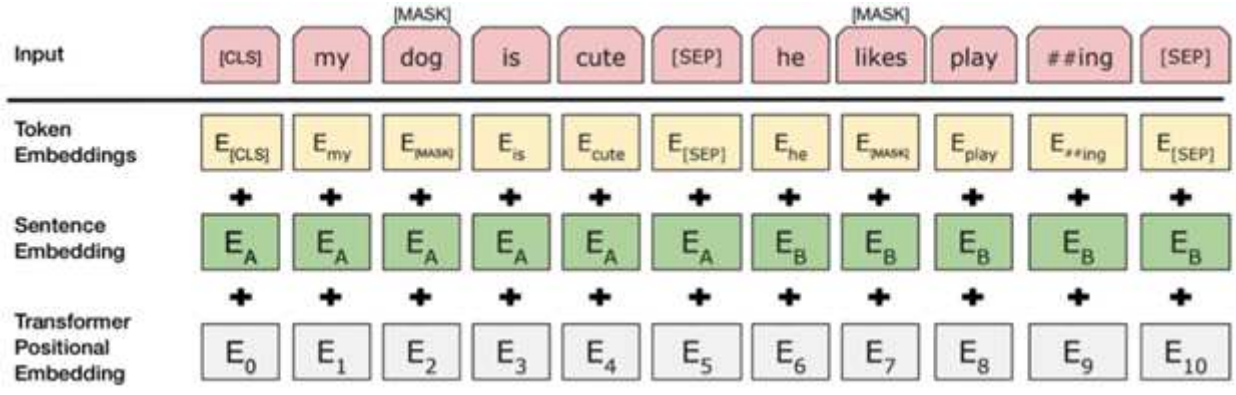
We utilize deep learning BERT model to train our dataset. BERT is a superior performer Bidirectional Encoder, which superimposes 12 or 24 layers of multiheaded attention in a Transformer [1]. Bert stands for Bidirectional(read from left to right and vice versa with the purpose of an accurate understanding of the meaning of each word in a sentence or document) Encoder Representations from Transformers(the used of transformers and bidirectional models allows the learning of contextual relations between words). Notice that BERT uses two training strategies MLM and NSP.

Masked Lenguage Model (MLM) process is made by masking around 15% of token making the model predict the meaning or value of each of the masked words. In technical word it requires 3 steps Adding a classification layer on top of the encoder output, Multiplying the output vectors by the embedding matrix, transforming them into the vocabulary dimension. And lastly calculating the probability of each word in the vocabulary with SoftMax. Here we can see an image of the process [6].



MLM Bert Figure: "Masked Language Model Figure Example" [6].

Next Sentence Prediction (NSP) process is based in sentence prediction. The model obtains pair of sentences as inputs, and it is train to predict which is the second sentence in the pair. In The training process 50% of the input sentences are in fact first and second sentence and in the other 50% the second sentences are random sentences used for training purposes. The model is able to distinguish if the second sentence is connected to the first sentence by a 3-step process. An CLS (the reserved token to represent the start of sequence) is inserted at the beginning of the first sentence while the SEP (separate segments or sentence) is inserted at the end of each sentence. And embedding indicating sentence A or B is added to each token, and lastly a positional embedding is added to each token to indicate its position in the sequence like is shown on the image [6].



NSP Bert Figure: "Next Sentence Prediction Figure Example" [6].

By the trained model Parameter learning we obtains the word embeddings of the input sentence or input sentence pair in the unsupervised learning framework proceeds by solving the following two tasks: Masked Language Model and Next Sentence Prediction.

We try to use Bert model in the small dataset Hashimoto without any success because the BERT model was overfitting the data points. We use LDA model to train the Hashimoto dataset which allow us to find topic probabilities that we compare with the thyroiditis dataset that was trained with the BERT model-framework.

We used Natural Language Toolkit (NLTK) which is a module that uses the process of splitting sentences from paragraph, split words, recognizing the meaning of those words, to highlighting the main subjects, with the purpose to help to understand the meaning of the document [7]. For instance, in our NLTK model we used two data sets Hashimoto and thyroiditis and we were able to identify the top 30 topics connected to these disorders. From the information collected we were able to identify general information like association of the disorder with other health issues. The impact of Hashimoto patient with covid19, long term consequences of untreated Hashimoto, recommendation for advance cases, and diet suggestion for improvement. The used of Natural Language tool kit made a precise and less time consuming research process.

2. Summary Tables

We can observe in this table the differences between this two similar disorders that are frequently misunderstood.

Summary Table 1: "Differences Between Hashimoto's Thyroiditis and Grave's Disease" [8].

Autoimmune disorder	Hashimoto's thyroiditis[2,4]		Graves' disease[2,5]	
Antibodies	• Antithyroid peroxidase antibodies • and/or Antithyroglobulin antibodies • (Normally no anti-TSH receptor antibodies)		• Anti-TSH receptor antibodies • (Antithyroid peroxidase antibodies (ir 80% of patients)) • (Antithyroglobulin antibodies (in 50%))	
Thyroid state	Progresses into hypothyroidism		Hyperthyroidism	
Serum TSH	Variable	Progresses into hypothyroidism	Low	Hyperthyroidism
Free T3	Variable		High	
Free T4	Variable		High	
Clinical features	• Hypothyroid, • Euthyroid, • or (more rarely) Hyperthyroid		• Hyperthyroid signs and symptoms with sometimes pretibial myxedema, often thyroid eye disease	
Thyroid ophthalmopathy	• Rarely exophthalmos (<2%), • More frequently hypothyroid signs, such as suborbital edema		• Exophthalmos (in 25%), • Other signs: lid lag, eyelid edema, chemosis, extraocular muscle weakness	
Thyroid gland	• Thyroid atrophy and (micro) nodules are more common, • Occasionally goiter		• Goiter	
Histological findings	• Mononuclear cells, mostly lymphocytes (especially T cells)		• Activated follicular cells • Increased vascularization	
	• Destroyed thyroid follicles		• Hyperplastic follicles with colloidal absorption	
Thyroid ultrasound	• Reduced echogenicity (progresses to atrophic hypovascular thyroid)		• Increased echogenicity (diffusely enlarged hypervascular thyroid gland	

Summary Table 2: "Hashimoto's thyroiditis is associated with other important disorders" [8].

Table 2: Hashimoto's Thyroiditis: Associated Pathologies

TYPE OF DISORDER	RISK, SEVERITY*
Fitness deficits	
• Reduced brain perfusion[51-53]	*2.1x ↑ perfusion defects*[53]
• Low quality of life, fatigue[35,54-66]	*+66% greater fatigue*[55]
• Reduced physical functioning[67]	*-11% reduction*
• Low cardiac function[68-69]	*-29% reduction*[69]
Psychological disorders	
• Hypothyroid symptoms[35]	*+50% increase*
• Reduced mental health[67]	*-11% reduction*
• Depression[67,70-85]	**1.5-9x higher risk**[79]
• Suicide (death by)[86]	**1.4x higher risk**
• Anxiety[67,75,79-81,87]	**4x higher risk**[80]
• Panic disorder[75,79]	**9x higher risk**[79]
• Obsessive compulsive disorder[60-75]	**1.5x higher risk**[61]
• Neuroticism[88-89]	**1.3x higher risk**[88]
• Paranoia[90]	in 40% in Hashimoto's encephalopathy
• Psychosis[91-92]	in 25%[91] in Hashimoto's encephalopathy
Mental -Neurological disorders[90]	
• Neuropathy (peripheral)[93]	in 11% of HT patients
• Orbitopathy[94-96]	in 2% of HT patients[96]
• Encephalopathy[97-99]	Rare
• Attention deficit[100-101]	**2.9x higher risk**[101]
• Cognitive impairment[102]	in 28% of HT Patients
• Alzheimer's disease[103]	Rare
• Other dementias[104-107]	Rare
• Multiple sclerosis[108]	9% of men with MS have HT
Sleep disorders	
• Sleep apnea[109-110]	47% of patients have HT[109]
Sexual /reproductive disorders	
• Sexual dysfunction[35,72]	**1.4x higher risk**[72]
• Infertility[111-112]	HT women have
	-41% lower pregnancy rate[112]
• Miscarriages[113]	**1.2-2.5x higher risk**
Endocrine disorders	
• Prolactinomas[114-116]	30% of patients have HT[114]
• Thyroid nodules, goiter[117-121]	in 36% of HT patients[120]
• Hypothyroidism[10-18]	**10x higher risk**[10]
• Polycystic ovarian syndrome[122-123]	**11x higher risk**[123]
• Premature ovarian failure[124-125]	20-30% of patients have HT[125]
• Addison's disease[126]	**Higher risk**
• Type I diabetes[127-146]	35% of patients have HT[127]
• Type II diabetes[147]	19% of patients have HT
• High estradiol-low testosterone (men)[148]	Men with HT have +10% higher E2/T ratio
Metabolic disorders	
Overweight, obesity[149-152]	*4x higher ATPO and 10x higher ATG levels*[152]
Digestive disorders	
• Celiac disease[153-160]	**4x higher risk**[160]

TYPE OF DISORDER	RISK, SEVERITY*
General disorders	
• Reduced general health[67]	*-15%*
• Oxidative stress, low antioxidant capacity[161-165]	*↑ ROS, ↓ antioxidant potential, +18% ↑ AGEs*
• Other autoimmune diseases	in 20-40% of HT patients[126,166-177]
Hair, skin, mucosa disorders	
• Alopecia areata, totalis[178-180]	5-25% of patients have HT[178,180]
• Ichthyosis[43,181]	**Higher risk**
• Atopic dermatitis[182-183]	10% of patients have HT[182]
• Chronic urticaria[184-185]	18% of patients have HT[184-185]
• Psoriasis[186]	25-30% of patients have HT[184]
• Vitiligo[187-194]	34% have HT (vs 9% in the general population)[190]
• Hirsutism[43]	**Higher risk**
• Sjögren syndrome, reduced salivary output[170-174]	*4x more have HT (than the general population)*[171]
Cardiovascular disorders	
• Lipid disorders[195-200]	*High total-LDL cholesterol, triglycerides, low HDL*
• Hyperhomocysteinemia[201]	*+22% ↑ homocysteine*
• Arterial stiffness[202-204]	*+10% ↑ pulse wave velocity*[204]
• Atherosclerosis[205-208]	*Thicker intima media*
• Thrombosis[209-210]	*Fibrinolytic deficit*
• Pulmonary hypertension[211-212]	**3x higher risk**[211]
• Mitral valve prolapse[213]	**3x higher risk**
• Coronary vasospasm[214]	**5x higher risk (w/ ↑ ATPO)**
• Coronary heart disease[215-224]	**1.4x higher risk**[219]
• Myocardial infarction[225-226]	**2x higher risk**[225]
• Stroke[227]	**1.1-1.3x higher risk**
• Hepatitis C virus-related cryoglobulinemia[228-231]	**3-6x ↑ risk for high cytokines CXCL 9,10 & 11****[228]
Bone, joint, and tendon disorders	
• Temporomandibular arthritis (TMA)[232]	**100% risk** to have TMA symptoms
• Myopathy (proximal)[93,233]	in 13% of HT patients[93]
• Polymyalgia rheumatica[169,234]	**Higher risk**
• Body pains[235-236]	**Higher risk**
• Fibromyalgia[235,237-243]	in 31% of HT patients[241]
• Rheumatoid arthritis[235,244-248]	**2.5x higher risk**[247]
• Systemic lupus Erythematosus[169,175-177]	2.3x more patients have HT than controls[175]
• Systemic sclerosis[249]	20% of patients have HT
• Spinal disc degeneration[250]	**1.8x higher risk**
Cancer	
• Thyroid (papillary) cancer[117-118,251-261]	**3x higher risk**,[258] but more favorable outcome[262-264]
Premature death by	
• Suicide[86]	**1.4x higher risk**
• Unknown matters[86]	**1.4x higher risk**
• Cardiovascular causes[265]	**1.7x higher risk**

Notes: Symbols: * Approximate risks, severity found in various studies and compared to the general population or control subjects without Hashimoto's thyroiditis; **heart failure markers

Abbreviations: HT = Hashimoto's thyroiditis; ATPO = antithyroperoxidase antibodies; ATG = antithyroglobulin antibodies; x = times or -fold; ROS: reactive oxygen species = free radicals; AGE's = Advanced Glycation End Products; ↑= increased = higher

Summary Table 3: "Overview of the main dietary recommendations for patients with Hashimoto" [8].

Table 3: Dietary Changes That Reduce Antithyroid Antibody Levels

DIETARY CHANGE	THE STUDIES; WHAT IT DOES	PRACTICAL TIPS
• Paleo diet[281-283]	This diet is easier to digest, as it consists of foods that our ancestors ate in the Paleolithic period before agriculture was discovered (the Neolithic) and humans started to consume foods less fit for their guts, such as cereal and milk products. A study showed that regular intake of fruit and berry juices increases the likelihood of autoimmune diabetes.	• Consume fresh vegetables, fruits, meat, fish, poultry, and eggs. • Consume organic foods. • Eat food raw, unprocessed[283] or steamed, boiled, or cooked at low temperature without oil. • Reducing fruit and fruit juice consumption to below-average levels might have value.[284]
• Protein-rich foods at breakfast	When protein-rich foods are eaten in the morning, there is time enough to digest them in the stomach and small intestine and absorb them as amino acids in the small intestine. At the end of the day, the gut can rest and recover better during sleep without undigested proteins remaining in the gut. A protein-rich breakfast also increases satiety.[285-286]	• Eat the main protein-rich meal (meat, poultry, fish, and eggs) in the morning. • Moderate protein intake at lunch. • Avoid consuming protein-rich foods in the evening.
• Small fatty fishes, rich in omega-3 polyunsaturated fatty acids	Regular intake of fatty (oily) fishes has been reported to reduce the incidence of postpartum thyroiditis by more than four times[287-288] and of autoimmune diabetes two-fold![289] When fish is eaten in the first part of the day, it also reduces food intake at supper, helping to not overload the gut during the night.[290]	• Increase the intake of small fatty fishes, such as sardines, mackerel, eels, and herring. • Limit the consumption of big fatty fishes, such as salmon, tuna, and trout, as they often contain too much mercury.
• Intermittent fasting • Supper: to skip or eat minimally	Intermittent fasting has been shown to reduce the production of autoimmune antibodies of a variety of diseases.[291-294] It provides a rest for the gut, allowing it to be temporarily free of new aggressors. It also permits the abdomen to get flat.	• In my experience, the best plan is to skip one meal daily, especially the evening meal, and eat proteins only in the morning. • Consuming boiled or steamed vegetables at supper is an acceptable alternative.
• Avoid soy milk	Soy milk intake is associated with a higher risk of autoimmune thyroiditis in children.[295]	
• Avoid cow-milk protein	Cow-milk protein is known to trigger autoimmune diabetes[296-299] and is suspected of triggering other autoimmune diseases.	• Avoid milk, yogurt, cheese, etc. • Clarified butter (also called ghee) is okay, as it has lost the white layer of allergenic proteins.
• Low-carbohydrate diet	• A low-carb diet alone has been reported to reduce thyroid antibody levels by 50%.[300]	Stop consumption of bread, porridge, and other cereals, particularly gluten-containing cereals, as well as high-sugar foods and drinks
	• Gluten-containing cereals can trigger autoimmune thyroiditis[301-303] and celiac disease. Celiac disease itself is often associated with autoimmune thyroiditis[153-160] and autoimmune diabetes.[304-306]	• No wheat. • Sprouted rice and other sprouted grains can be acceptable alternatives.
• Avoid artificial sweeteners	The consumption of artificial sweeteners is associated with a higher risk of autoimmune thyroiditis[307] and diabetes.[308]	Avoid aspartame, cyclamate, and other artificial sweeteners.
• Avoid sugar	The consumption of sugar and sweetened beverages is associated with a higher risk of autoimmune diabetes.[309-310]	Avoid sugar and soft drinks.
• Probiotics	Probiotic supplementation has been shown to reduce autoimmune antibody production in various autoimmune disorders, including autoimmune enteropathy, diabetes and multiple sclerosis.[311-314]	In the case of dysbiosis: • The addition of probiotics (with at least 10 billion germs of lactobacilli and bifidus bifidi strands per capsule) is recommended. • Take various types of probiotics in alternation to restore the variety in the strands better.

