C951

Task 3

Machine Learning Project Proposal

01/30/2022

A. Project Overview.....................................................................................................................pg 2

A.1. Organizational Need.............................................................................................................pg 2

A.2. Context and Background......................................................................................................pg 2

A.3. Outside Works Review........................................................................................................pg 3

A.4. Solution Summary............................................................................................................... pg 4

A.5. Machine Learning Benefits...................................................................................................pg 4

B. Machine Learning Project Design............................................................................................pg 4

B.1. Scope ...................................................................................................................................pg 4

B.2. Goals, Objectives, and Deliverables ..................................................................................pg 4-5

B.3. Standard Methodology .......................................................................................................pg 5

B.4. Projected Timeline.............................................................................................................pg 5-6

B.5. Resources and Costs............................................................................................................ pg 6

B.6. Evaluation Criteria .............................................................................................................. pg 6

C. Machine Learning Solution Design..........................................................................................pg 7

C.1. Hypothesis........................................................................................................................... pg 7

C.2. Selected Algorithm.............................................................................................................. pg 7

C.2.a Algorithm Justification....................................................................................................... pg 7

C.2.b. Algorithm Advantage ....................................................................................................... pg 7

C.2.c. Algorithm Limitation ........................................................................................................ pg 7

C.3. Tools and Environment ...................................................................................................... pg 7

C.4. Performance Measurement............................................................................................... pg 8

D. Description of Data Sets ....................................................................................................... pg 8

D.1. Data Source .......................................................................................................................pg 8

D.2. Data Collection Method...................................................................................................... pg 8

D.2.a.i. Data Collection Method Advantage............................................................................... pg 8

D.2.a.ii. Data Collection Method Limitation .............................................................................. pg 8

D.3. Data Normalization............................................................................................................. pg 9

D.4. Data Security....................................................................................................................... pg 9 Sources…................................................................................................................................... pg 10

1. **Project Overview**

This proposal describes providing a machine learning project that will analyze images and be able to spot patterns in skin conditions. Currently, the Healthcare Softworks company that provides medical applications and support to smaller medical facilities around the US has an older solution that has been in use for many years. However, this solution is based off hand-picked data that must be manually selected for matches and does not provide much in the sense of expansion. The new proposed system would implement a Convolutional Neural Network to automate the identification process and produce results faster than a manual solution.

**1. Organizational Need**

As it stands now, the Healthcare Softworks company provides a software solution in the design of an application, website, and partial applications that are implemented into third party systems and care providers. Part of this system is to assist with diagnosing skincare conditions and diseases, but it is a fully manual process that is reviewed by a team in order to produce accurate results. The data is reviewed, matched with internal references for a certain condition, and then sent back once it has been seen as a possible match. In the past, this system was acceptable in the sense that the company would only maintain a smaller client base that utilized this feature. Implementing a machine learning solution that can identify an image and produce a result within seconds would allow for growth of this system.

**2. Context and Background**

The context surrounding the purpose of this proposal would be to fully replace a slow and outdated existing solution with a smarter, faster one. This would allow expansion from only the care provider side to the user side with support for providing information over the possible skin condition they have, or advice on whether to schedule to see a care provider. All information would be privately retained on existing internal servers for use only by this solution to minimize HIPPA risks and provide a better benefit to our users.

**3. Outside Works**

Many large applications and platforms implement some form of image analyzation in one way or another. Easily recognizable ones would be Google with their image search and searching by a “related image”, or with Translate and being able to identify characters of a language to be translated into your language of choice.

