Artificial Intelligence-Driven Detection of Genetic Mutations in *Glioblastoma Multiforme* Using Genomic Data

Pranav

Shrewsbury Public High School

3058: Research Methods & Biology Honors

Ms. Jennifer Lambert-Peloquin

October 10, 2024

Genes are what make us up; they contain the code for our body, DNA, and without them we are nothing (MedLine Plus, 2020). But oftentimes genes may have deformity or errors and this is what causes cancer (National Cancer Institute, 2021). Cancer is one of the most well-known medical conditions in the world, notorious for having no solutions and being extremely deadly. While there can be cancer in any organ due to it being related to genes, it is often the most deadly in vital organs such as the heart, lungs, and brain. The most common tumor found in the brain is known as glioblastoma (Penn Medicine, 2018). Formally known as *glioblastoma multiforme* (GBM), also referred to as a *grade IV astrocytoma*, GBM is an extremely fast-growing brain tumor (Thakkar et al., 2024). Most patients with GBM often only have 12–15 months to survive if it isn't detected early enough (Glioblastoma Foundation, 2021). There are no known causes of GBM, though in some cases it can occur to people with certain genetic syndromes such as *neurofibromatosis type 1*, *Turcot syndrome*, and *Li Fraumeni syndrome*. Because there is no known cause of GBM, the only well-known way to detect it is after it has grown and you are experiencing extreme symptoms such as headaches, vomiting, and nausea (Mayo Clinic Staff, 2024). With artificial intelligence becoming more prominent in the medical scene, it would be a disservice to not try to use it in fighting our battle against cancer (Bohr & Memarzadeh, 2020). By changing the dependent variables, the artificial intelligence algorithm, training dataset, and what specific data used to train (whole gene, specific type of gene, specific part of gene, etc.); we can see the effect on the dependent variables, the accuracy of the artificial intelligence algorithm, by cross-testing with known cases of GBM and plugging it into this mathematical equation $accuracy = \frac{correct\ detections}{total\ detections}$. Specifically, I will be looking into the mutations found in the following: *tumor protein 53* (TP53), *phosphatase and tensin homolog* (PTEN), *epidermal growth factor receptor* (EGFR), and *neurofibromin 1* (NF1). The overall purpose of this project is to find the mutations in *tumor protein 53*, *phosphatase and tensin homolog*, *epidermal growth receptor*, and *neurofibromin 1* early so that we can improve treatment and outcomes. Traditional methods are slow and complicated, so we need an AI-based approach to quickly and accurately find these important mutations. Not only that, but *glioblastoma multiforme* has many genetic changes, making it hard to treat and very aggressive.

Artificial intelligence has become something that most people use in their everyday lives, whether it is through ChatGPT or Google Bard; we all have used it. In fact, ChatGPT receives nearly 100,000,000 users per week; now most of the things ChatGPT does are not anything entirely complex, but rather simple tasks such as documenting or planning out the day for someone (Porter, 2023). But artificial intelligence has the ability to do much more complex things, such as analysis of data, which can help society. But to really understand how artificial intelligence works, we need to understand what it is. Artificial intelligence (AI) is the simulation of human thinking by a system or machine (Xu et al., 2021). The goal of AI is to develop a system or algorithm that can think like humans or mimic human behaviors, such as perceiving, reasoning, learning, planning, predicting, and so on (Xu et al., 2021). In this project, I will be focusing on the human behavior of predicting or finding. But this wasn't always the definition of AI; in the past, it has carried many different definitions that have changed as the amount we know about AI has increased. The concept of machines being able to do what humans are able to do is nothing new; philosophers back in the 1700's also thought about algorithms that could mimic human intelligence (Maryville University, 2023). The two main people who are regarded as the inventors of AI are Alan Turing and John McCarthy. Alan Turing is often regarded as the "father of AI," largely because of his development of the Turing Test in 1950 (Maryville University, 2023). This test offered a theoretical way to distinguish between human intelligence and artificial intelligence by evaluating whether a machine could think through a series of questions (Maryville University, 2023). In 1955, John McCarthy introduced the term “artificial intelligence” in a research proposal, aiming to explore the idea that a machine could replicate the fundamental aspects of human intelligence (Maryville University, 2023). There are often many terms thrown around in the AI world, like neural networks and gradient descent, but those are often specific toward certain AI algorithms.