3. Datasets

Silobreaker software was used to obtain scientific information related to the Hashimoto disease coming from different sources such as journals, proceedings, tweets, and news. Our date consists in the fallowing feature: ID, cluster Id, Description, publication date, Source URL, publisher. And the purpose is to analyze the preform of the proposed approach to discover the hiding semantic structures related with Hashimoto and thyroiditis the description from the gather data is used to study the frequency of Hashimoto and thyroiditis appears in the documents and detecting words and phrases patterns within them to automatically clustering work groups.

The dataset was obtained from Silobreaker database which is a commercial database. We got access through Florida A&M University who provided me the right to query the data. the link for the silobreaker information is [Here] (https://www.silobreaker.com/) [9].

This data was preprocessed dropping the columns 'Id', 'ClusterId', 'Language', 'LastUpdated','CreatedDate','FirstReported'. Also, stop words and punctuation were removed, we convert to lower case all the titles.

The dataset already query can be download in my personal drive [Here] (https://drive.google.com/drive/u/0/folders/1Omtnn5e-yH3bbhW0-5fIbLgi8SEyfYBP.

4. Results

The following figures were creating with the help of libraries like gensim. Gensim stands for Generate similar and is an unsupervised library wide used for topic modeling and natural language processing based on modern statistical machine learning [10]. it can handle large text

collections of data and can preforms task like corpora, building document, word vectors and topic identification, which is one of the technics we used here, and we can observe it in some of the images. Each figure is described and explains the method we used to created it along with the relationship of the key word or major topic to the Hashimoto disorder.

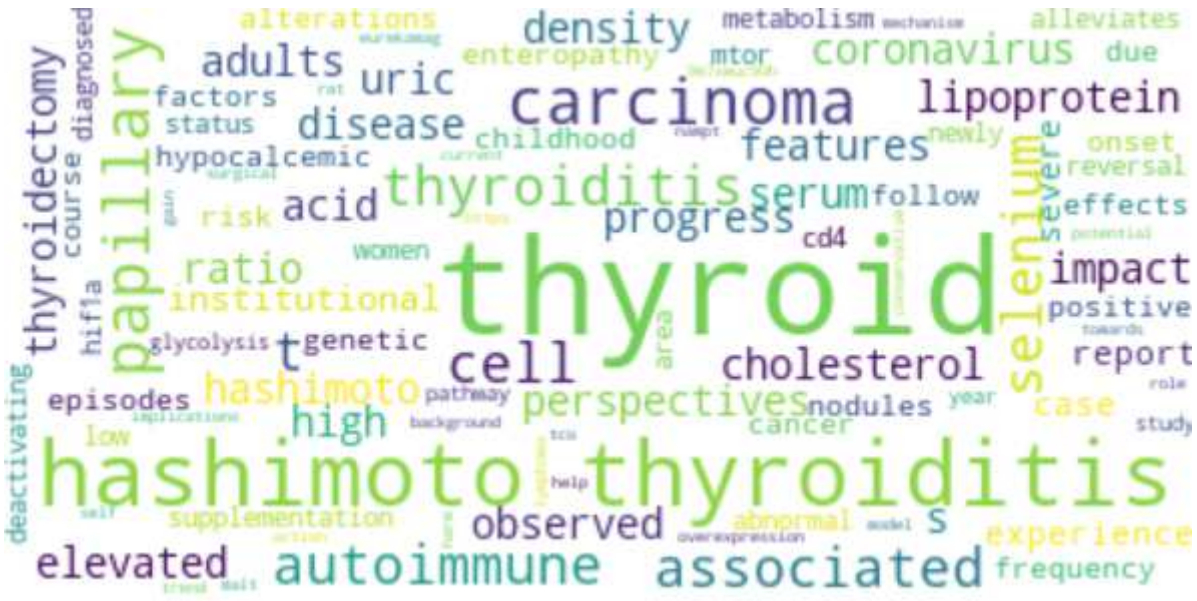


Figure 1: "Example of a Word Cloud Object"

On Figure 1 we observed an example of a word cloud object and represent the difference words found in our dataset and the size of the words means the frequency of the given words in the document. Meaning that the size of the words is proportional to the frequency of its used.

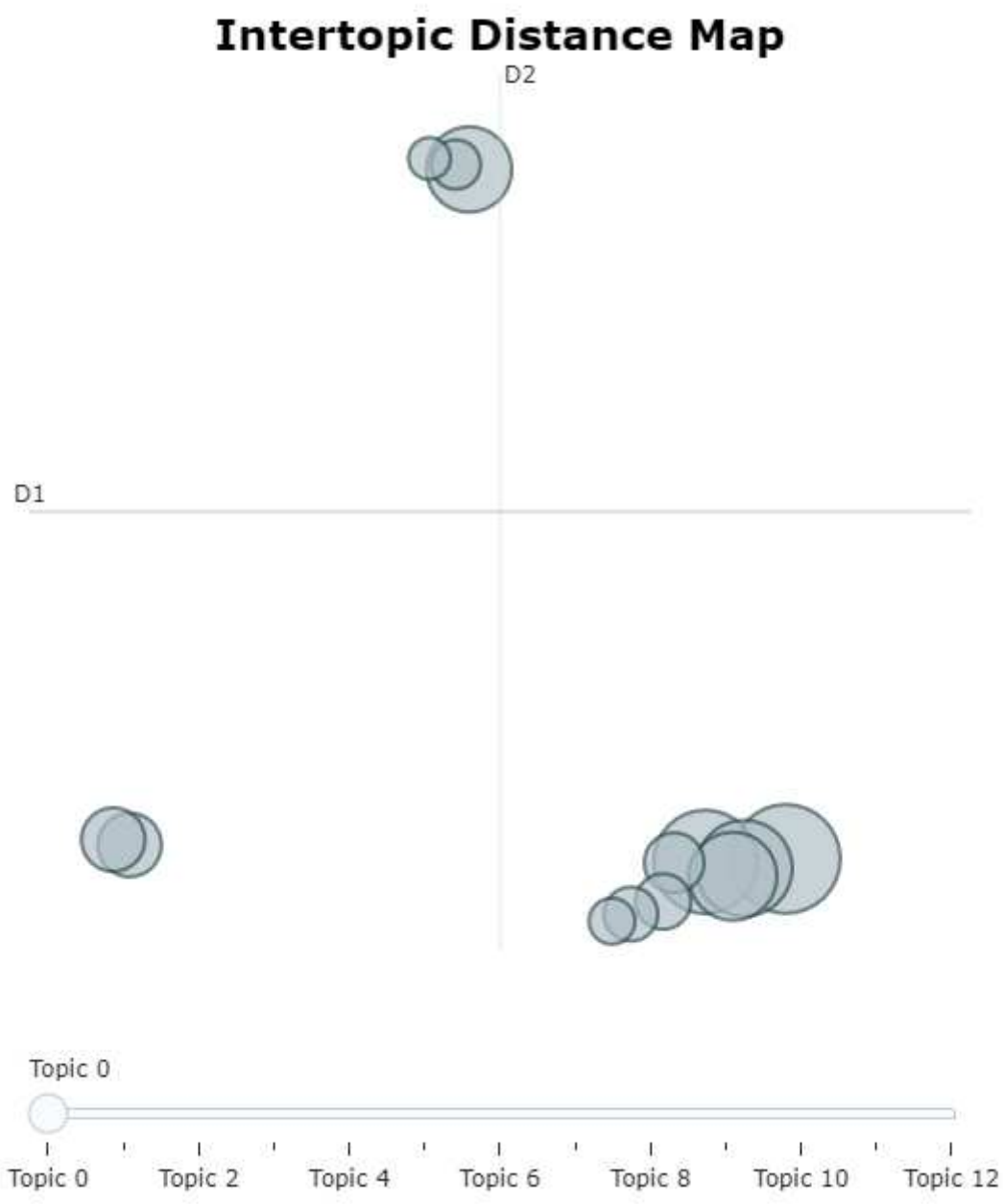


Figure 2: "Example of a Intertopic Distance Map"

Figure 2 shows an Intertopic Distance Map which is a two-dimensional space filled with circles representing the proportional number of words that belongs to each topic making the distance to each other represent the relation between the topics, meaning that topics that are closer together have more words in common. For instance, in topic 1 we observed word like hypothyroidism, Morgan, symptoms after a small search we were able to find that Morgan is a well-known writer that presented thyroiditis symptoms after giving birth which is something that happen to some women's and then recover after a couple of months, however this increments the risk of developing the syndrome later in their lives [11]. On topic 4 we see words like food, levothyroxine, liothyronine, selenium and dietary. the relationship between these words is symptom control, symptoms relive, some natural remedies and supplements [1213].

Figure 3: "Top 30 major Topics"

On figure 3 we observed a bar chart that shows 30 major terms. The bars indicate the total frequency of the term across the entire corpus. The size of the bubble measures the importance of the topics, relative to the data. for example, for visualization purposes we used the first topic that include Hashimoto, thyroiditis, and selenium. Saliency is a measure of how much the term talks about the topic. And in terms of findings is important to mentions the relationship between Hashimoto thyroiditis and selenium. Selenium is a suplement recomended for patients with this disorder that have shown a reduction on antibody levels [13].

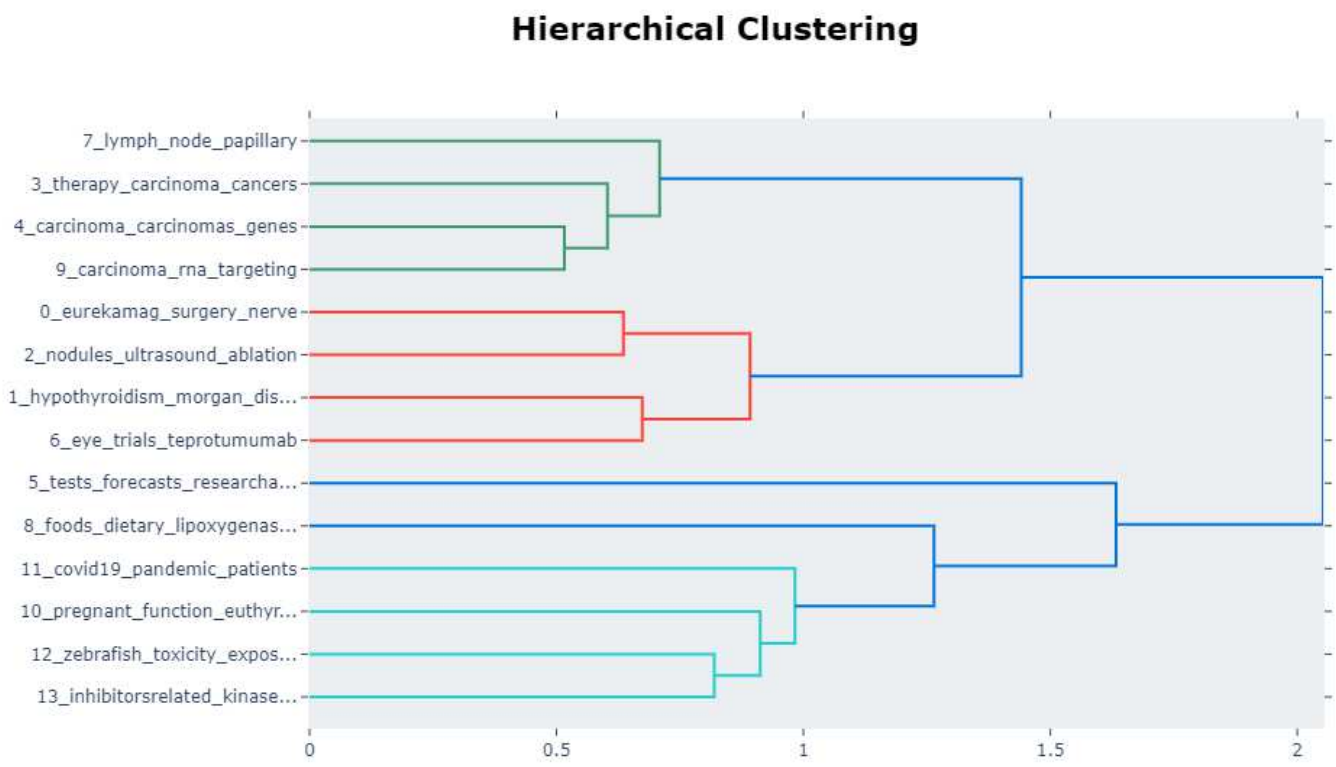


Figure 4: "Example of Hierarchical Clustering chart"

On figure 4 we can see that the dendrograms have been created joining points 4 with 9, 0 with 2, 1 with 6, and 12 with 13. The vertical height of the dendrogram shows the Euclidean distances between points. It is easy to see that Euclidean distance between points 12 and 13 is greater than the distance between point 4 and 9. This is because the algorithm is clustering by similarity, differences, and frequency of words. We observed in the dark green dendrogram topic 7,3,4,9 which are all related to an advance stage of the disorder. we can find the information about certain treatments, causes of the disorder, level of damage at certain stages. On the reds dendrograms we observe topics 0,2,1,6 which are closely related to diagnosis, early symptoms and procedures used for the diagnosis of the disorder.