1. The first work that relates to what we aim to implement is by Google. This is as stated by Google, a “deep learning approach to assessing skin conditions.” It aims to reach certification by dermatologists and assist with diagnosis and recognition. This appears based on a Python-implemented model using TensorFlow, similar to the presented solution in this document. However, this solution has not achieved FDA certification and is not available within the United States, where Healthcare Softworks is headquartered and supports all their clients. (Peggy Bui, 2021) (<https://blog.google/technology/health/ai-dermatology-preview-io-2021/>)
2. A second work was proposed using Google’s EfficientNet-B4 CNN algorithm that was developed specifically for psoriasis, eczema, and atopic dermatitis. In its current state, it achieved 89.46% accuracy for psoriasis and 92.57% accuracy for eczema and atopic dermatitis. This was proposed and developed in mainland China to test with provided images from the Xiangya Hospital at the Central South University in China. The training network for the algorithm used ImageNet for the source images before being applied to the local dataset. While this implementation looks promising and relates similarly to the work we aim to achieve, what is proposed in this document would be mainly for internal use and licensing by the company. (Wu et al., 2020) (<https://pubmed.ncbi.nlm.nih.gov/32566608/>)
3. The third and final work is another CNN that was trained specifically for use with psoriasis and has a claimed accuracy that is on par with 230 dermatologists. This proposed model appears to also be implementing the EfficientNet-B4 algorithm but was instead trained by using 7033 images from 1166 patients provided by the Peking Union Medical College in China. Regarding the results, 90 images were used to compare with dermatologist diagnosis. Perceived results were listed as “generally comparable performance to the average level of dermatologists.” In relation to a business need, this is like the second work, where it is for now only a model that has been proposed and tested. (Yang et al., 2021) (<https://www.sciencedirect.com/science/article/abs/pii/S0010482521007186?via%3Dihub>)

In conclusion, each work provided above presented a machine learning application very similar to what is proposed in this document and has achieved a certain level of accuracy that has been very promising with limited testing. What we aim to achieve is a level of accuracy comparable to if not greater than the listed works and implement a method of re-training the model at set intervals to measure accuracy.

**4. Solution Summary**

This proposed solution aims to reduce the time needed in order to diagnose or provide more information and advice over a skin condition. It will use machine learning and image analyzation in order to better adapt to various types of images with ever-growing accuracy. This will then allow expansion of the solution from current care providers out to end-users due to results being available within seconds.

**5. Machine Learning Benefits**

The benefits of machine learning for this project are that using image recognition, we will be able to not only identify a skin condition, but also be able to continuously have the algorithm improving by re-training at certain points using the newer, identified pictures. The main benefit and purpose of this proposal is to reduce the amount of time to achieve a diagnosis as well as implement an automated solution that is maintained at a minimum level of accuracy.

1. **Machine Learning Project Design**
2. **Scope**

* **In-Scope:** Create a machine learning system that can identify images within a reasonable certainty and a minimum level of accuracy
* **In-Scope:** Develop a training model that can be “fed” after trained with confirmed images of certain conditions to improve accuracy
* **In-Scope:** Utilize company data and retain all data in company servers, as well as maintain a certain level of privacy due to potential HIPPA regulations
* **In-Scope:** Develop an addition to the application and web portal to be utilized by several smart devices with a camera, or a web-enabled device that can upload a photo
* **Out-of-Scope:** Optimize application for use outside of Healthcare Softworks’ main application

1. **Goals, Objectives, and Deliverables**

**Goals:**

The main goal of the project would be to create an efficient machine learning system/application that is able to process images, recognize a certain skin condition, and either produce a diagnosis or provide advice on next steps to receive proper care. A secondary goal once the main goal has been established will be to further optimize the processing time of the algorithm to reduce the amount of time taken to produce results.

**Objectives:**

Main objectives for this project would be first establishing an algorithm for use, then creating the framework for use specifically with the company’s data types. Afterwards, it would be optimizing the application for use across multiple systems and a range of devices. In addition, a “portal” or user interface will be designed for front-end use with the main processing load being handled with the company’s existing servers. Finally, it will have a training method that can be re-trained with newer images to increase accuracy.

**Deliverables:**

The firstdeliverable would be the system and/or application that is able to utilize image processing and recognition for identifying skin conditions. A second would be the final algorithm and training system, modified to allow use of the company’s specific data types with storage and use within the company’s system. Thirdly, the addition to the application and web portal

1. **Standard Methodology**

The development process will follow the SEMMA methodology:

* **Sample:** In the sample phase, we will want to take known good photos of different types of skin conditions for organization and use by the machine learning system. From there, we can categorize between related and unrelated types.
* **Explore:** In this stage, we will want to draw similarities between the data and images that we already have in order to further the accuracy process.
* **Modify:** In this stage, we will take the categorized photos and data and organize it into specific data types for use in the flow of the application.
* **Model:** In this stage, we will take the data from the modification stage and pass it through to the final stage, such as the final produced result from a “raw” image into being categorized.
* **Assess:** In this stage, the final image will be examined to see how accurate the result is based on known good photos and other results.

