The Potential of Artificial Intelligence to Identify Cancer and Aid Research

Author(s) Wilson B. Peguero Rosario

Master of Science in Data Science Capstone Project Proposal

Grand Canyon University

Instructor: Professor TBA

Revision: 5

Date: November 9, 2022

**ABSTRACT**

Artificial Intelligence (AI), the science and engineering of creating intelligent computer programs that can think rationally and without human interaction. This field has rapidly expanded within the last five years, causing massive changes within all industries, from Medicinal, civil, finance, to even government agencies that currently leverage AI to minimize long and arduous tasks to simple calculations.  These AI models can think on a much faster level than the human mind can process and can make assumptions from existing data and even predict what may occur in the future. Because of these capabilities, AI is one of the most invested skills within the field of medicine. With the ability to classify certain images based on different stages of cancer, a doctor may be able to pick out early signs that can save a patient’s life. With the ability to predict the reaction between chemical combinations and the human body, AI can assist pharmaceuticals in non-destructive drug discovery. The potential of AI, specifically in cancer research, cannot be ignored any longer.

**PROJECT SYNOPSIS**

[This section is to be written only when the entire project is complete. If not applicable yet, delete this section. The project synopsis should summarize the entire project, but briefly provide a broader context of the major sections of the project (design, development, implementation, testing and overall functionality). Once complete, remove this section and update the abstract accordingly.]

|  |
| --- |
| **HISTORY AND SIGN-OFF SHEET** |

**Change Record**

|  |  |  |
| --- | --- | --- |
| **Date** | **Author** | **Revision Notes** |
| 5/30/2022 | Wilson Peguero Rosario | Project Overview and Abstract |
| 6/11/2022 | Wilson Peguero Rosario | Initial draft for review/discussion |
| 6/22/2022 | Wilson Peguero Rosario | Final Draft |
| 11/3/2022 | Wilson Peguero Rosario | Updated objectives to represent DICOM files. |
| 11/9/2022 | Wilson Peguero Rosario | * Updated the project scope to use image data AND header file data * Included assumption about header files * Included overfitting and sample size to risk table * Reworked the schedule hours and included web design in work breakdown table |

|  |
| --- |
| **Overall Instructor Feedback/Comments** |

|  |
| --- |
| **Overall Instructor Feedback/Comments** |

**Integrated Instructor Feedback into Project Documentation**

Yes  No

**Project Approval**

*<Insert name of instructor here>*

**TABLE OF CONTENTS**

Project Overview and Project Objectives 4

Project Scope 5

Project Success Measures 6

Project High-Level Solution 7

Project Controls 8

Project Cost and Schedule 9

Appendix A – References 10

Appendix B – Copyright Compliance 11

Project Overview and Project Objectives

**State the Problem and Background**

Cancer has existed since the ancient Egyptians (Husaini et al., 2020). As technology advanced over the eras, so did our knowledge of the potential cancer have as a disease. Multiple ways of identifying cancer have been utilized over time. These modalities are called Mammogram, Ultrasound, CT scan, and Thermography (Devita et al., 2016). These imaging modalities can be supplemented with machine learning algorithms to detect early signs of cancer (Cardoso et al., 2020), and the correct form of treatment (Hadjiyski, 2020). Multiple cancer treatments have been developed through drug discovery and gene sequencing. Machine learning can supplement cancer treatments through non-destructive drug discovery, which will save resources and materials used to develop the medicine up to the clinical trial stage. Through AI-powered gene sequencing, cancer treatment can be customized on a patient level to provide efficient treatment while avoiding other more harmful treatments (Kulkarni et al., 2019). AI-powered gene sequencing can go as far as working on a genetic level to suppress certain genome sequences that can cause cancer to develop in the patient (Iyer et al., 2019).

**Christian Worldview**

Cancer is the second cause of death in the United States. This is a painful experience both for the patient and the family that can be traumatizing for those who are misdiagnosed and for those who do not require special treatment as the tumors are not malignant but rather benign. As a Christian, I am meant to associate and empathize with the pain of my fellow man or woman. To attempt to help them in any way that I can is something that God has called us to do. As someone who had a family member to die from breast cancer, I find that largest factor to cause this family members death was not due to the doctors, but rather the limitations that they have from having to handle multiple people suffering from different ailments and proper customized treatment. I can’t bring back my family member back to life, but I can help others from suffering the same thing that my family did.