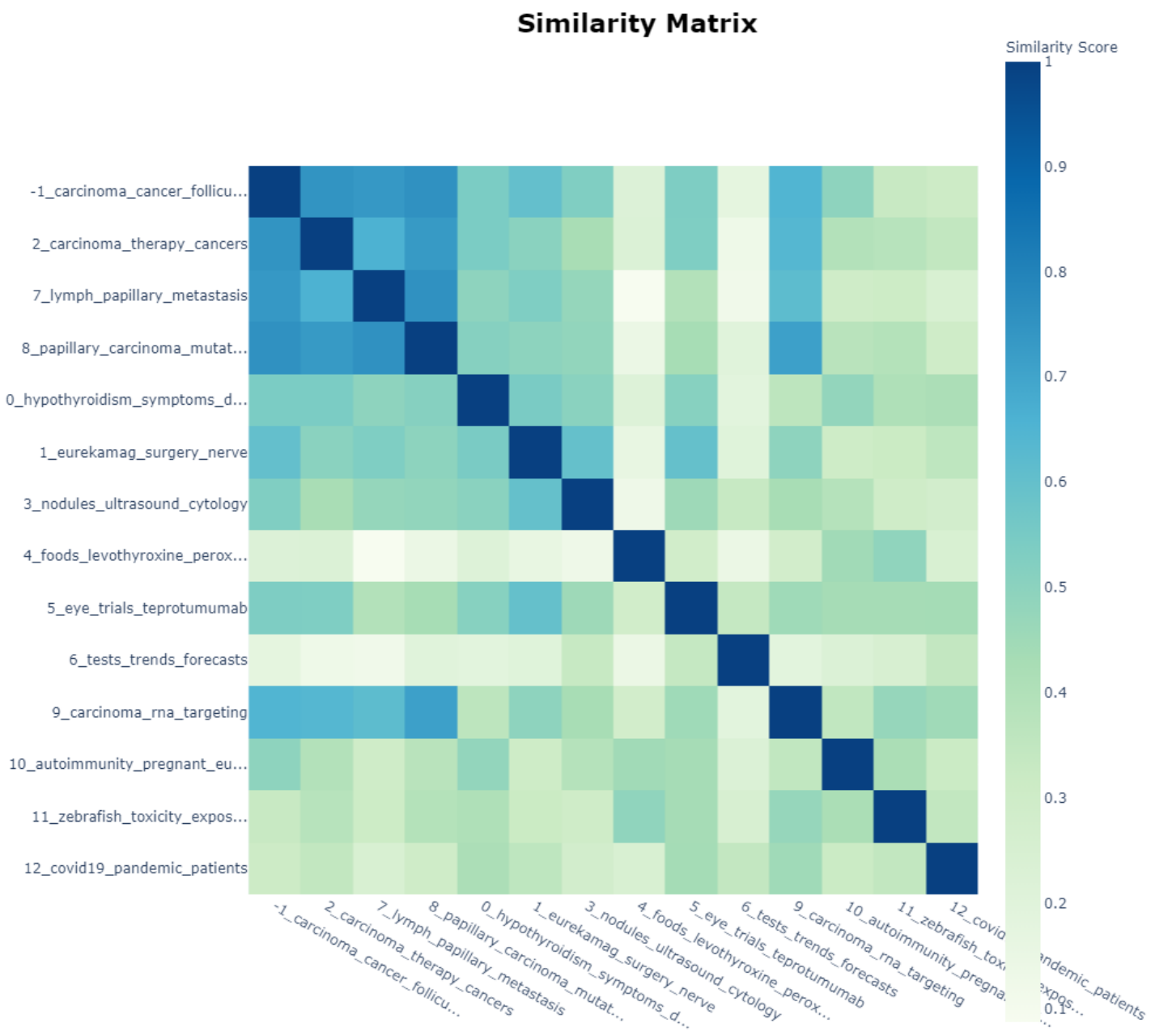


Figure 5: “Exaple of Similarity Matrix Chart”

On figure 5 we can see a similarity matrix chart, the graph is build based on similarity reached from the volume of topic and association by document, therefore the graph show groups of documents that are cluster together based on similarities. in this case the blue square is an indication of a strong similarity, and the green and light green is an indication of different topics. for instance, we are able to derive as a conclusion that carcinoma cancer, carcinoma therapy, lymph papillary metastasis and hypothyroidism are closely related. in facts they are advance stages of the disorder. E.g. Carcinoma therapy is a type of treatment that can be used for this disorder [14].

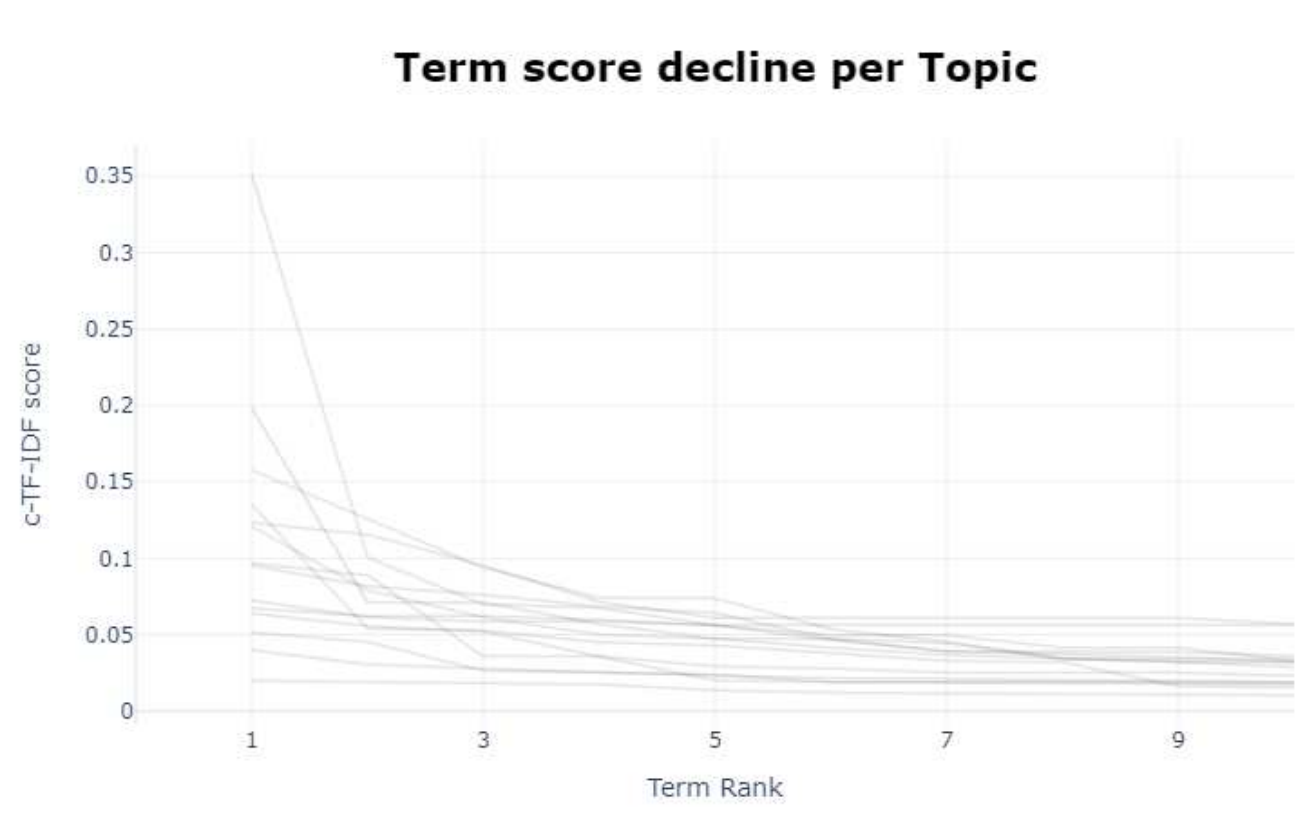


Figure 6: “Example of Term Score Decline Per Topic Chart”

On figure 6 we observed TF-IDF which is an interesting technic used on machine learning that have the ability to give weight to those words that are not frequent in the document but can carry important information. In this example we can see how topic 12, covid19 pandemic patients is the at the top of the chart and then start declining when the rank term increase. The science behind this behave is explain by the TF-IDF which is term frequency - Inverse document frequency. Therefore, covid 19 was a relative new disease, and we do not expect to have a high frequency used in the document. In this case we were able to find information about Hashimoto patients and covid19 which it seems not to causes any extreme symptoms for patient with this disorder others than the ones expected from a healthy person in other words Hashimoto patients have the same risk of a healthy person [15].

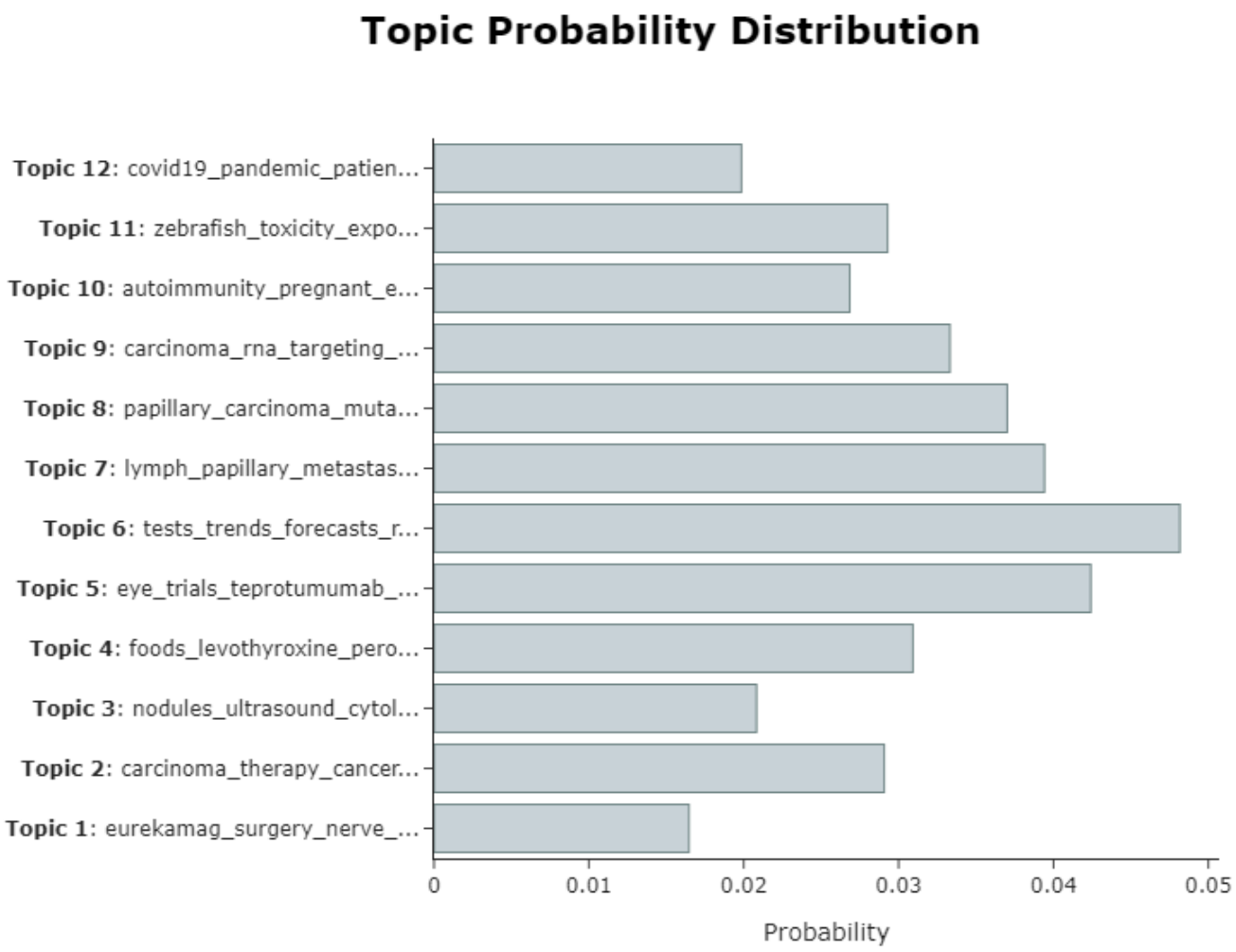


Figure 7: "Example of Topic Probability chart"

On figure 7 we see a probability distribution chart based on each topic frequency and its relationship with the main topic: Hashimoto thyroiditis causes or cure. We can see that topic 12 is the least frequent or least related since most of its content is about covid19. Then we have topic 11 zebrafish which is related to the investigation of the disorder but most of its content is about the research made on zebrafish and how had help researchers to understand thyroid diseases in other no mammals' animals, but is not closely related to the major point of this project, however, is an interesting research which have provide useful information about thyroiditis [16].

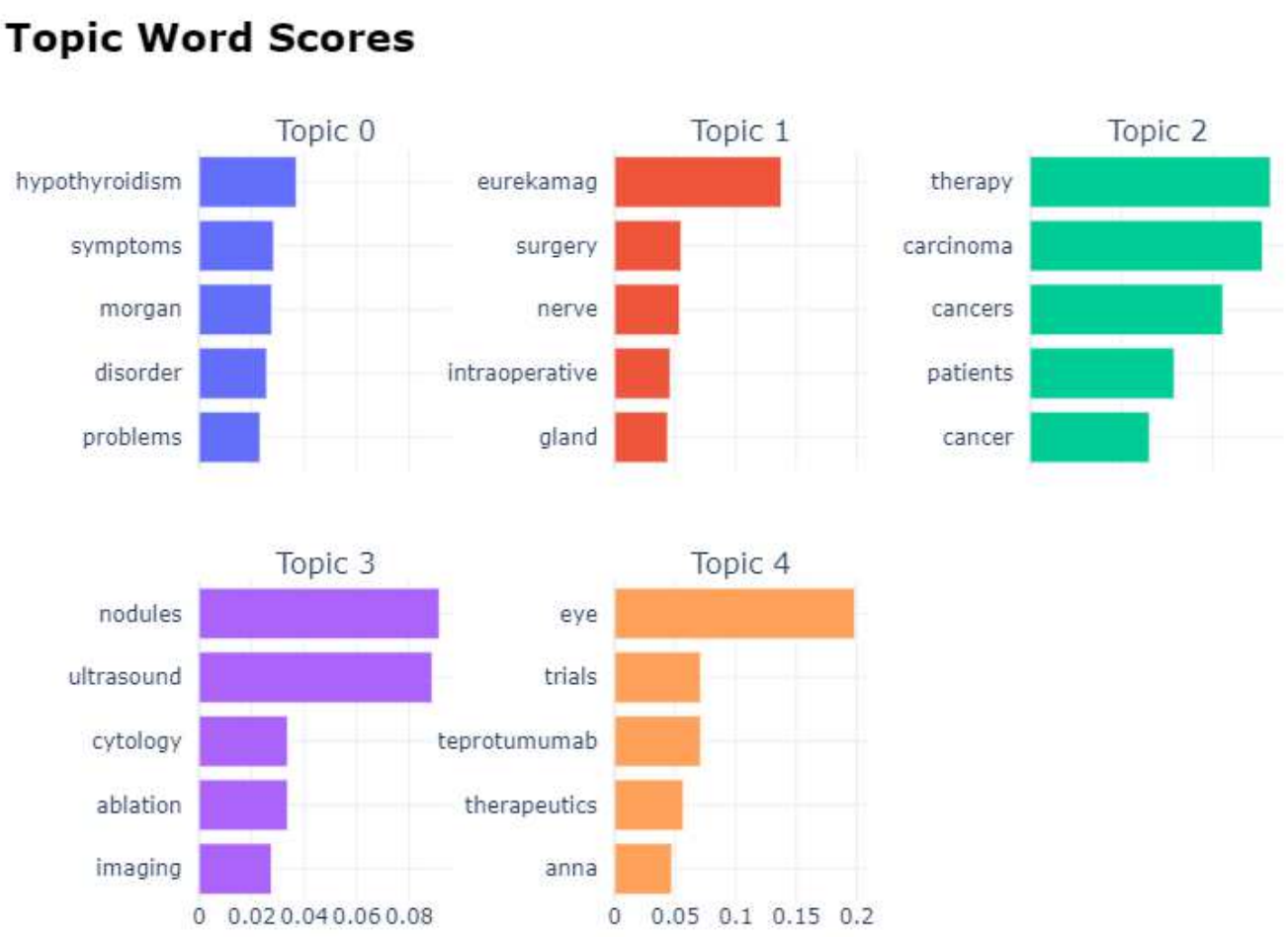


Figure 8: “Example of a Topic Word Score Chart”