1. **Projected Timeline**

* **1/31/22:** The stated problem, projected solution, requirements and costs are proposed.
* **2/1/22 - 2/4/22:** The project and teams are approved for development and implementation.
* **2/7/22 - 2/25/22:** The model is being worked on and is then presented with a working prototype.
* **2/28/22 - 3/11/22**: Further optimizations and testing has concluded on the final model, and it has been submitted for review.
* **3/14/22 - 4/1/22:** Presentation for implementation of the model and UX model begins, it is tested and revised.
* **04/04/2022 – 04/15/2022:** The model is finalized and is slowly integrated into the existing system.

**Sprint Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | 02/07/2022 | 02/25/2022 | Data is collected, cleaned, organized and tested. The algorithm is developed and tested to reach a certain level of accuracy. |
| 2 | 02/28/2022 | 03/11/2022 | Further optimizations are implemented, and the final algorithm is developed. |
| 3 | 03/14/2022 | 04/01/2022 | The UX model is started, tested, and revised. |
| 4 | 04/04/2022 | 04/15/2022 | Final touches are placed on the algorithm and model, and the design is integrated into the existing application. |

1. **Resources and Cost**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
|  |  |  |
| Existing internal network/servers | For use of hosting the main machine learning system | $10,000 |
| Internal Development Team | For use of creating main application. 3 team members will dedicate approximately 30 hrs/week to the project | $12,000 |
| Online analytics and visualization tools | For use in creating the machine learning system | $1500 |
| UX Designer | For creating the UI in the application and web portal, roughly 60-70 hours of work overall with revisioning | $2500 |
|  | **Total** | $26,000 |

1. **Evaluation Criteria**

The criteria that will primarily be used to identify whether this project is successful are:

* Accuracy of diagnosis
* Efficiency of program and algorithm
* Ease of use by end-users
* Ease of implementation into possible third-party applications

We will need to deliver a diagnosis that is accurate at least 95% of the time, that runs within a reasonable timeframe (at most 5-10 seconds for each image) and allows us to propose this solution to possible other business for use in their systems. We will also maintain the overall same ease of use provided by the existing application and web portal.

1. **Machine Learning Solution Design**
2. **Hypothesis**

The hypothesis of this project is that if there are cases where someone is unsure what possible skin condition, they or another may have, then we will provide a system that recognizes the type of condition based on AI and machine learning with image processing. The specific purpose of this proposal’s hypothesis is to present a solution to Healthcare Softworks’ older solution that only allows for so much expansion, and to implement a machine learning solution that will analyze images to provide a much faster result than what was possible before, leading to greater expansion opportunities.

1. **Machine Learning Algorithm**

The algorithm of choice will be an unsupervised algorithm based on the Convolutional Neural Network. This was chosen based on the overall accuracy that can be achieved with image processing.

1. **Justifications**

This algorithm was chosen mostly since it has been compared “to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex.” It has also been known to require less pre-processing, but it will require more training to present a solution as compared to other methods. (Saha, 2018)

1. **Advantages**

A main advantage is that a CNN can recognize multiple layers of an image as compared to “flattening” an image for use by other methods. This allows the CNN “to reduce images into a form which is easier to process, without losing features which are critical for getting a good prediction.” Essentially, we’re wanting to create an algorithm and process that can examine features and diagnose within a small margin of error. (Saha, 2018)

1. **Limitations**

Due to a CNN requiring a tensor-based image/dataset, this can have rather heavy performance requirements. In order to solve this, we will host all processing on the company’s internal network and provide the results after processing has finished. A CNN also requires a large amount of data in order to be trained properly, to which we will provide a large dataset to achieve the level of accuracy we aim to maintain.

1. **Tools and Environment**

For this project, we will primarily be using Windows for development, for deployment across iOS, Android, Windows and Mac devices. The IDE in use will be PyCharm with Python, for use with the Keras libraries for developing the CNN network. PyCharm has a multitude of features that make using Python much easier, and Keras primarily supports faster development with support for several devices.

1. **Performance Measurement**

The performance metrics that this project will adhere to are accuracy of the result as well as the efficiency. Due to being health-related, we will want the result to be delivered in a timely manner, or within a matter of seconds. Secondly, we will want to make sure the accuracy of the result is above 95%, and possibly having a small margin of error in the testing phases. As a base, we will want to establish a known set of good values or images that can be compared to with the results of the algorithm. In order to achieve an accurate diagnosis, we will have a team of dermatologists and physicians within Healthcare Softworks provide anonymous feedback on the results to compare with their diagnoses.