**Project Objectives**

* Gather data set of DICOM files with images that are labeled as either benign or malignant (these images will contain tumors).
  + The DICOM file will contain textual, numerical, and categorical data within the header file and image data.
* Develop a base model to use as a baseline for comparison.
* Develop model class with potential to extract features from DICOM files.
  + Learn ideal features to extract from DICOM file.
* Train model based on gathered data and use some of the gather data as a test set.

**Challenges**

Data Gathering will be the biggest challenge. Although there are multiple medical image data sets, majority of the data sets are not reliable enough to surpass current misdiagnosis rate (95%). There is also the case that the images are using different modalities (i.e., X-ray, MRI, Ultrasound, etc.).

**Benefits and Opportunities**

This project may be utilized to start a biomedical software company or to use as project for hire in a medical institution.

Project Scope

Develop a machine learning model that can classify patients on whether a tumor is malignant or benign using DICOM file data with the same accuracy as the annual misdiagnosis rate.

In Scope Features:

* Model with ability to classify based on medical image data AND patient metadata.
* Able to receive raw image, and patient metadata to classify the patient.

Out of Scope Features:

* Highlights regions of interest.
* Generates details regarding the image.

**Table <Insert #>. Stakeholders**

|  |  |  |
| --- | --- | --- |
| Stakeholder Name | Role(s) | Responsibilities |
| Self | Data Scientist | Gathering data; developing model; training model; test model; interpret results |
| Operator | Data Analyst | Developing the model; Training and testing the model |
| User | Radiologist | Interpreting the results |
| Maintainer | Data Engineer | Gathering the data |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Work Breakdown Structure | | | | | | | | | | |
| ID | Task | Dependencies | Status | Effort Hours | Cost | Start Date | Planned Completion | Estimate to Completion | Actual Completion | Resource |
| 1 | Collect Data | Data availability | Finished | 40 | Free |  | Fully Labeled Data Set | 7 Days | 9 days | Kaggle; UCI Machine Learning Repository; GitHub |
| 2 | Process Data | Data Availability | Finished | 20 | Free |  | Fully Filtered Data Set | 3 Days | 5 Days | Data Collected |
| 3 | Research Machine Learning Models | Journals with model designs | Finished | 40 | Free |  |  | 7 Days | 3 Days | Journal Articles |
| 4 | Design Model | Libraries available;  Programming languages;  Data type;  Research articles | Finished | 100 | Free |  | Completed Model Design | 15 Days | 5 Days | TensorFlow;  Python;  NIH; |
| 5 | Training & optimize hyperparameters | Libraries available;  Research Articles; Model Design | In Progress | 80 | Free |  | Optimized Model | 14 Days |  | TensorFlow;  Python;  NIH |
| 6 | Test Model | Data Collected | In Progress | 10 | Free |  | Fully Tested Model | 2 |  | TensorFlow;  Python;  NIH |
| 7 | Evaluate Model | Data Availability; Research Articles | In Progress | 20 | Free |  | Model with real life evaluation | 4 Days |  | TensorFlow;  Python |
| 8 | Develop Dashboard | Research articles; Model performance; Libraries available | Finished | 8 | Free |  |  | 1 Day | 5 Hours | Python; Dash library |

Project Success Measures

The metrics used to measure project success will be based on the model’s ability to label the medical images as either malignant or benign. The loss, accuracy, recall, and precision metrics will be used to compare with how well the model performs in the diagnosis aspect of this project. Other metrics, such as AUC will be used to visually explain the performance of the machine learning model.