On figure 8 we have Topic Word Scores chart that provides a deep understanding of large corpus of texts trough topic extraction. for instance, the data used in this project provide 5 fundamental topics from 0 to 4. Essentially each topic provided closely related words with deep information about the disorder itself, treatments, diagnosis, and symptoms. E.g. in topic number 4 we find a specific word “eye” which it does not seem to have a close relationship with Hashimoto thyroiditis but in facts is related to one of the early symptoms that the human body experiment most likely when is still undiagnosed [16]. In the same topic we also find the word teprotumumab which is an eye relieve medication recommended from doctors to relive the symptoms, in other word is not the cure but it helps [17].

5. Hashimoto Findings

As we can see our findings are wide in aspects of causes which is one of the main keys, because if we know the cause of something most likely we will be able to avoid it. However, this disorder is considered relative knew and have been around for some decades only, but it is necessary to point out the relation of diseases with the environment. Environmental changes are a fact and are affecting us every day even when we don’t notice it. We have seen an exponential increase of Hashimoto cases in the last five decades, and at the same time the last five decades have been potentially related to climate change, high levels of pollution, less fertile soils, increased use of pesticides on food, etc. It would be a good idea to think about our environment and how to help it heal since it will bring benefits for all of us[18].

Table Summary: “Finding summary on causes, descriptions and recommendations.”

Possible Causes	Description	Recomendations
Genetic predispositions	Genetically linked	Manage stress
Dietary errors	Imbalance of iodine intake	Balance is key
Nutritional deficiencies	not enough veggies, vitamins and minerals	eat more veggies
Hormone deficiencies	lover levels of vit D	Enough sleep, Take some sun light

Possible Causes	Description	Recomendations
Viral, bacterial, yeast, and parasitic infections.	H pylori, Bad guts microbes	food hygiene
Enviromental Factors	Pollution, Pesticides used etc.	Human footprint on environment

Possible causes

Genetic predispositions, Dietary errors, Nutritional deficiencies, Hormone deficiencies, Viral, bacterial, yeast, and parasitic infections [8].

Hashimoto Trigger Food

Some Food that can trigger Hashimoto are gluten, dairy, some type of grains, eggs, nuts or nightshades, sugar, sweeteners, sweet fruits, including honey, agave, maple syrup, and coconut sugar and high-glycemic fruits like watermelon, mango, pineapple, grapes, canned and dried fruits. Vegetable oil, specially hydrogenated oils, ad trans-fat. Patient with this disorder may experience symptoms of fatigue, rashes, joint pain, digestive issues, headaches, anxiety, and depression after eating some of these foods [19].

Hashimoto recommended Diets

The recommended foods are healthy fats like coconut, avocado, and olive oil, ghee, grass-fed and organic meat, wild fish, healthy fats, fermented foods like coconut yogurt, kombucha, fermented cucumbers and pickle ginger, and plenty of vegetable like Asparagus, spinach, lettuce, broccoli, beets, cauliflower, carrots, celery, artichokes, garlic, onions [19].

Environmental causes of Hashimoto

There have been an increase in the number of Hashimoto cases in the United States since 1950s. These is one of the reason research explain that Hashimoto disorder can be closely related to environmental causes since the rapid increase of cases can not only be related to family gens as it takes at least two generations to acquire and transfer gen mutation. Adding to this that for generation thru history human have been fitting microorganisms than enter our body but for the past centuries our environment has become very hygienic consequently our immune system suddeling was left without aggressors therefore humane start developing more allergies and autoimmune diseases. Another important factor is the balance of iodine intake because too much is as dangerous for people with genetic Hashimoto predisposition but too littler can be also dangerous for patients with the disorder to reduce goiter which is the enlargement of the thyroid glands [18].

It is still not enough research to state that low vitamin D levels are a cause or a consequence of the Hashimoto disorder, but it is a fact that most patients with this disorder have low levels of vitamin D this insufficient this is closely related to insufficient sun exposure [18].

The exposure to certain synthetic pesticide. An important fact is that 9 out of 12 pesticides are dangerous and persistent pollutants [18].

Symptoms of Hashimoto's

Some of the symptoms are fatigue and sluggishness, sensitivity to cold, constipation, pale and dry skin, dry eyes, puffy face, brittle nails, hair loss, enlargement of the tongue, unexplained weight gain, muscle aches, tenderness and stiffness, joint pain and stiffness, muscle weakness, excessive or prolonged menstrual bleeding, depression, memory lapses, Another symptom reported by some patients was ablation, some patient described as an acceleration of the heart rhythm [20].

Complications

Tissue damage, Abnormal look of the thyroid gland (figure 2), goiter, Heart problems, mental health issues, myxedema, birth defects [20], Nodule (figure 4 Similarity Matrix topic 3), and High antibody level. It is important to mention an association between high levels of thyroid autoantibodies and the increased of mood disorders, thyroid autoimmunity disease, celiac disease, panic disorder and major depressive disorder [8].

Recomendations

Healthy diets, exercising, selenium supplementation [8], healthy sun exposure at an adequate time, getting enough sleep is primordial for the human body, in special for the metabolism regulation and the creation of normal hormones that the human body needs, [8] lowering stress levels by physical exercise is a good idea, exercise like yoga and reiki are valuable because it also exercise you brain with meditation which is a great stress reliever.

6. Benchmark

We used benchmark to perform the process time to get topics frequency in parallel using google colab with run type: GPU and TPU. We can observe that TPU machines take less time to classify topic 1. Tensor Processor Unit (TPU) is designed to run cutting-edge machine learning models with AI services on Google Cloud [21]

Benchmark Topics Frequency:

parallel	Topic	Status	Time	processor
164	1_cancer_follicular_carcinoma_autoimmune	ok	0.53	GPU
190	1_cancer_follicular_carcinoma_autoimmune	ok	0.002	TPU

7. Conclusion

As expected, we were able to derive helpful information of the Hashimoto thyroiditis disorder. we attempted to summarize our findings concerning Hashimoto thyroiditis in aspects of causes, symptoms, recommended diets and supplements and used medication.

Our findings highlight the great potential of the model we used. certainly, topic modeling method was a precise idea for the optimization of the research process. We also used various features of genism, which allows to manipulate data texts on NPL projects. The use of clustering technics was very useful to label our findings on the large datasets. Each used graph provided useful details and key words that later help us to review each important topic in a faster manner and develop the research project with accurate results.

8. Acknowledgments

Gregor von Laszewski

Yohn J Parra

Carlos Theran

9. References

1. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, [Online resource] https://arxiv.org/abs/1810.04805 ↩
2. Helicobacter pylori infection in women with Hashimoto thyroiditis, [Online resource] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5265752/ ↩
3. I. Voloshyna, V.I Krivenko, V.G Deynega, M.A Voloshyn, Autoimmune thyrod disease related to helicobacterer pylori contamination, [Online resource] https://www.endocrine-abstracts.org/ea/0041/eposters/ea0041gp213_eposter.pdf ↩
4. How your diet can trigger Hashimoto's, [Online resource] https://www.boostthyroid.com/blog/2019/4/5/how-your-diet-can-trigger-hashimotos ↩

5. Hypothyroidism in Context: Where We've Been and Where We're Going, [Online resource] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6822815/ ↩

6. BERT Explained: State of the art language model for NLP https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 ↩

7. Pavan Sanagapati, Knowledge Graph & NLP Tutorial-(BERT,spaCy,NLTK), [Online resource] https://www.kaggle.com/pavansanagapati/knowledge-graph-nlp-tutorial-bert-spacy-nltk ↩

8. Hashimoto's Thyroiditis, A Common Disorder in Women: How to Treat It, [Online resource] https://www.townsendletter.com/article/441-hashimotos-thyroiditis-common-disorder-in-women/ ↩

9. Silobreaker: Intelligent platform for the data era https://www.silobreaker.com ↩

10. Gensim Tutorial – A Complete Beginners Guide, [Onile resource] https://www.machinelearningplus.com/nlp/gensim-tutorial/ ↩

11. Julia Haskins, Thyroid Conditions Raise the Risk of Pregnancy Complications, [Online resource] https://www.healthline.com/health-news/children-thyroid-conditions-raise-pregnancy-risks-052913 ↩

12. How your diet can trigger Hashimoto's, [Online resource] https://www.boostthyroid.com/blog/2019/4/5/how-your-diet-can-trigger-hashimotos ↩

13. Selenium Supplementation for Hashimoto's Thyroiditis, [Online resource] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4005265/ ↩

14. Thyroid Cancer Treatment, [Online resource] https://www.cancer.gov/types/thyroid/patient/thyroid-treatment-pdq ↩

15. Hashimoto's Disease And Coronavirus (COVID-19), [Online resource] https://www.palomahealth.com/learn/coronavirus-and-hashimotos-disease ↩

16. How zebrafish research has helped in understanding thyroid diseases, [Online resource] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5730863/ ↩

17. Teprotumumab for the Treatment of Active Thyroid Eye Disease, [Online resource] https://www.nejm.org/doi/full/10.1056/nejmoa1910434 ↩

18. 11 environmental triggers of Hashimoto's, [Online research] https://www.boostthyroid.com/blog/11-environmental-triggers-of-hashimotos ↩

19. Hashimoto's low thyroid autoimmune, [Online research] https://www.redriverhealthandwellness.com/diet-hashimotos-hypothyroidism/ ↩

20. Hashimoto's disease, [Online research] https://www.mayoclinic.org/diseases-conditions/hashimotos-disease/symptoms-causes/syc-20351855 ↩

21. TPU: Tensor Processor Unit https://cloud.google.com/tpu ↩

Report: Dentronics: Classifying Dental Implant Systems by using Automated Deep Learning

Artificial intelligence is a branch of computer science that focuses on building and programming machines to think like humans and mimic their actions. The proper concept definition of this term cannot be achieved simply by applying a mathematical, engineering, or logical approach but requires an approach that is linked to a deep cognitive scientific inquiry. The use of machine-based learning is constantly evolving the dental and medical field to assist with medical decision making process. In addition to diagnosis of visually confirmed dental caries and impacted teeth, studies applying machine learning based on artificial neural networks to dental treatment through analysis of dental magnetic resonance imaging, computed tomography, and cephalometric radiography are actively underway, and some visible results are emerging at a rapid pace for commercialization. Researchers have found deep convolutional neural networks to have a future place in the dental field when it comes to classification of dental implants using radiographic images.

Tags: project reu ai health

9 minute read

Check Report passing Status passing Status: final, Type: Project

Jamyla Young, su21-reu-376, Edit

Abstract

Artificial intelligence is a branch of computer science that focuses on building and programming machines to think like humans and mimic their actions. The proper concept definition of this term cannot be achieved simply by applying a mathematical, engineering, or logical approach but requires an approach that is linked to a deep cognitive scientific inquiry. The use of machine-based learning is constantly evolving the dental and medical field to assist with medical decision making process.In addition to diagnosis of visually confirmed dental caries and impacted teeth, studies applying machine learning based on artificial neural networks to dental treatment through analysis of dental magnetic resonance imaging, computed tomography, and cephalometric radiography are actively underway, and some visible results are emerging at a rapid pace for commercialization.

Contents

Keywords: Dental implants, Deep Learning, Prosthodontics, Implant classificiation, Artificial Intelligence, Neural Networks.

1. Introduction

Dental implants are ribbed oral protheses typically made up of biocompatible titanium to replace the missing root(s) of an absent tooth. These dental protheses are used to support the jaw bone to prevent deterioration due to an absent root[1]. This is referred to as bone resorption which can result to facial malformation as well as reduced oral function such as biting and chewing. These devices are composed of three elements that imitates a natural tooth function and structure.The implant which are typically ribbed and threaded to promote stability while integrating within the bone tissue. The osseointegration process usually takes 6-8 months to rebuild the bone to support the implant. An implant abutment is fixed on top of the implant to act as a base for prosthetic devices [2]. Prefabricated abutments are manufactured in many shapes, sizes and angles depending on the location of the implant and the types of prothesis that will be attached. Dental abutments support a range of prothetic devices such as dental crowns, bridges, and dentures [3].

Osseointegrated dental implants depend on various factors that affect the anchorage of the implant to the bone tissue. Successful surgical anchoring techniques can contribute to long term success of implant stability. Primary stability plays a role 2 week postoperatively by achieving mechanical retention of the implant. It helps establish a mechanical microenvironment for gradual bone healing, or osseointegration-This is secondary implant stability. Bone type, implant length, implant and diameter influences primary and secondary implant stability. Implant length can range from 6mm to 20mm; however, the most common lengths are between 8mm to 15mm. Many studies suggest that implant length contribute to decreasing bone stress and increasing implant stability. Bone stress can occur at both the cortical and cancellous part of the bone. Increasing implant length will decrease stress in the cancellous part of the bone while increasing the implant diameter can decrease stress in the cortical part of the bone[4]. Bone type can promote positive bone stimulation around an implant improving the overall function. There are four different types: Type I, Type II, Type III, and Type IV. Type I is the most dense of them which provides more cortical anchorage but has limited vascularity. Type II is the best for osseointegration because it provides good cortical anchorage and has better vascularity than type I. Type III and IV have a thin layer of cortical bone which decrease the success rate of primary stability[5].