In my case, I will be using a convolutional neural network (CNN). But I will also look at other different algorithms, such as Deep Learning (DL) and Machine Learning (ML) to see if there would be any better output. Genomic data, unlike image data, consists of sequences of nucleotides (T, C, G, A), and CNNs can be applied to detect patterns and motifs within these sequences. Instead of images, the input to the CNN will be nucleotide sequences. The sequences need to be converted into a numerical format that the CNN can process. A common way to represent DNA sequences is through one-hot encoding, where each nucleotide (T, C, G, A) is encoded as a binary vector (Piotr Skalski, 2019).

$$T = [0, 1, 0, 0]$$

$$C = [0, 0, 1, 0]$$

$$G = [0, 0, 0, 1]$$

$$A = [1, 0, 0, 0]$$

For a DNA sequence of length L, the input to the CNN will be a matrix of size $L \times 4$, where each row represents one nucleotide in one-hot encoding format (Ganji, 2019). The convolution layer is used to identify patterns or motifs in the genomic sequence. These motifs could represent biologically significant features such as transcription factor binding sites or promoter regions. Unlike images where we apply 2D convolutions, genomic data can be treated as a one-dimensional sequence. Therefore, we use 1D convolutions to scan the genomic sequences. The one-hot encoded sequence is a matrix of size $L \times 4$ (sequence length) by nucleotide encoding. Instead of a 2D filter as used for images, here you apply 1D filters of size $n \times 4$, where n represents the length of the pattern you're trying to detect (such as a DNA motif) (Tariq, 2023). DNA motifs are different recurrences in nucleotides (D'haeseleer, 2006). The filter slides over the sequence, identifying regions that match the pattern. Though this is one type of AI, I will experiment with many different types of AI algorithms.

To program this AI to make a model, I need to use a programming language. Computers only understand the numbers 1 and 0 (Rahul Awati, 2022). But it is impossible for humans to understand the language that computers use (Rahul Awati, 2022). And code is often what is used to write these models. A compiler is a program that translates source code into object code, processing the entire source code at once to reorganize and optimize the instructions (*Introduction to Compilers*, 2020). This is different from an interpreter, which executes source code line by line (*Introduction to Compilers*, 2020). While interpreters can run code immediately, compilers take time to produce an executable, but the compiled programs generally run faster (*Introduction to Compilers*, 2020). The translation from source code to object code involves several steps, including lexical analysis, parsing, and code generation, and the final executable is created using a linker and loader (*Introduction to Compilers*, 2020). There are many different programming languages that I can use to create the code for my project, such as Structured Language Query (SQL), Python, Java, JavaScript, C#, C++, R, C, Go, and Perl (Staff, 2020). Libraries are collections of pre-written code that programmers can call to perform specific tasks, helping to avoid the need to write repetitive code from scratch (Woke, 2023). They can be static, included at compile-time, or dynamic, loaded during runtime, allowing for greater modularity, code reuse, and efficient memory management (Woke, 2023). AI libraries and frameworks, such as TensorFlow and PyTorch, are essential tools for developing machine learning and deep learning applications, offering flexibility and scalability for various tasks (Melnik, 2023). Others like Scikit-Learn, Keras, and XGBoost provide interfaces for beginners, while specialized libraries such as Hugging Face and OpenAI, though it costs money for OpenAI, focus on natural language processing and advanced model deployment (Melnik, 2023).

At first glance, it may seem like choosing a language like C, C#, or C++ might be the best language due to it being close to assembly language, allowing for minimal abstraction and control over memory management, which results in faster execution compared to higher-level languages like Java or Python (*Why Is C Considered Faster than Other Languages?*, 2017). Libraries like NumPy, Pandas, and TensorFlow provide tools for data and artificial intelligence, making Python a great choice for building artificial intelligence models quickly (McFarland, 2022). Python is popular for AI and machine learning because its open-source libraries make complex tasks easier and boost productivity across different platforms (McFarland, 2022). Its ability to work well with other languages and a strong, supportive community offer lots of resources for developers (McFarland, 2022). Python is a flexible programming language that executes code line by line, meaning I can write and test code quickly without needing to compile it first (*Python Introduction* | *Python Education*, n.d.). It automatically determines the type of data you're using at runtime, which allows for easier and faster coding, but it also means that some errors might not show up until the code is run (*Python Introduction* | *Python Education*, n.d.). Additionally, Python uses indentation to define code blocks instead of braces or keywords, making the code generally cleaner and easier to read (*Python Introduction* | *Python Education*, n.d.). PyTorch is an open-source tool for deep learning that helps developers create and train neural networks easily (Ng & Katanforoosh, n.d.). It lets users change their models on the fly, making it user-friendly for testing and fixing issues (Ng & Katanforoosh, n.d.). PyTorch also includes many libraries for different tasks, like image recognition and language processing, so I can use it for various applications (Ng & Katanforoosh, n.d.).

