THE CANADIAN SCIENCE FAIR JOURNAL

ARTICLE

Doctor X: An AI Powered Personal Medical Assistant

By: Gary Zhu | Edited by: Luka Zivkovic | Layout by: Ahmed Nadeem

Age: 16 | London, ON

As noted by Reisman (2017), data collection is currently a huge problem in healthcare. Reisman states that it is often a lengthy process to request documents from multiple different departments and sources. Only 6% of healthcare providers with Electronic Health Record (EHR) systems can send data to other clinicians that use a different system. These systems can even struggle to send data to systems built on the same platform. She concludes that one of the greatest problems for this is a lack of standardized health care records. Furthermore, she finds that EHR systems are often tedious and complex, causing physicians to spend double the time interacting with the EHR system and clerical work than actual clinical work, which leads to both physician burnout and a bogged-down system (Reisman, 2017).

Furthermore, Mirin (2021) notes that one of the biggest obstacles to AI in healthcare is efficient and clean data collection. They discovered that there is a lack of high-quality data in the medical field, as most of the data found in EHR systems are unusable for training AI models and must be cleaned by hand. As discussed previously, medical data from one organization can be incompatible with other platforms, making it harder to integrate datasets into AI models (Mirin, 2021). Additionally, there are many unconscious biases in data that come from different frames of reference which can make it hard for AI models to be accurate (Norori et al, 2021). For example, skin lesions may exhibit different characteristics between different genders and races. If the AI is predominantly trained using data gathered on members of one racial group, it may exhibit bias when applied to another.

Doctor-X is a standardized medical records system that helps people track their medical symptoms daily. If a patient is sick, they open the application through the web and input their symptoms into the system. This way, when they go into the doctor's office, they do not struggle when asked to recall their symptoms. Furthermore, it organizes all this information into specific categories, like hematology, biochemistry, etc., enabling other algorithms and applications to be run on the data. Doctor-X also provides some auxiliary features, including the graphing of symptoms, a note tracker, and an AI assistant algorithm. The AI assistant employed by Doctor-X helps strengthen the correlations found in patient data by providing an objective second opinion to healthcare professionals. Using the patient's symptoms, it attempts to find the disease that best matches the patient's condition. This reminds doctors of potential possibilities that may have been overlooked.

Since the pandemic started, our healthcare system has been overwhelmed by the sheer number of extra patients. With Doctor-X, doctors can get more accurate data quickly, allowing them to make a faster diagnosis with more effective treatments. It also helps patients understand the medical jargon and visualize the progression of their illness in real time. A centralized system like Doctor-X can help prevent that in the future, as all the data relating to the patient is stored in one place, making it easy to grant access and manage patient data. Additionally, consistency is built into the Doctor-X ecosystem, as all health metrics are stored with the same units, labelled with the same names, etc.

DESIGN AND IMPLEMENTATION

For the application, a front-end and a back end had to be created. The front end is also known as the user interface - it will display information to the user and allow them to interact with the application. This portion was created using ReactJS, a popular open-source web development framework created by those at Meta.

When a user enters data through the front end, it will be sent to the back end. The back end will process all the input from the user into meaningful information. It will organize it into tables and graphs and store it for future use. It will also analyze the data for any red flags in the user's information. To build the back end, Golang was used as the foundation, which is a fast and multi-platform language.

The AI embedded into the Doctor-X system is also part of the back end. Analyzing the information stored can extrapolate the possible diseases the patient could be suffering from. Many different AI algorithms were tested to find the most suitable one for this program. Two popular classification algorithms, the KNN and the Bayesian Network (see Table 1.0 for descriptions of each

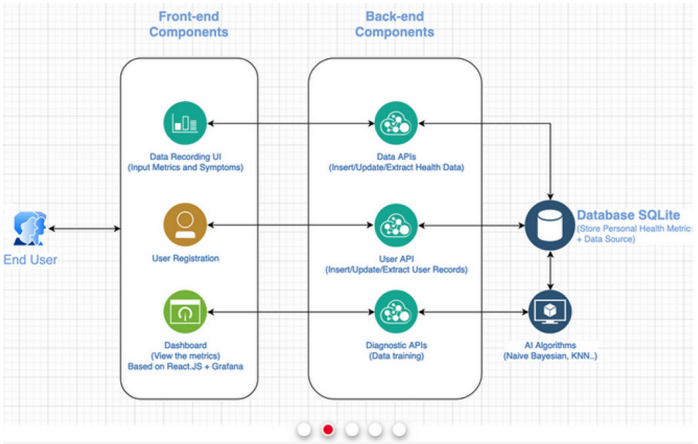

Figure 1: Doctor-X Architecutre Design

mode) were both tested on a dataset of about 300 different diseases. In the end, the Bayesian Network was chosen to be used in the final version for its higher accuracy on the testing dataset and ability to return multiple results easily.