Implant stability can be measured using the Implant Stability Quotient (ISQ) as an indirect indicator to determine the time frame for implant loading and prognostic indicator for implant failure [4]. This can be measured by resonance frequency analysis (RFA) immediately after the implant has been placed. Resonance frequency analysis is the measurement in which a device vibrates in response to frequencies in the range of 5-15 kHz. The peak amplitude of the response is then encoded into the implant stability quotient (ISQ). The clinical range of ISQ is from 55-80. High stability is >70 ISQ while medium stability is between 60-69 ISQ. Low stability is <60 ISQ[6].

There are over 2000 types of dental implant systems (DIS) that differs in diameter, length, shape, coating, and surface material properties. These devices have more than a 90% long termed survival rate which ranges more than 10 years. Inevitably, biological and mechanical complications such as fractures, low implant stability, and screw loosening can occur. Therefore, identifying the correct Dental Implant System is essential to repair or replace the existing system. Methods and techniques that enables clear identification is insufficient [7].

Artificial intelligence is a branch of computer science that focuses on building and programming machines to think like humans and mimic their actions. A deep convolutional neural network (DCNN) is a brach of artificial intelligence that applies multiple layers of nonlinear processing units for feature extraction, transformation, and classification of high dimensional datasets. Deep convolutional neural networks are commonly used to identify patterns in images and videos. The structure typically consist of four types of layers: convolution, pooling, activation, and fully connected. These neural networks use images as an input to train a classifier which employs a mathematical operation called a convolution. Deep neural networks have been successfully applied in the dental field and demonstrated advantages in terms of diagnosis and prognosis. Using automated deep convolutional neural networks is highly efficient in classifying different dental implant systems compared to most dental professionals[7].

2. Data sets

Researchers at Daejon Dental Hospital used automated deep convolutional neural networks to evaluate the efficacy of its ability to classify dental implant systems and compare the performance with dental professionals using radiographic images.

11,980 raw panoramic and periapical radiographic images of dental implant systems were collected. these images were then randomly divided into 2 groups: 9584 (80%) images were selected for the training dataset and the remaining 2396 (20%) images were used as the testing dataset.

2.1 Dental implant classification

Dental implant systems were classified into six different types with a diameter of 3.3-5.0mm and a length of 7-13mm.

- Astra OsseoSpeed TX (Dentsply IH AB, Molndal, Sweden), with a diameter of 4.5–5.0 mm and a length of 9–13 mm;
- Implantium (Dentium, Seoul, Korea), with a diameter of 3.6–5.0 mm and a length of 8–12 mm;
- Superline (Dentium, Seoul, Korea), with a diameter of 3.6–5.0 mm and a length of 8–12 mm;
- TSIII (Osstem, Seoul, Korea), with a diameter of 3.5–5.0 mm and a length of 7–13 mm;
- SLActive BL (Institut Straumann AG, Basel, Switzerland), with a diameter of 3.3–4.8 mm and a length of 8–12 mm;
- SLActive BLT (Institut Straumann AG, Basel, Switzerland), with a diameter of 3.3–4.8 mm and a length of 8–12 mm.

2.2 Deep Convulutional Neural Network

Using Neuro-T to automatically select the model and optimize hyper-parameter. During training and inference, the automated DCNN automatically creates effective deep learning models and searches the optimal hyperparameters. An Adam optimizer with L2 regularization was used for transfer learning. The batch size was set to 432, and the automated DCNN architecture consisted of 18 layers with no dropout.

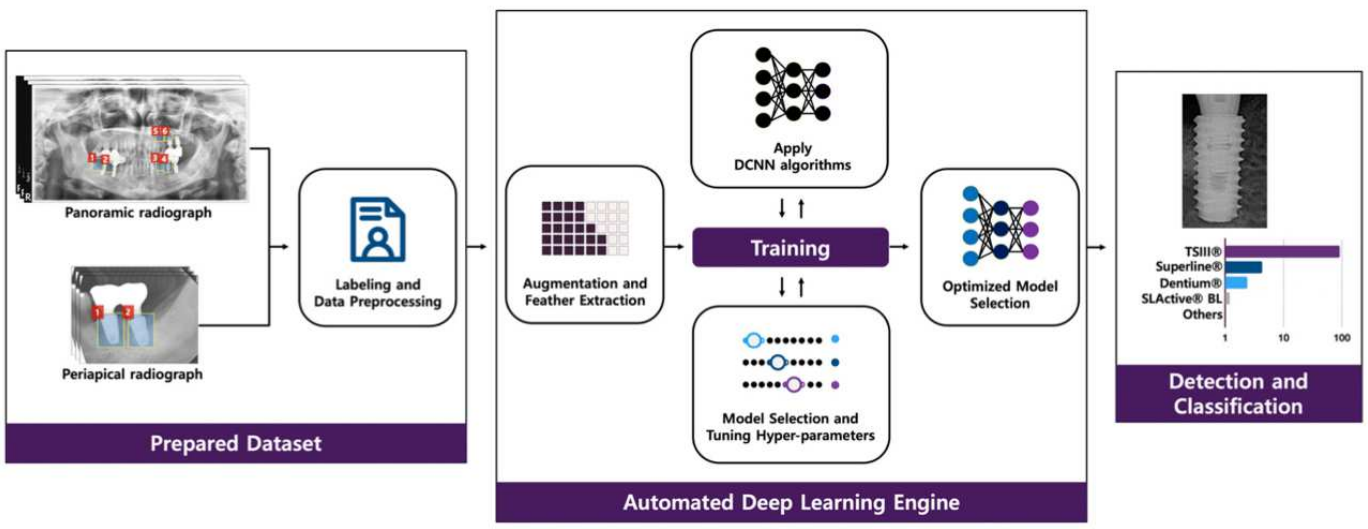


Figure 1: Overview of an automated deep convolutional neural network [7].

3. Results

For the evaluation, the following statistical parameters were taken into account: receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), 95% confidence intervals (CIs), standard error (SE), Youden index (sensitivity + specificity − 1), sensitivity, and specificity, which were calculated using Neuro-T and R statistical software . Delong's method was used to compare the AUCs generated from the test dataset, and the significance level was set at $p < 0.05$.

The accuracy of the automated DCNN abased on the AUC, Youden index, sensitivity, and specificity for the 2,396 panoramic and periapical radiographic images were 0.954(95% CI = 0.933–0.970, SE = 0.011), 0.808, 0.955, and 0.853, respectively. Using only panoramic radiographic images (n = 1429), the automated DCNN achieved an AUC of 0.929 (95% CI = 0.904–0.949, SE = 0.018, Youden index = 0804, sensitivity = 0.922, and specificity = 0.882), while the corresponding value using only periapical radiographic images (n = 967) achieved an AUC of 0.961 (95% CI = 0.941–0.976, SE = 0.009, Youden index = 0.802, sensitivity = 0.955, and specificity = 0.846). There were no significant differences in accuracy among the three ROC curves.

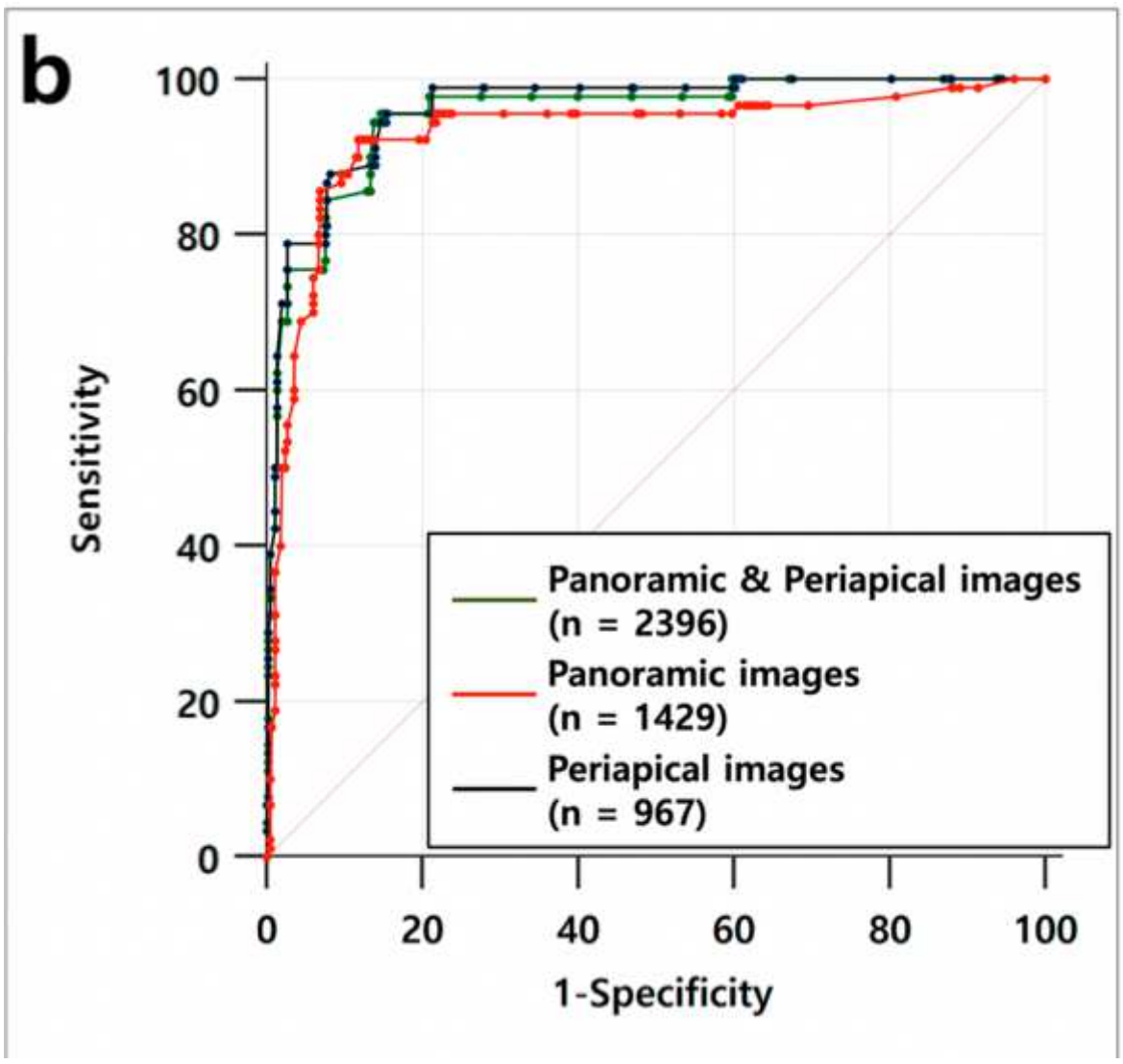


Figure 2: The accuracy of the automated DCNN for the test dataset did not show a significant difference among the three ROC three ROC curves based on DeLong's method [7].

The Straumann SLActive BLT implant system has a relatively large tapered shape compared to other types of DISs. Thus, the automated DCNN (AUC = 0.981, 95% CI = 0.949–0.996). However, for the Dentium Superline and Osstem TSIII implant systems that do not have conspicuous characteristic elements with a tapered shape, the automated DCNN classified correctly with an AUC of 0.903 (95% CI = 0.850–0.967) and 0.937 (95% CI = 0.890–0.967)

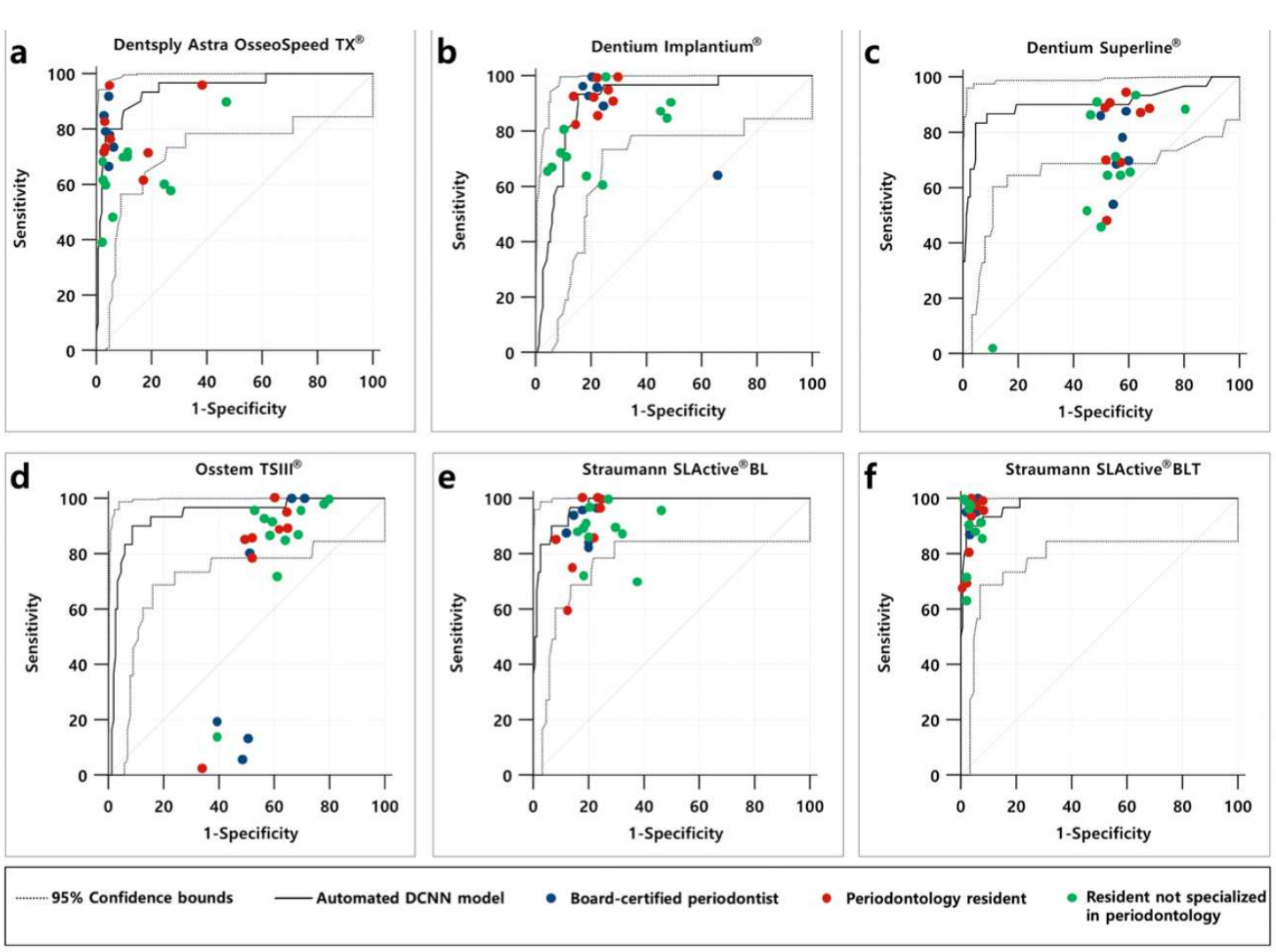


Figure 3 (a-f): Performance of the automated DCNN and comparison with dental professionals for classification of six types of DIS [7]

4. Conclusion

Nonetheless, this study has certain limitations. Although six types of DISs were selected from three different dental hospitals and categorized as a dataset, the training dataset was still insufficient for clinical practice. Therefore, it is necessary to build a high-quality and large-scale dataset containing different types of DISs. If time and cost are not limited, the automated DCNN can be continuously trained and optimized for improved accuracy. Additionally, the automated DCNN regulates the entire process, including appropriate model selection and optimized hyper-parameter adjustment. The automated DCNN can help clinical dental practitioners to classify various types of DISs based on dental radiographic images. Nevertheless, further studies are necessary to determine the efficacy and feasibility of applying the automated DCNN in clinical practice.