The TP53 gene makes a protein called tumor protein p53 (or p53), which acts as a tumor suppressor by controlling cell division and stopping cells from growing too quickly or uncontrollably (Medline Plus, 2020). This protein is found in the nucleus of cells and attaches directly to DNA (Medline Plus, 2020). When DNA is damaged by things like harmful chemicals, radiation, or UV light, p53 helps decide whether the damage should be repaired or if the cell should self-destruct (Medline Plus, 2020). The TP53 gene produces a protein that can activate other genes and bind to DNA, allowing it to react to different types of stress within cells (TP53 Tumor Protein P53 [Homo Sapiens (Human)] - Gene - NCBI, 2019). Mutations in this gene are linked to many types of cancer, including inherited forms like Li-Fraumeni syndrome (Olivier et al., 2009). The PTEN protein, made by the PTEN gene, also plays a key role in managing cell growth and survival by blocking the Akt pathway (Lu et al., 2016). It serves as a tumor suppressor, and when it becomes hypermethylated, this change can be connected to different cancers (Lu et al., 2016). While PTEN hypermethylation has been seen in both inherited and non-inherited cancers, its specific connection to certain types of cancer, such as breast cancer, is still not well understood (Lu et al., 2016). The *epidermal growth factor receptor* (EGFR) is a protein found on certain cells that attaches to a substance called epidermal growth factor (National Cancer Institute, 2011). This receptor is part of cell signaling pathways that help control how cells divide and survive (National Cancer Institute, 2011). Sometimes, changes (mutations) in the EGFR gene can lead to an increased production of EGFR proteins on some cancer cells, which causes those cells to divide more quickly (National Cancer Institute, 2011). The *neurofibromin 1* (NF1) gene gives instructions for making a protein called neurofibromin, which is found in many cell types, including nerve cells and specialized cells like oligodendrocytes and Schwann cells (*NF1 Gene: MedlinePlus Genetics*, n.d.). These specialized cells create myelin sheaths that protect and insulate certain nerve cells (*NF1 Gene: MedlinePlus Genetics*, n.d.). If one copy of the NF1 gene is inactive, it can increase the risk of juvenile myelomonocytic leukemia (JMML), a rare type of cancer that usually affects children under 2, leading to an overproduction of immature white blood cells (*NF1 Gene: MedlinePlus Genetics*, n.d.).

As artificial intelligence has become extremely important in our day-to-day lives, it is important that we use it in fields like medicine, where any new breakthroughs can be revolutionary for patients. Specifically in dangerous conditions such as *glioblastoma multiforme* and cancer in general. Previous research has proven that artificial intelligence has the potential to and already has made changes that have saved people's lives. This project will be a stepping stone towards using artificial intelligence to detect dangerous mutations that lead to cancer, not only in *glioblastoma multiforme* but in the entirety of medicine. The integration of artificial intelligence in analyzing genomic data will significantly improve the detection accuracy of key genetic mutations in *tumor protein 53*, *phosphatase and tensin homolog*, *epidermal growth receptor*, and *neurofibromin 1* associated with *glioblastoma multiforme*, ultimately enhancing early diagnosis and treatment outcomes.

References

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, *1*(1), 25–60. NCBI. https://doi.org/10.1016/B978-0-12-818438-7.00002-2

D'haeseleer, P. (2006). What are DNA sequence motifs? *Nature Biotechnology*, *24*(4), 423–425. https://doi.org/10.1038/nbt0406-423

Ganji, L. (2019, June 12). *One Hot Encoding in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-one-hot-encoding/

Glioblastoma Foundation. (2021, August 30). *Glioblastoma: What Every Patient Needs to Know*. Glioblastoma Foundation; Glioblastoma Foundation. https://glioblastomafoundation.org/news/glioblastoma-multiforme

Huang, Y.-F., Chiao, M.-T., Hsiao, T.-H., Zhan, Y.-X., Chen, T.-Y., Lee, C.-H., Liu, S.-Y., Liao, C.-H., Cheng, W.-Y., Yen, C.-M., Lai, C.-M., Chen, J.-P., Shen, C.-C., & Yang, M.-Y. (2024). Genetic mutation patterns among glioblastoma patients in the Taiwanese population – insights from a single institution retrospective study. *Nature*. https://doi.org/10.1038/s41417-024-00746-y

Introduction To Compilers. (2020, May 24). Geeks for Geeks; GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-compilers/

Krishnamurthy, B. (2022, October 28). *ReLU Activation Function Explained | Built In*. Built In. https://builtin.com/machine-learning/relu-activation-function

Lu, Y.-M., Cheng, F., & Teng, L.-S. (2016). The association between phosphatase and tensin homolog hypermethylation and patients with breast cancer, a meta-analysis and literature review. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep32723

Maryville University. (2023, May 19). *History of AI: Timeline and the Future*. Maryville Online. https://online.maryville.edu/blog/history-of-ai/