To specifically measure the results of the CNN algorithm, we will compare with similar image recognition solutions if available for use in the diagnosis and speed of result. Internal testing further in development or post-deployment could be compared to previous results, in order to maintain the level of accuracy we want.

1. **Description of Data Sets**
2. **Data Source**

Initial data collection will come from the images that the company already has, possible public sources, plus any images that are taken during the development process for testing. Afterwards, pictures that are relevant and can provide further information will be utilized to provide more accurate results.

1. **Data Collection Method**

Specifically, the data collection will be primarily focused on the characteristics in an image of the type of condition, based on patterns. As mentioned above, a public source such as ImageNet or Google’s Open Images could be used for a large dataset in order to train the CNN we are wanting to implement. The data collected by the application will either be stored by the company in an internal database if it is a known good picture of a condition for re-use or discarded if no useful information can be assessed.

1. **Data Collection Method Advantages**

With the method of providing a large dataset to the CNN, we are mitigating the downside of not having a large bank of images for training the algorithm that is being implemented. This should lead to a more accurate algorithm that will hopefully increase in accuracy over time after being trained with more accurate pictures.

1. **Data Collection Limitation**

A limitation of this dataset is that a large dataset is needed to train the CNN, and it will take a lengthy amount of time to make sure the algorithm is presenting accurate results. Other methods would allow us to present the training model with hand-picked images.

1. **Quality and Completeness of Data**

With training a CNN using images, we could help test with a smaller batch of images for the use of testing the algorithm’s efficacy. In the case of mitigating pulling bad information from an image in the testing phase, an image can be set aside or flagged for possible deletion if no useful information can be pulled in a reasonable amount of time. In the case of missing data, a potential solution could be a reference image either pulled from an outside source for the sole purpose of training the model, with the right licensing and agreements of use of the image.

1. **Precautions for Sensitive Data**

Since the data we will be working with will most likely be images of skin and potential body parts, certain restrictions will have to be placed on the access and storage of the images. We will need to comply with all HIPPA regulations, as well as data privacy regulations worldwide if we are looking to deploy this solution outside of the United States. If any information has patient data tied to it, it will be retained in a secure location. Due to Healthcare Softworks already providing an application for use by smaller medical facilities, most of the organizational framework is in place.

As a baseline for security over this information, we will most likely want to implement a sub-domain from the company’s primary domain for solely hosting and providing access to this application and its data. In this sub-domain, the database that will retain this information will be fully encrypted, adhere to the most updated security standards, and any data that is not needed to train the algorithm will be purged within a certain timeframe. Only authorized individuals will have access to the database, and the sub-domain will remain hidden to anyone outside of that.

**Sources**

Saha, S. (2018, December 17). *A comprehensive guide to Convolutional Neural Networks - the eli5 way*. Medium. Retrieved January 30, 2022, from <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>

Bui, P., & Liu, Y. (n.d.). *Using AI to help find answers to common skin conditions*. Retrieved January 30, 2022, from <https://blog.google/technology/health/ai-dermatology-preview-io-2021/>)

Wu, H., Yin, H., Chen, H., Sun, M., Liu, X., Yu, Y., Tang, Y., Long, H., Zhang, B., Zhang, J., Zhou, Y., Li, Y., Zhang, G., Zhang, P., Zhan, Y., Liao, J., Luo, S., Xiao, R., Su, Y., Zhao, J., … Lu, Q. (2020). A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. Annals of translational medicine, 8(9), 581. Retrieved January 30, 2022, from <https://doi.org/10.21037/atm.2020.04.39>

Yang, Y., Wang, J., Xie, F., Liu, J., Shu, C., Wang, Y., Zheng, Y., & Zhang, H. (2021, October 6). *A convolutional neural network trained with dermoscopic images of psoriasis performed on par with 230 dermatologists*. Computers in Biology and Medicine. Retrieved February 4, 2022, from <https://www.sciencedirect.com/science/article/abs/pii/S0010482521007186?via%3Dihub>

Pokhrel, S. (2020, February 19). *Collecting data for custom object detection*. Medium. Retrieved January 30, 2022, from <https://towardsdatascience.com/collecting-data-for-custom-object-detection-e7d888c1469b>