5. Acknowledgments

1. Carlos Theran, REU Instructor
2. Yohn Jairo Parra, REU Instructor
3. Gregor von Laszewski, REU Instructor
4. Victor Adankai, Graduate Student
5. Jacques Fleischer, REU peer
6. Florida Agricultural and Mechanical University

6. References

[^3] Gregor von Laszewski, Cloudmesh StopWatch and Benchmark from the Cloudmesh Common Library, [GitHub https://github.com/cloudmesh/cloudmesh-common

---

1. Karras, Spiro, Look at the structure of dental implants.(2020, September 2). https://www.drkarras.com/a-look-at-the-structure-of-dental-implants/ ↩
2. Ghidrai, G. (n.d.). Dental implant abutment. Stomatologia pe intelesul tuturor. https://www.infodentis.com/dental-implants/abutment.php __
3. Bataineh, A. B., & Al-Dakes, A. M. (2017, January 1). The influence of length of implant on primary stability: An in vitro study using resonance frequency analysis. Journal of clinical and experimental dentistry. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5268121/ __
4. Huang, H., G, Wu., & E, Hunziker. (2020). The clinical significance of implant Stability QUOTIENT (ISQ) MEASUREMENTS: A literature review. Journal of oral biology and craniofacial research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7494467/ __
5. Li, J., Yin, X., Huang, L., Mouraret, S., Brunski, J. B., Cordova, L., Salmon, B., & Helms, J. A. (2017, July). Relationships among Bone QUALITY, IMPLANT Osseointegration, and WNT SIGNALING. Journal of dental research https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5480808/ __
6. Möhlhenrich, S. C., Heussen, N., Modabber, A., Bock, A., Hölzle, F., Wilmes, B., Danesh, G., & Szalma, J. (2020, July 24). Influence of bone density, screw size and surgical procedure on orthodontic mini-implant placement – part b: Implant stability. International Journal of Oral and Maxillofacial Surgery. https://www.sciencedirect.com/science/article/abs/pii/S0901502720302496 ↩

7. Lee JH, Kim YT, Lee JB, Jeong SN. A Performance Comparison between Automated Deep Learning and Dental Professionals in Classification of Dental Implant Systems from Dental Imaging: A Multi-Center Study. Diagnostics (Basel). 2020 Nov 7;10(11):910. doi: 10.3390/diagnostics10110910. PMID: 33171758; PMCID: PMC7694989. ↩

Project: Analyzing the Advantages and Disadvantages of Artificial Intelligence for Breast Cancer Detection in Women

Breast Cancer is one of the most dangerous type of disease that affects many women. For detecting Breast Cancer, machine learning techniques are applied to improve the accuracy of diagnosis.

Tags: project reu ai health

8 minute read

Check Report failing Status failing Status: draft, Type: Project

RonDaisja Dunn, su21-reu-377, Edit

Abstract

The AI system is improving its diagnostic accuracy by significantly decreasing unnecessary biopsies. AI's algorithms for workflow improvement and outcome analyses are advancing. Although artificial intelligence can be beneficial to detecting and diagnosing breast cancer, there are some limitations to its techniques. The possibility of insufficient quality, quantity or appropriateness is possible. When compared to other imaging modalities, breast ultrasound screening offers numerous benefits, including a cheaper cost, the absence of ionizing radiation, and the ability to examine pictures in real time. Despite these benefits, reading breast ultrasound is a difficult process. Different characteristics, such as lesion size, shape, margin, echogenicity, posterior acoustic signals, and orientation, are used by radiologists to assess US pictures, which vary substantially across individuals. The development of AI systems for the automated detection of breast cancer using Ultrasound Screening pictures has been aided by recent breakthroughs in deep learning.

Contents

Keywords: project, reu, breast cancer, Artificial Intelligence, diagnosis detection, women, early detection, advantages, disadvantages

1. Introduction

The leading cause of cancer death in women worldwide is breast cancer. This deadly form of cancer has impacted many women across the globe. Specifically, African American women have been the most negatively impacted. Their death rates due to breast cancer have surpassed all other ethnicities. Serial screening is an essential part in detecting Breast cancer. Detecting the early stages of this disease and decreasing mortality rates is most effective by utilizing serial screening. Some women detect that they could have breast cancer by discovering a painless lump in their breast. Other women began to detect that there may be a problem due to annual and bi-annual breast screenings. Screening in younger women is not likely, because breast cancer is most likely to be detected in older women. Women from the age 55 to 69 are likely to be diagnosed with breast cancer. Women who frequently participate in receiving mammograms reduce the chance of breast cancer mortality.

Artificial Intelligence is the branch of computer science dedicated to the development of computer algorithms to accomplish tasks traditionally associated with human intelligence, such as the ability to learn and solve problems. This branch of computer science coincides with diagnosing breast cancer in individuals because of the use of radiology. Radiological images can be quantitated and can inform and train some algorithms. There are many terms that relate to Artificial Intelligence such as artificial neural networks (ANNs), machine and deep learning (ML, DL). These techniques complete duties in healthcare, including radiology. Machine learning interprets pixel data and patterns from mammograms. Benign or malignant features for inputs are defined by microcalcifications. Deep learning is effective in breast imaging, where it can identify several features such as edges, textures, and lines. More intricate features such as organs, shapes, and lesions can also be detected. Neural networks algorithms are used for image feature extractions that cannot be detected beyond human recognition.

A computer system that can perform complicated data analysis and picture recognition tasks is known as artificial intelligence (AI). Both massive processing power and the application of deep learning techniques made this possible, and are increasingly being used in the medical field. Mammograms are the x-rays used to detect breast cancer in women. Early detection is important to reduce deaths, because that is when the cancer is most treatable. Screenings have presented a 15%-35% false report in screened women. Errors and the ability to view the cancer from the human eye are the reasons for the false reports. Artificial Intelligence offers many advantages when detecting breast cancer. These advantages include less false reports, fewer cases missed because the AI program does not get tired and it reduces the effort of reading thousands of mammograms.

2. Methods From Literature Review

The goal was to emphasize the present data in terms of test accuracy and clinical utility results, as well as any gaps in the evidence. Women are screened by getting photos taken of each breast from different views. Two readers are assigned to interpret the photographs in a sequential order. Each reader decides whether the photograph is normal or whether a woman should be recalled for further examination. Arbitration is used when there is a disagreement. If a woman is recalled, she will be offered extra testing to see if she has cancer.

Another goal is to detect cancer at an earlier stage during screening so that therapy can be more successful. Some malignancies found during screening, on the other hand, might never have given the woman symptoms. Overdiagnosis is a term used to describe a situation in which a person has caused harm to another person during their lifetime. As a result, overtreatment (unnecessary treatment) occurs. Since some malignancies are overlooked during screening, the women are misled.

The methods in diagnostic procedures vary between radiologists and Artificial Intelligence networks. In a breast ultrasound exam, radiologists look for abnormal abnormalities in each image, while AI networks analyze each image in an exam that is processed separately using a ResNet-18 model, and a saliency map is generated, identifying the most essential sections. With radiologists, the focus is on photos with abnormal lesions and with AI networks the image is given an attention score based on its relative value. To make a final diagnosis, radiologists consider signals in all photos, and AI computes final predictions for benign and malignant results by combining information from all photos using an attention technique.

3. Results From Literature Review

Using pathology data, each breast in an exam was given a label indicating the presence of cancer. Image-guided biopsy or surgical excision were used to collect tissues for pathological tests. The AI system was shown to perform comparably to board-certified breast radiologists in the reader study subgroup. In this reader research, the AI system detected tumors with the same sensitivity as radiologists, but with greater specificity, a higher PPV, and a lower biopsy rate. Furthermore, the AI system outperformed all ten radiologists in terms of AUROC and AUPRC. This pattern was replicated in the subgroup study, which revealed that the algorithm could correctly interpret Ultrasound Screening examinations that radiologists considered challenging.

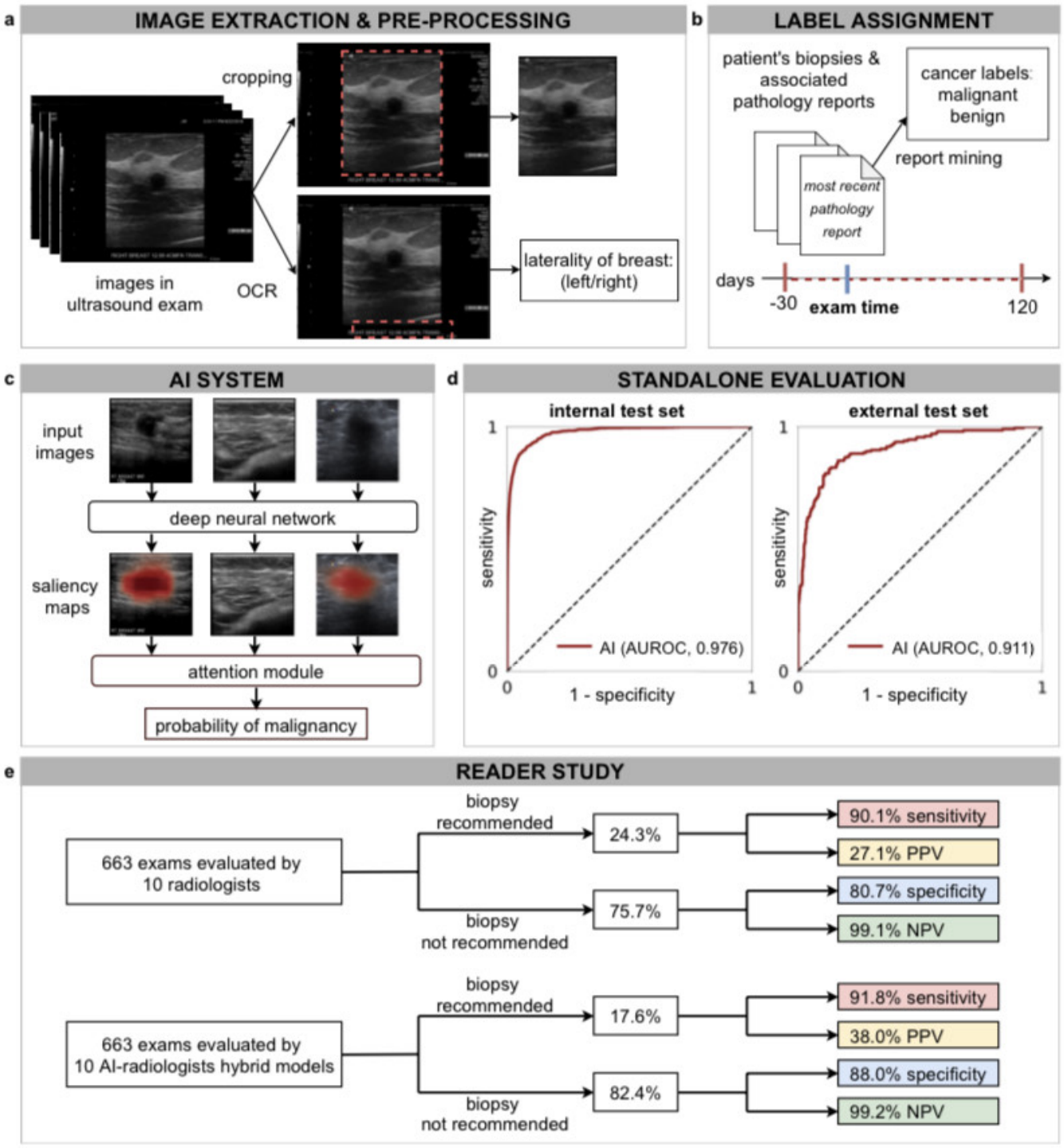


Figure 1: Analysis of saliency maps on a qualitative level- This figure displays the sagittal and transverse views of the lesion (left) and the AI's saliency maps indicating the anticipated sites of benign (center) and malignant (right) findings in each of the six instances (a-f) from the reader study.

4. Datasets

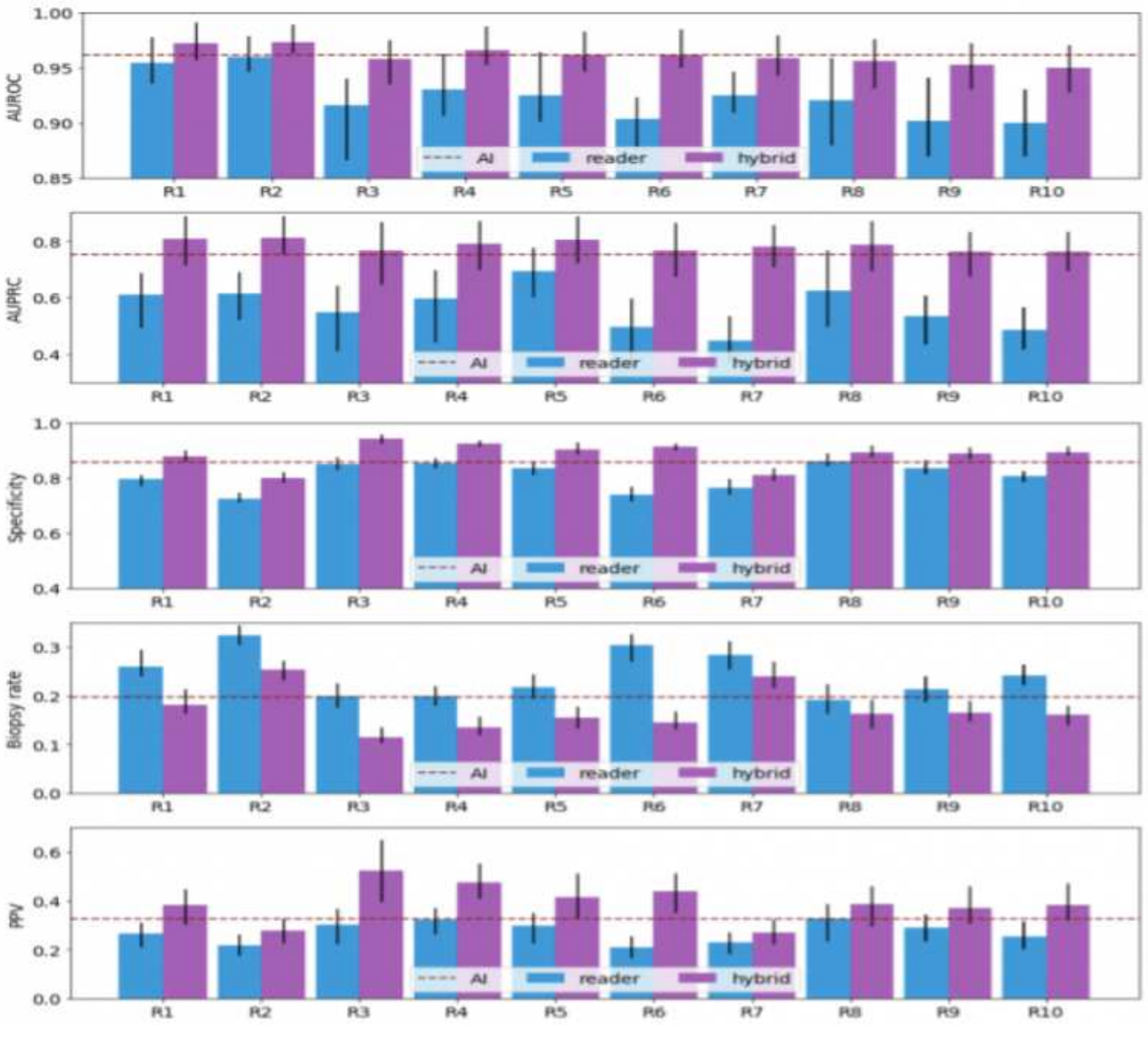


Figure 2: The probabilistic forecasts of each hybrid model were randomly divided to fit the reader's sensitivity. The dichotomization of the AI's predictions matches the sensitivity of the average radiologists. Readers' AUROC, AUPRC, specificity, and PPV improve as a result of the collaboration between AI and readers, whereas biopsy rates decrease.