|  |
| --- |
| Project Completion Criteria |
| 1. Accuracy matches the range provided within articles containing similar models (60%-90%) |
| 2. Accuracy of the model is equivalent or higher than the yearly correct diagnosis rate (95%) |
| 3. false positive rate of the model is minimal (5%-10%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Assumptions and Constraints | | | | | |
| ID | Description | Comments | Type | Status | Date Entered |
| 1 | The data sets from multiple hospitals are not significantly different from each other in terms of imaging modality | There are some parameters used to take MRI images and even different devices to take images of the tumors which may impact the model’s ability to classify tumors. There are also changes in resolutions related to the image which may also impact the model efficiency, these are assumed to have minimal impact. | Assumption |  | 6/11/2022 |
| 2 | The data sets will stem from at most 20 different hospitals |  | Constraint |  | 6/22/2022 |
| 3 | The images will be taken using at most three (3) imaging modalities (MRI, Computed Tomography, SPECT) | The three main imaging modalities used are Magnetic Resonance (MRI), X-rays (Computed Tomography), and Nuclear Medicine (SPECT) | Constraint |  | 6/22/2022 |
| 4 | All available header file data pertaining to the patient correlates with the malignancy of the tumor | Most of the information gathered regarding the patient pertains to risk factors of cancer. | Assumption |  | 11/5/2022 |

Project High-Level Solution

**Introduction**

Current Cancer treatment has reached its limits. Cancer screening is very effective at discovering cancer in its early stages, yet the mortality rate has not significantly changed because of increased cancer screening (Devita et al., 2016). The efficiency of early screening regarding mortality is heavily dependent on the treatments available for the types of cancer (Devita et al., 2016). There are too many factors leading to cancer that can be observed in a typical lab environment. Tobacco alone has enough of an effect to be considered as the cause of cancers located in the bladder, cervix, colon, and rectum, esophagus, kidney larynx, leukemia, liver, lung, oral cavity, and pharynx, pancreas, and stomach (Devita et al., 2016). Although current cancer treatments are limited, catching cancer at its earliest stages yields high survival rates (Kulkarni et al., 2019).  These factors cause the survival rate of cancer to stagnate at very low percentages, as cancer is not caught on time, as more cancerous devices spread, such as the popularity of the e-cigarette, or vaping, so do the chances of developing cancer. The highest probability of survival occurs in the earlier stages (Devita et al., 2016), creating a need for AI to better identify cancer.

**Solution**

The project can be split into three sections, data gathering/mining, data validation/analysis, and AI modeling. The first step in the procedure is to gather imaging data as well as medical test results from multiple sources (such as Kaggle, IEEE, government websites, associate organizations) to create a valid dataset that can be leveraged to prove the potential in AI through testing and comparisons with raw data. Prelabelled or filled in data would be required to use comparative analysis and observe the efficiency of the machine learning model to identify or fill in for missing labels or values. Once the data is gathered, the next step is to properly validate the data. This would include removing outliers, finding importance in features, identifying independent and dependent features, and minimize the dataset to its most useful and functional version. Then, the machine learning model will be developed based on research articles and peer-reviewed journals and trained, its hyperparameters will also be optimized to develop the best model possible. This will require majority of the project time as the development of the model will require fine tuning and the most efficient model used may require a combination from multiple algorithms. Lastly, the model will be tested extensively on pre-labeled test data. Comparative analysis is done to see the efficiency of the model against pre-labeled data.