Mayo Clinic Staff. (2024, June 20). *Glioblastoma - Symptoms and causes*. Mayo Clinic; Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/glioblastoma/symptoms-causes/syc-20569077

McFarland, A. (2022, April 13). *10 Best Python Libraries for Machine Learning & AI*. Unite.AI. https://www.unite.ai/10-best-python-libraries-for-machine-learning-ai/

MedLine Plus. (2020, September 17). *What is a gene?* MedLine Plus; National Library of Medicine. https://medlineplus.gov/genetics/understanding/basics/gene/

Medline Plus. (2020, August 18). *TP53 gene: MedlinePlus Genetics*. Medlineplus.gov. https://medlineplus.gov/genetics/gene/tp53/

Melnik, Y. (2023, September 29). *The Top 16 AI Frameworks and Libraries: A Beginner's Guide*. Datacamp; DataCamp. https://www.datacamp.com/blog/top-ai-frameworks-and-libraries

National Cancer Institute. (2011, February 2). *NCI Dictionary of Cancer Terms*. National Cancer Institute. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/epidermal-growth-factor-receptor

National Cancer Institute. (2021, October 11). *What Is Cancer?* National Cancer Institute; National Institutes of Health. https://www.cancer.gov/about-cancer/understanding/what-is-cancer

NF1 gene: MedlinePlus Genetics. (n.d.). Medlineplus.gov. Retrieved October 9, 2024, from https://medlineplus.gov/genetics/gene/nf1/#conditions

Ng, A., & Katanforoosh, K. (n.d.). *Introduction to Pytorch Code Examples*. Cs230.Stanford.edu; Stanford. Retrieved October 8, 2024, from https://cs230.stanford.edu/blog/pytorch/

Olivier, M., Hollstein, M., & Hainaut, P. (2009). TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, *2*(1), a001008–a001008. https://doi.org/10.1101/cshperspect.a001008

Penn Medicine. (2018, November 5). *What Are the Most Common Types of Brain Tumors? – Penn Medicine*. Penn Medicine; The Trustees of the University of Pennsylvania. https://www.pennmedicine.org/updates/blogs/neuroscience-blog/2018/november/what-are-the-most-common-types-of-brain-tumors

Piotr Skalski. (2019, April 12). *Gentle Dive into Math Behind Convolutional Neural Networks*. Medium; Towards Data Science. https://towardsdatascience.com/gentle-dive-into-math-behind-convolutional-neural-networks-79a07dd44cf9

Porter, J. (2023, November 6). *ChatGPT continues to be one of the fastest-growing services ever*. The Verge. https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference

Python Introduction | Python Education. (n.d.). Google Developers. Retrieved October 8, 2024, from https://developers.google.com/edu/python/introduction

Rahul Awati. (2022). *What is Binary*. WhatIs; TechTarget. https://techtarget.com/whatis/definition/binary

Staff, G. P. (2020, June 18). *The 10 Most Popular Programming Languages to Learn in 2021*. Graduate Blog. https://graduate.northeastern.edu/resources/most-popular-programming-languages/

Tariq, F. (2023, May 2). *Breaking Down the Mathematics Behind CNN Models: A Comprehensive Guide*. Medium. https://medium.com/@beingfarina/breaking-down-the-mathematics-behind-cnn-models-a-comprehensive-guide-1853aa6b011e

Thakkar, J., Peruzzi, P., & Prabhu, V. (2024, April 15). *Glioblastoma Multiforme*. American Association of Neurological Surgeons; AANS. https://www.aans.org/patients/conditions-treatments/glioblastoma-multiforme/

TP53 tumor protein p53 [Homo sapiens (human)] - Gene - NCBI. (2019). Nih.gov. https://www.ncbi.nlm.nih.gov/gene/7157

Why is C considered faster than other languages ? (2017, July 6). Geeks for Geeks; GeeksforGeeks. https://www.geeksforgeeks.org/c-considered-faster-languages/

Woke, G. (2023, March 24). *The difference between libraries and frameworks*. Simple Talk. https://www.red-gate.com/simple-talk/development/other-development/the-difference-between-libraries-and-frameworks/

Xu, Y., Wang, Q., An, Z., Wang, F., Zhang, L., Wu, Y., Dong, F., Qiu, C.-W., Liu, X., Qiu, J., Hua, K., Su, W., Xu, H., Han, Y., Cao, X., Liu, E., Fu, C., Yin, Z., Liu, M., & Roepman, R. (2021). Artificial Intelligence: a Powerful Paradigm for Scientific Research. *The Innovation*, *2*(4). Sciencedirect. https://doi.org/10.1016/j.xinn.2021.100179