5. Conclusion

There are some benefits of AI help with mammogram screenings. The reduction in treatment expenses is one of the advantages of screening. Treatment for people who are diagnosed sooner is less invasive and expensive, which may lessen patient anxiety and improve their prognosis. One or all human readers could be replaced by AI. AI may be used to pre-screen photos, with only the most aggressive ones being reviewed by humans. AI could be employed as a reader aid, with the human reader relying on the AI system for guidance during the reading process.

However, there is also fear that AI could discover changes that would never hurt women. Because the adoption of AI systems will alter the current screening program, it's crucial to determine how accurate AI is in breast screening clinical practice before making any changes. It's uncertain how effective AI is at detecting breast cancer in different sorts of women or in different groups of women (for example different ethnic groups). AI could significantly minimize staff workload, as well as the proportion of cancers overlooked during screening, and the amount of women who are asked to return for more tests despite the fact that they do not have cancer. According to the findings of the reader survey, such teamwork between AI systems and radiologists increases diagnosis accuracy and decreases false positive biopsies for all 10 radiologists. This research indicated that integrating the Artificial intelligence system's predictions enhanced the performance of all readers.

6. Acknowledgments

Thank you to the extremely intellectual, informative, patient and courteous instructors of the Research Experience for Undergraduates Program.

1. Carlos Theran, REU Instructor
2. Yohn Jairo Parra, REU Instructor
3. Gregor von Laszewski, REU Instructor
4. Victor Adankai, Graduate Student
5. REU Peers
6. Florida Agricultural and Mechanical University

7. References

1. Coleman C. Early Detection and Screening for Breast Cancer. Semin Oncol Nurs. 2017 May;33(2):141-155. doi: 10.1016/j.soncn.2017.02.009. Epub 2017 Mar 29. PMID: 28365057
2. Freeman, K., Geppert, J., Stinton, C., Todkill, D., Johnson, S., Clarke, A., & Taylor-Phillips, S. (2021, May 10). Use of Artificial Intelligence for Image Analysis in Breast Cancer Screening. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/987021/AI_in_BSP_Rapid_review_consultation_2021.pdf
3. Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., & Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. PloS one, 16(4), e0250370.
4. Mendelsonm, Ellen B., Artificial Intelligence in Breast Imaging: Potentials and Limitations. American Journal of Roentgenology 2019 212:2, 293-299
5. Seely, J. M., & Alhassan, T. (2018). Screening for breast cancer in 2018-what should we be doing today?. Current oncology (Toronto, Ont.), 25(Suppl 1), S115–S124.
6. Shamout, F. E., Shen, A., Witowski, J., Oliver, J., & Geras, K. (2021, June 24). Improving Breast Cancer Detection in Ultrasound Imaging Using AI. NVIDIA Developer Blog. https://developer.nvidia.com/blog/improving-breast-cancer-detection-in-ultrasound-imaging-using-ai/

Project: Detection of Autism Spectrum Disorder with a Facial Image using Artificial Intelligence

This project uses artificial intelligence to explore the possibility of using a facial image analysis to detect Autism in children. Early detection and diagnosis of Autism, along with treatment, is needed to minimize some of the difficulties that people with Autism encounter. Autism is usually diagnosed by a specialist through various Autism screening methods. This can be an expensive and complex process. Many children that display signs of Autism go undiagnosed because there families lack the expenses needed to pay for Autism screening and diagnosing. The development of a potential inexpensive, but accurate way to detect Autism in children is necessary for low-income families. In this project, a Convolutional Neural Network (CNN) is utilized, along with a dataset obtained from Kaggle. This dataset consists of collected images of male and female, autistic and non-autistic children between the ages of two to fourteen years old. These images are used to train and test the CNN model. When one of the images are received by the model and importance is assigned to various features in the image, an output variable (autistic or non-autistic) is received.

Tags: project reu ai health

11 minute read

Check Report passing Status passing Status: final: Project

Myra Saunders, su21-reu-378, Edit

- Utilized CNN Code: autism_classification.ipynb

Abstract

This project uses artificial intelligence to explore the possibility of using a facial image analysis to detect Autism in children. Early detection and diagnosis of Autism, along with treatment, is needed to minimize some of the difficulties that people with Autism encounter. Autism is usually diagnosed by a specialist through various Autism screening methods. This can be an expensive and complex process. Many children that display signs of Autism go undiagnosed because there families lack the expenses needed to pay for Autism screening and diagnosing. The development of a potential inexpensive, but accurate way to detect Autism in children is necessary for low-income families. In this project, a Convolutional Neural Network (CNN) is utilized, along with a dataset obtained from Kaggle. This dataset consists of collected images of male and female, autistic and non-autistic children between the ages of two to fourteen years old. These images are used to train and test the CNN model. When one of the images are received by the model and importance is assigned to various features in the image, an output variable (autistic or non-autistic) is received.

Contents

Keywords: Autism Spectrum Disorder, Detection, Artificial Intelligence, Deep Learning, Convolutional Neural Network.

1. Introduction

Autism Spectrum Disorder (ASD) is a broad range of lifelong developmental and neurological disorders that usually appear during early childhood. Autism affects the brain and can cause challenges with speech and nonverbal communication, repetitive behaviors, and social skills. Autism Spectrum Disorder can occur in all socioeconomic, ethnic, and racial groups, and can usually be detected and diagnosed from the age of three years old and up. As of June 2021, the World Health Organization has estimated that one in 160 children have an Autism Spectrum Disorder worldwide[1]. Early detection of Autism, along with treatment, is crucial to minimize some of the difficulties and symptoms that people with Autism face[2]. Symptoms of Autism Spectrum Disorder are normally identified based on psychological criteria[3]. Specialists use techniques such as behaivoral observation reports, questionaires, and a review of the child's cognitive ability to detect and diagose Autism in children.

Many researchers believe that there is a correlation between facial morphology and Autism Spectrum Disorder, and that people with Autism have distinct facial features that can be used to detect their Autism Spectrum Disorder[4]. Human faces encode important markers that can be used to detect Autism Spectrum Disorder by analyzing facial features, eye contact,facial movements, and more[5]. Scientists found that children diagnosed with Autism share common facial feature distinctions from children who are not diagnosed with Autism[6]. Some of these facial features are wide-set eyes, short middle region of the face, and a broad upper face. Figure 1 provides an example of the facial feature differences between a child with Autism and a child without.

Figure 1: Image of Child with Autism (left) and Child with no Autism (right)[7].

Due to the distinct features of Autistic individuals, we believe that it is necessary to explore the possiblities of using a facial analysis to detect Autism in children, using Artificial Intelligence (AI). Many researchers have attempted to explore the possibility of using various novel algorithms to detect and diagnose children, adolescents, and adults with Autism[2]. Previous research has been done to determine if Autism Spectrum Disorder can be detected in children by analyzing a facial image[7]. The author of this research collected approximately 1500 facial images of children with Autism from websites and Facebook pages associated with Autism. The facial images of non-autistic children were randomly downloaded from online and cropped.The author aimed to provide a first level screening for autism diagnosis, whereby parents could submit an image of their child and in return recieve a probability of the potential of Autism, without cost.

To contribute to this previous research[7], this project will propose a model that can be used to detect the presence of Autism in children based on a facial image analysis. A deep learning algorithm will be used to develop an inexpensive, accurate, and effective method to detect Autism in children. This project implements and utilizes a Convolutional Neural Network (CNN) classifier to explore the possibility of using a facial image analysis to detect Autism in children, with an accuracy of 95% or higher. Most of the coding used for this CNN model was obtained from the Kaggle dataset and was done by Fran Valuch[8]. We made changes to some parts of this code, which will be discussed further in this project. The goal of this project is not to diagnose Autism, but to explore the possibility of detecting Autism at its early stage, using a facial image analysis.

2. Related Work

Previous work exists on the use of artificial intelligence to detect Autism in children using a facial image. Most of this previous work used the Autism kaggle dataset[7], which was also used for this project. One study utilized MobileNet followed by two dense layers in order to perform deep learning on the dataset[6]. MobileNet was used because of its ability to compute outputs much faster, as it can reduce both computation and model size. The first layer was dedicated to distribution, and allowed customisation of weights to input into the second dense layer. The second dense layer allowed for classification. The architecture of this algorithm is shown below in Figure 2.

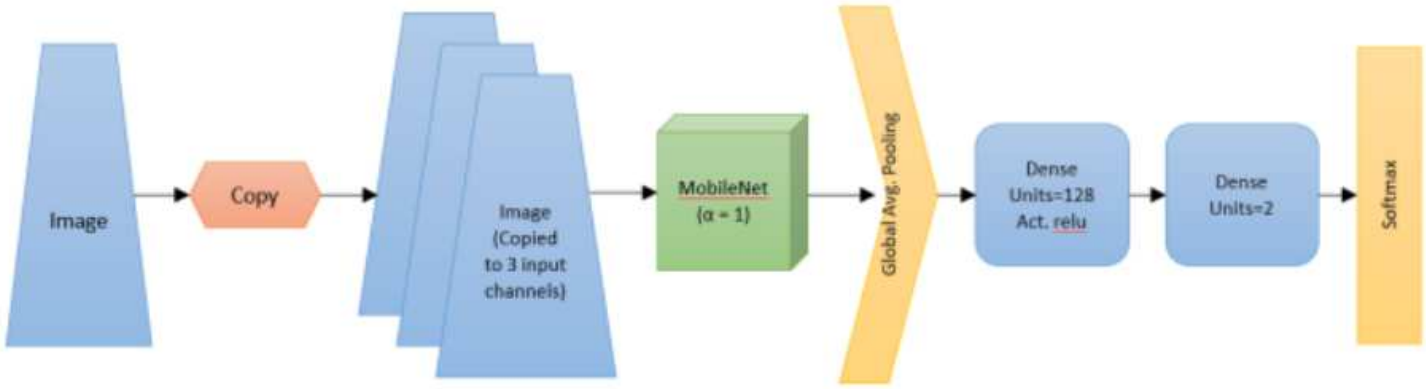


Figure 2: Algorithm Architecture using MobileNet[6].

Training of this model completed after fifteen epochs, which resulted in a test accuracy of 94.64%. In this project we utilize a classic Convolutional Neural Network model using tensorflow. This will be done in hopes of obtaining a test accuracy of 95% or higher.

3. Dataset

The dataset used for this project was obtained from Kaggle[7]. This dataset contained approximately 1500 facial images of children with Autism that were obtained from websites and Facebook pages associated with Autism. The facial images of non-autistic children were randomly downloaded from online. The pictures obtained were not of the best quality or consistency with respect to the facial alignment. Therefore, the author developed a python program to automatically crop the images to include only the extent possible for a facial image. These images consist of male and female children that are of different races and range from around ages two to fourteen.

This project uses version 12 of this dataset, which is the latest version. The dataset consists of three directories labled test, train, and valid, along with a CSV file. The training set is labeled as train, and consists of 'Autistic' and 'Non-Autistic' subdirectories. These subdirectories contain 1269 images of autistic and 1269 images of non-autistic children respectively. The validation set located in the valid directory are separated into 'Autistic' and 'Non-autistic' subdirectories. These subdirectories also contain 100 images of autistic and 100 images of non-autistic children respectively. The testing set located in the test directory is divided into 100 images of autistic children and 100 images of non-autistic children. All of the images provided in this dataset are in 224 X 224 X 3, jpg format. Table 1 provides a summary of the content in the dataset.

Table 1: Summary Table of Dataset.

Dataset Directory	Organization/ Composition	Percentage/ Overall data Composition
Train	1269 autistic 1269 non-autistic	86%
Test	100 autistic 100 non-autistic	6.8%
Validation	100 autistic 100 non-autistic	6.8%

4. Proposed Methodology

Convolutional Neural Network (CNN)

This project utilizes a Convolution Neural Network (CNN) to develop a program that can be used to detect the presence of Autism in children from a facial image analysis. If successful this program can be used an inexpensive method to detect Autism in children at its early stages. We believed that a CNN model would be the best way create this program because of its little dependence on preprocessing data. A Convolutional Neural Network was also used becuase of its ability to take in an image and assign importance to, and identify diferent objects within the image. CNN also has very high accuracy when dealing with image recognition. The dataset used contains 1269 training images that were used to train and test this Convolution Neural Network model. The architecture of this model can be seen in Figure 3.