Project Controls

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RISK MANAGEMENT | | | | |
| **Event Risk** | **Risk Probability**  **(high, medium, low)** | **Risk Impact** | **Risk Mitigation** | **Contingency Plan** |
| Data from different image modalities cause inaccuracy | medium | The model will inaccurately label medical images | adding hidden nodes to separate and classify images based on imaging modality. | Use only data sets using only one imaging modality. |
| Random noise | High | Minimal | Can be mitigated by filtering out noise | Remove data from the training set. |
| Overfitting | High | High | Add dropout layer(s) which will turn off the output of | Use smaller learning rate for the optimizer |
| Low Sample Size | Medium | Low | Include different data sets | Data augmentation techniques |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ISSUES LOG | | | | | | | | |
| **ID** | **Description** | **Project Impact** | **Action Plan/Resolution** | **Owner** | **Importance** | **Date Entered** | **Date to Review** | **Date Resolved** |
| 1 | Anonymized data limits model design | Will minimize the scope of the project to simple, and separate algorithms | Create separate models to evaluate different metrics or sets of data | Wilson Peguero Rosario | *High* | *6/22/2022* |  |  |
| 2 | Limitations of hardware equipment | Will extend the schedule past its targets. | Fine tune the training process to maximize model accuracy while minimize time spent training | Wilson Peguero Rosario | High | 6/22/2022 |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CHANGE CONTROL LOG | | | | | | | | | |
| **ID** | **Change Description** | **Priority** | **Originator** | **Date Entered** | **Date Assigned** | **Evaluator** | **Status** | **Date of Decision** | **Included in Rev. #** |
| 1 |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| ROLES AND RESPONSIBILITIES | | | |
| Name | Team | Project Role | Responsibility |
| Radiologist | N/A | Input on GUI | To comment and provide feedback on the model and its results. |
| Data Analyst | N/A | Analyze data | Review the feedback on the models and retrain based on comments |
| Data Engineer | N/A | Maintain Pipelines | Make improvements based on the comments that Data Analyst sends their way in regards to the data. |

Project Cost and Schedule

There will be no project cost as this project can be achieved using already purchased computer hardware. The kind of hardware that one possesses will only impact the time the model takes to be trained and fine-tuned.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Work Breakdown Structure | | | | | | | | | | |
| ID | Task | Dependencies | Status | Effort Hours | Cost | Start Date | Planned Completion | Estimate to Completion | Actual Completion | Resource |
| 1 | Collect Data | Data availability | Complete | 40 | Free | 9/30/2022 | Fully Labeled Data Set | 7 Days | 14 Days | Kaggle; UCI Machine Learning Repository; GitHub |
| 2 | Design Model | Libraries available;  Programming languages;  Data type;  Research articles | Complete | 100 | Free | 10/15/2022 | Completed Model Design | 15 Days | 30 Hours | TensorFlow;  Python;  NIH; |
| 3 | Training & optimize hyperparameters | Libraries available;  Research Articles; Model Design | Complete | 80 | Free | 11/01/2022 | Optimized Model | 14 Days | 2 Days | TensorFlow;  Python;  NIH |
| 4 | Design Dashboard | Libraries available | Complete | 10 | Free | 10/25/2022 | Dashboard | 1 Day | 5 Hours | Python; Dash; Plotly |
| 5 | Evaluate Model | Data Availability; Research Articles | In Progress | 20 | Free | 11/5/2022 | Model with real life evaluation | 4 Days |  | TensorFlow;  Python |

Appendix A – References

Bongiolatti, S., Gonfiotti, A., Borgianni, S., Crisci, R., Curcio, C., Voltolini, L., Alloisio, M., Amore, D., Ampollini, L., Andreetti, C., Argnani, D., Baietto, G., Bandiera, A., Benato, C., Benvenuti, M. R., Bertani, A., Bertolaccini, L., Bortolotti, L., Bottoni, E., & Breda, C. (2021). Post-operative outcomes and quality of life assessment after thoracoscopic lobectomy for Non-small-cell lung cancer in octogenarians: Analysis from a national database. *Surgical Oncology*, *37*, 101530. https://doi.org/10.1016/j.suronc.2021.101530

Data from a database regarding Video-Assisted Thoracic Surgery Lobectomy (VATS-L) for Non-Small-Cell Lung Cancer (NSCLC) was collected key metrics to observe the outcomes for the elderly. The authors were able to derive metrics from the dataset, such as “...30-day and 90-day postoperative mortality...[, ]any complication...[, ]complication rate...” and more to determine whether there is a significant risk that the elderly between 80-89 have when undergoing VATS-L for lung cancer through machine learning. Elsevier first received this article on 22 October 2020, revised on 27 December 2020, and accepted on 25 January 2021. The authors of this article utilize machine learning algorithms (logistic regression) to derive metrics related to complications that did not exist within the dataset, demonstrating the potential that AI possesses to supplement existing data and provide a new narrative on the risks of surgery for a specific age group.