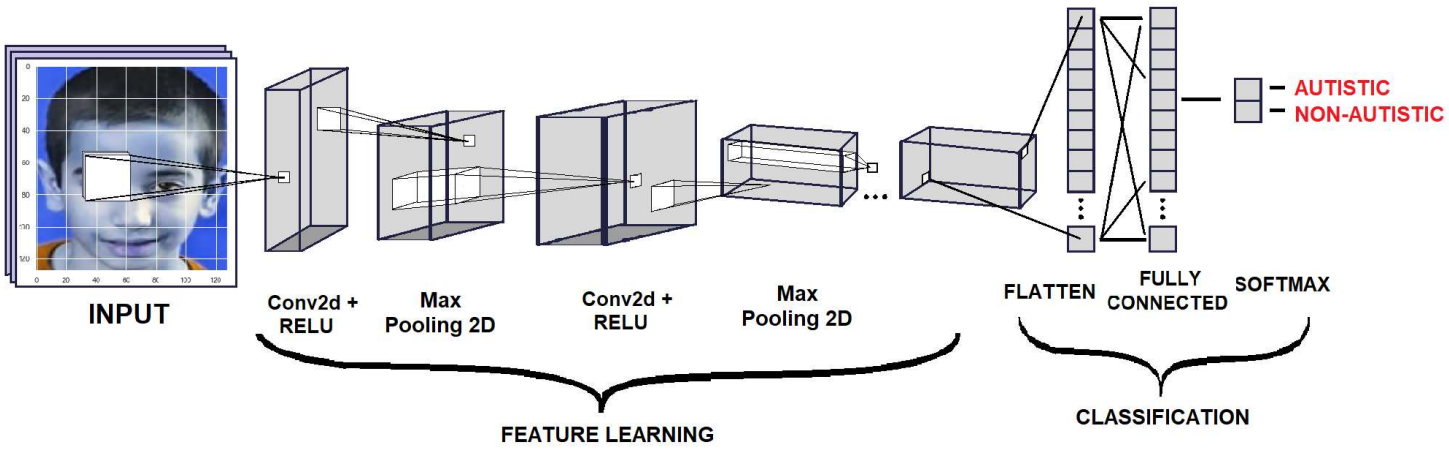


Figure 3: Architecture of utilized Convolutional Neural Network Model.

5. Results

The results of this project is estimated by affectability and accuracy by utilizing the Confusion Matrix CNN. The results also rely on how correct and precise the model was trained. This model was created to explore the possibility of detecting Autism in children at its early stage, using a facial image analysis. A Convolutional Neural Network classifier was used to create this model. For this CNN model we utilized max pooling and Rectified Linear Unit (ReLU), with two epochs. This resulted in an accuracy of 71%. These results can be seen below in Figure 4. Figure 5 displays some of the images that were classified and labeled correctly (right) and the others that were labeled incorrectly (left).

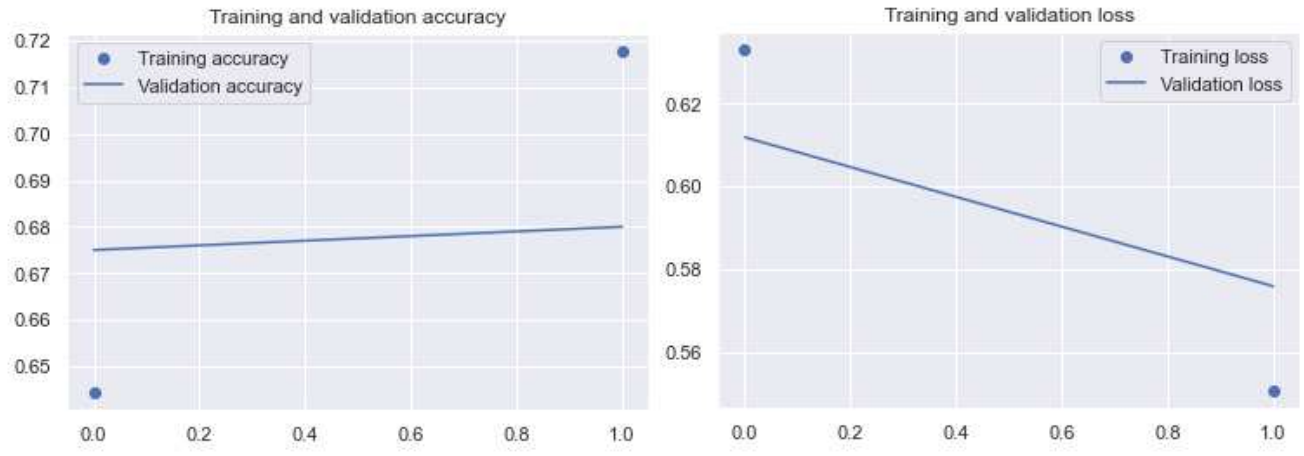


Figure 4: Results after Execution.

validation loss: 57% - validation accuracy: 68% - training loss: 55% - training accuracy: 71%

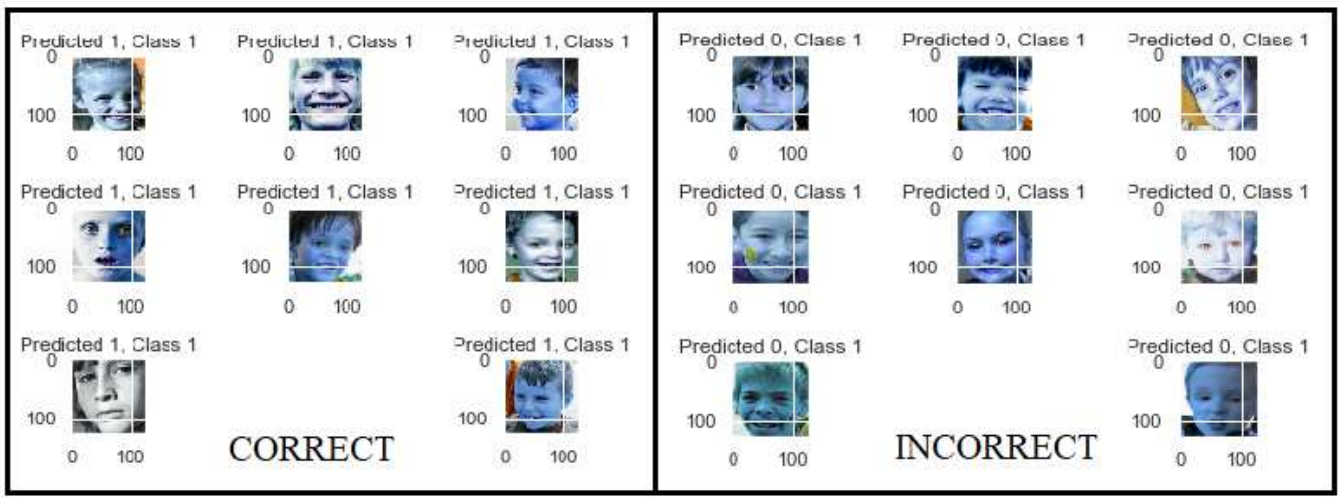


Figure 5: Correct Labels and Incorrect Labels.

6. Benchmark

Figure 6 shows the Confusion Matrix of the Convolutional Neural Network model used in this project. The Confusion Matrix displays a summary of the model's predicted results after its attempt to classify each image as either autistic or non-autistic. Out of the 200 images, 159 of the images were labled correctly and 41 of the images were labled incorrectly.

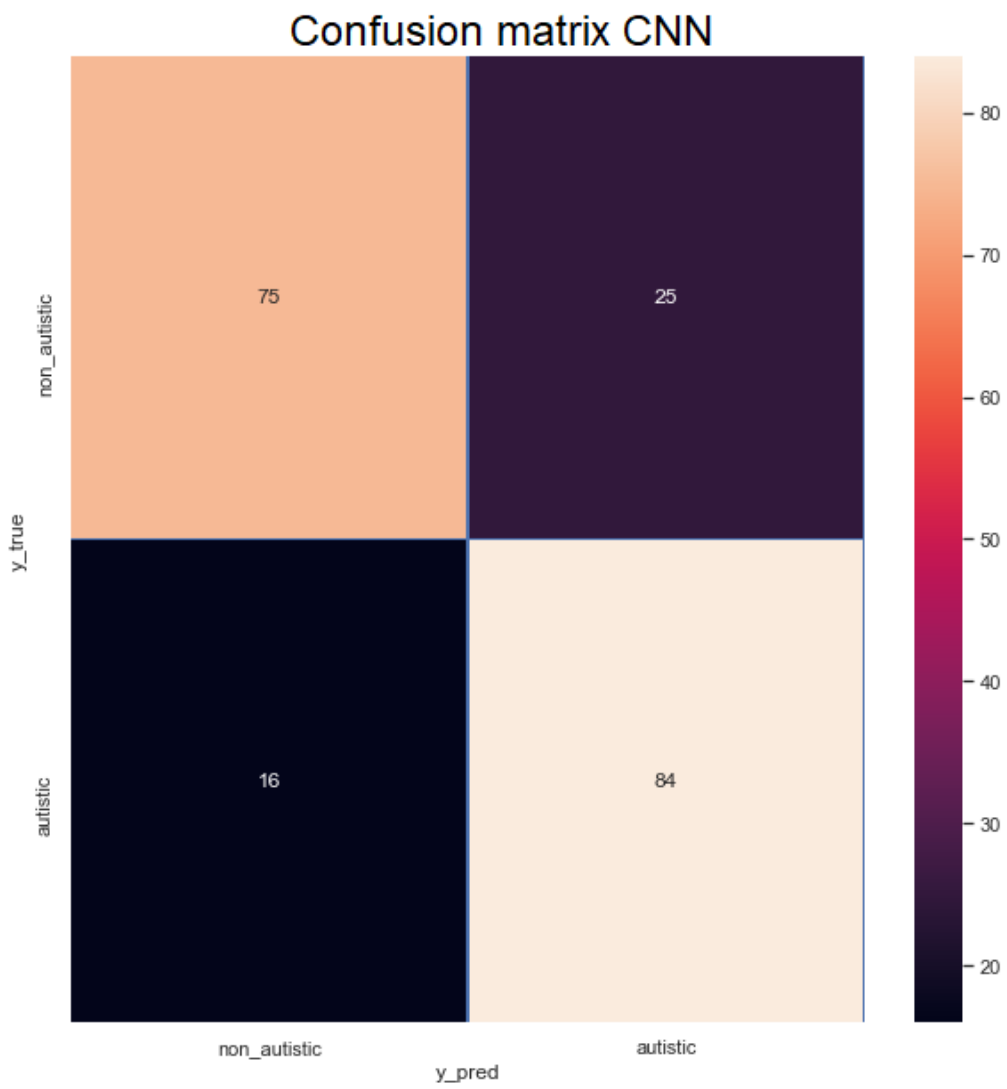


Figure 6: Confusion Matrix of the Convolutional Neural Network model.

Cloudmesh-common[9] was used to create a Stopwatch module, that was used to measure and study the training and testing time of the model. Table 2 shows the cloudmesh benchmark output.

Table 2: Cloudmesh Benchmark

Name	Status	Time	Sum	Start	tag	msg	Node	User	OS	Vers
Train	ok	3745.28	3745.28	2021-08-10 16:08:57			dab8db0489cd	collab	Linux	#1 S Sat J 5 09:5 PDT 202
Test	ok	2.088	2.088	2021-08-10 17:43:09			dab8db0489cd	collab	Linux	#1 S Sat J 5 09:5 PDT 202

7. Conclusions and Future Work

Autism Spectrum Disorder is a broad range of lifelong developmental and neurological disorders that is considered one of the most growing disorders in children. The World Health Organization has estimated that one in 160 children have an Autism Spectrum Disorder worldwide[1]. Techniques that are used by specialists to detect autism can be time consuming and inconvenient for some families. Considering these factors, finding effective and essential ways to detect Autism in children is a neccesity. The aim of this project was to create a model that would analyze facial images of children, and in return determine if the child is Autistic or not. This was done in hopes of receiving 95% accuracy or higher. After executing the model we received an accuracy of 71%.

As shown in the results section above, some of the pictures that were initially labeled as Autistic, were labeled incorrectly after running the model. This low accuracy rate could be improved if the CNN model is combined with other algorithms such as transfer learning and VGG-19. This low accuracy could also be improved by using a dataset that includes a wider variety and larger amount of images. We could also ensure that images in the dataset includes children that are of a wider age range. These improvements could possibly increase our chances of obtaining an accuracy of 95% or higher. When this model is improved and an accuracy of atleast 95% is achieved, furture work can be done to create a model that can be used for Autistic individuals outside of the dataset age range (2 - 14 years old).

8. Acknowledgments

The author of this project would like to express a vote of thanks to Yohn Jairo, Carlos Theran, and Dr. Gregor von Laszewski for their encouragement and guidance throughout this project. A special vote of thanks also goes to Florida A&M University for funding this wonderful research program. The completion of this project could not have been possible without their support.

9. References

1. World Health Organization. 2021. Autism spectrum disorders, [Online resource] https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders ↩
2. Raj, S., and Masood, S., 2020. Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques, [Online resource https://reader.elsevier.com/reader/sd/pii/S1877050920308656?token=D9747D2397E831563D1F58D80697D9016C30AAC6074638AA926D06E86426CE4CBF7932313AD5C3504440AFE0112F3868&originRegion=us-east-1&originCreation=20210704171932 ↩
3. Khodatars, M., Shoeibi, A., Ghassemi, N., Jafari, M., Khadem, A., Sadeghi, D., Moridian, P., Hussain, S., Alizadehsani, R., Zare, A., Khosravi, A., Nahavandi, S., Acharya, U. R., and Berk, M., 2020. Deep Learning for Neuroimaging-based Diagnosis and Rehabilitation of Autism Spectrum Disorder: A Review. [Online resource] https://arxiv.org/pdf/2007.01285.pdf ↩
4. Musser, M., 2020. Detecting Autism Spectrum Disorder in Children using Computer Vision, Adapting facial recognition models to detect Autism Spectrum Disorder. [Online resource] https://towardsdatascience.com/detecting-autism-spectrum-disorder-in-children-with-computer-vision-8abd7fc9b40a ↩
5. Akter, T., Ali, M. H., Khan, I., Satu, S., Uddin, Jamal., Alyami, S. A., Ali, S., Azad, A., and Moni, M. A., 2021. Improved Transfer-Learning-Based Facial Recognition Framework to Detect Autistic Children at an Early Stage. [Online resource] https://www.mdpi.com/2076-3425/11/6/734 ↩
6. Beary, M., Hadsell, A., Messersmith, R., Hosseini, M., 2020. Diagnosis of Autism in Children using Facial Analysis and Deep Learning. [Online resource] https://arxiv.org/ftp/arxiv/papers/2008/2008.02890.pdf ↩
7. Piosenka, G., 2020. Detect Autism from a facial image. [Online resource] https://www.kaggle.com/gpiosenka/autistic-children-data-set-traintestvalidate?select=autism.csv ↩
8. Valuch, F., 2021. Easy Autism Detection with TF.[Online resource] https://www.kaggle.com/franvaluch/easy-autism-detection-with-tf/comments ↩
9. Gregor von Laszewski, Cloudmesh StopWatch and Benchmark - Cloudmesh-Common, [GitHub] https://github.com/cloudmesh/cloudmesh-common ↩