Cardoso, M. J., Houssami, N., Pozzi, G., & Séroussi, B. (2020). Artificial intelligence (AI) in breast cancer care - Leveraging multidisciplinary skills to improve care. *The Breast*, 110–113. https://doi.org/10.1016/j.breast.2020.11.012

This article focuses on the potential of AI to automate repetitive tasks in the medicinal field and allows medical professionals to focus on providing a better patient experience. The article describes the different fields that AI can apply to, such as medical imaging, predictive diagnoses, and more. This article is a part of “The Breast” Journal, which is peer-reviewed by an editorial board comprising medical doctors and doctors from other professions (i.e. biomedical engineering). This article shows the potential that AI has in multiple fields related to medicine and can be utilized as a part of the introduction to AI in the medical field.

Devita, V. T., Lawrence, T. S., & Rosenberg, S. A. (2016). *Cancer : principles & practice of oncology. Prostate and other genitourinary cancers*. Wolters Kluwer. https://gcu-encore.iii.com/iii/encore/record/C\_\_Rb1321879\_\_SCancer%20\_\_P0%2C1\_\_Orightresult?lang=eng&suite=def

The book provides in-depth information about cancer, the medical techniques used to identify the type of cancer within the body, the risk factors that can increase or decrease the chances of developing cancer, as well as different cancer and their behavior. The book first starts on the genetic level, providing detail on cancer’s development from a genetic level and the highest risk factor. It then describes modern medicine used to combat cancer, to the factors that can either increase or decrease the probability of developing cancer. Finally, the book oversees the different techniques in medicine to combat cancer as well as provides detailed information about different cancer in existence. Being the tenth edition of the book, edited by three medical Doctors, released in 2016 by the organization Wolters Kluwer which has existed since 1836 makes this book a scholarly reference. This reference can describe cancer and provide broader information regarding the factors related to cancer used to develop machine learning algorithms.

Feng, Y., Yang, K., Sun, H., Liu, Y., Zhang, D., Zhao, Y., Shi, W., Lu, G., Zhang, Z., Jia, A., He, S., & Li, H. (2021). Value of preoperative gastroscopic carbon nanoparticles labeling in patients undergoing laparoscopic radical gastric cancer surgery. *Surgical Oncology*, *38*, 101628. https://doi.org/10.1016/j.suronc.2021.101628

This journal article provides insight on the effect that preoperative gastroscopic carbon nanoparticles labeling has on patients who are undergoing surgery for stomach-related cancer. The article then details how the data was gathered retrospectively. It was collected from a system containing electrical records and data from past surgeries done in their institution. It finally describes the results of it’s statistical analysis. This article was published by Elsevier, revised on 7 June 2021, and accepted on 11 June 2021 making this very recent journal article a scholarly reference. This article demonstrates the statistical aspect in determining the efficiency of certain methodologies which can be crucial to determine features within datasets that can potentially be utilized to develop machine learning algorithms.

Hadjiyski, N. (2020). Kidney Cancer Staging: Deep Learning Neural Network Based Approach. *2020 International Conference on E-Health and Bioengineering (EHB)*, pp. 1-4. https://doi.org/10.1109/ehb50910.2020.9280188

Article details the different stages of kidney cancer and how deep learning (a subset of AI) can determine the stage of kidney cancer to assist medical professionals in determining the correct treatment. Images extracted from CT scans of kidney cancer patients were taken and a deep learning algorithm was utilized to label, train, and test on the collected images. Then statistical analyses were used to determine the accuracy of the deep learning algorithm’s ability to classify the different stages of kidney cancer using images. This is an IEEE article published in the 2020 International Conference on e-Health and Bioengineering. These two factors (meaning that the article is backed by a large association related to all fields of science and that the article was published in a conference that may have required peer review) make this article a scholarly source. This is a clear example of the potential that AI has to aid cancer research.

Husaini, M. A. S. A., Habaebi, M. H., Hameed, S. A., Islam, Md. R., & Gunawan, T. S. (2020). A Systematic Review of Breast Cancer Detection Using Thermography and Neural Networks. *IEEE Access*, *8*, 208922–208937. https://doi.org/10.1109/access.2020.3038817

Using AI and deep learning to predict signs of breast cancer within images taken via thermography. Compares thermography techniques to other imaging modalities by providing its advantages and disadvantages. This article was published in March 2020 at an IEEE conference in New Delhi, India. This would make this article a scholarly reference as it is published by a multinational institution focused on research and innovation. This is a clear example where AI shows the potential to supplement new imaging techniques to identify cancer at its earlier stages.

IBM Cloud Education. (2020, June 3). *What is Artificial Intelligence (AI)?* Www.ibm.com. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

Contains basic information on Artificial Intelligence (AI). Can be utilized to define, and explain the origin of AI.

Iyer, V., Hima Vyshnavi, A. M., Iyer, S., & Namboori, P. K. K. (2019). An AI driven Genomic Profiling System and Secure Data Sharing using DLT for cancer patients. *2019 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 1-5. https://doi.org/10.1109/ibssc47189.2019.8973020

Using genetics and sample images of patients with different melanoma (skin cancers) to train a model that would predict early signs of melanoma along with the gene associated with the melanoma type. This article also dabbles in information sharing to keep patient confidentiality secret. It was published by IEEE in the 2019 IEEE Bombay section Signature Conference. As IEEE is a multinational institution that focuses on research, this article can be considered to be of the scholarly variety. This is yet another example of how AI can be utilized to research and identify cancer.

Kulkarni, S., Bhat, S., & Moritz, C. A. (2019). Reconfigurable Probabilistic AI Architecture for Personalized Cancer Treatment. *2019 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1-7. https://doi.org/10.1109/icrc.2019.8914697

This article tackles the aspect of cancer as a disease caused by genetic defects by developing an AI model that can personalize treatment for patients under the aforementioned circumstances. As IEEE is a multinational institution that focuses on research, this article can be considered to be of the scholarly variety. This is another example of how AI’s potential to personalize cancer patient’s treatments.

Moser, E. C., & Narayan, G. (2020). Improving breast cancer care coordination and symptom management by using AI driven predictive toolkits. *The Breast*, *50*, 25–29. https://doi.org/10.1016/j.breast.2019.12.006

This article tackles customized cancer treatment from primary health professionals through long periods of time by speaking about how AI can be used in cancer treatment to mitigate some issues that occur at different stages of care. This article was published by Elsevier on “The Breast” Volume 50, making this a scholarly article as it has been peer-reviewed. This article demonstrates the ideal situations where AI has great potential to aid Cancer research as well as its treatment.

Tew, K. D., & Fisher, P. B. (2019). *Advances in cancer research. Volume 142*. Academic Press. https://gcu-encore.iii.com/iii/encore/record/C\_\_Rb1373095\_\_SCancer%20\_\_P0%2C21\_\_Orightresult\_\_U\_\_X6?lang=eng&suite=def

This book shows advances in cancer research. It demonstrates the technology currently used to detect and treat cancer. This is a scholarly article because it was released on ScienceDirect, a platform that focuses on scientific research, and it was also edited by at least one medical Doctor. This provides further details on the advances of technology in cancer research and also illustrates how AI can boost the current advances in cancer research.

Yadav, K. K. (2018). How AI Is Optimizing the Detection and Management of Prostate Cancer. *IEEE Pulse*, *9*(5), 19–19. https://doi.org/10.1109/mpul.2018.2866354

This article provides a short description of prostate cancer for men, as well as some statistics associated with the mortality, costs, and diagnoses done yearly for prostate cancer in males. This is a scholarly article published in 2018 by IEEE (IEEE Pulse). this can be utilized to point out where AI has the most potential to aid in cancer research and its treatments.

Appendix B – Copyright Compliance

[For each external technical tool or code used, provide a reference to its copyright policy, clearly showing your right to use it. For each external technical tool or code used, detail how you used it, how you adapted it, how you modified it (if permitted), and why did you use it as opposed to write your own. Only a small portion of your project may rely on external code. When code libraries/packages are used, explain why this was necessary/required/recommended. Seek instructor approval for using external resources prior to beginning to work on the project.]