To turn the data into graphs, Grafana was used. Grafana is an open-source analytics application used to create data visualization frameworks. Using Grafana, the information corresponding to the patients could be better organized to make it easier for patients and doctors to see patterns or trends. Examples of the data visualization system are present in Figure 2.0 and Figure 3.0.

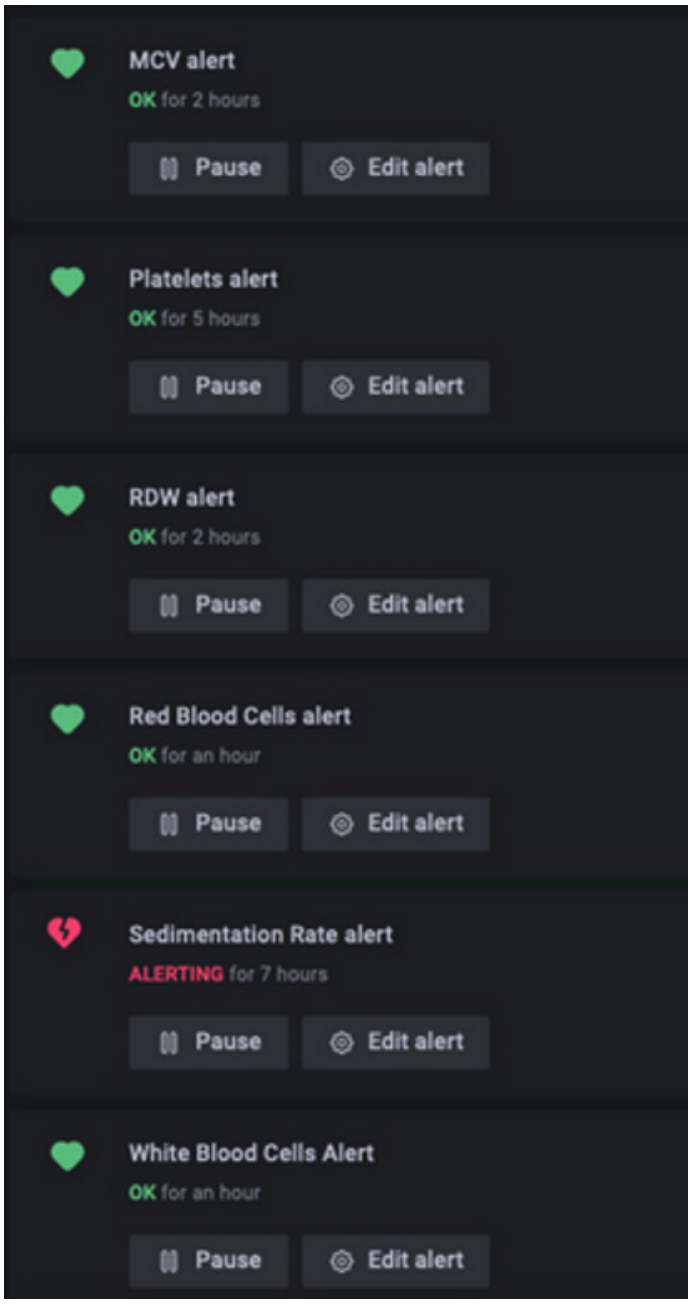

Figure 2: Alert Dashboard

Finally, to store all the information, it was necessary to have a data source. A data source takes in input and rearranges it in a way for it to be easily retrieved in the future. The data source used in this project was sqlite3, is a lightweight database that can also be used on mobile devices.

RESULTS AND OBSERVATIONS

By using the technologies above, a complete end-to-end ecosystem for data management, data visualization, and AI analysis was created. My ecosystem allows users to record their health metrics, including symptoms, vitals, and results from medical examinations. Doctor-X uses a modern UI framework, making it easy for users to enter complex data in seconds. The system can be accessed through the web on both computers and mobile devices. Most other systems either track symptoms or test results but are unable to do both at the same time. This allows Doctor-X users to store their personal data in one secure location instead of having to use multiple different software.

Doctor X also employs data visualization to display this data in easily readable formats (see Figure 2.0). Compared with other similar systems, my system shows the trends of multiple metrics simultaneously. Other systems can only track one at a time, forcing users to go back and forth, which can be very tedious for some. The stored information is displayed through multiple dashboards that allow both doctors and users to understand what they are looking at. Users can filter the metrics by category to narrow them down into specific areas. The system also incorporates alerts that show when the metrics are abnormal, allowing doctors to see what is wrong (see Figure 3.0) quickly. These alerts can be set up to automatically notify medical staff through a variety of channels like email, Slack, Discord, etc. This data visualization system will be extremely useful in the current pandemic as more and more appointments are being made through phones and other digital devices.

Finally, Doctor-X incorporates an AI recommender system that generates the three most likely diseases based on the symptoms the user displays (see Figure 4.0), which no comparable system was found to employ. After testing the AI model on a verification dataset, the AI model achieved an accuracy of 70% (see Figure 5.0), which is enough to be considered a good model (Barkved, 2022). As discussed in the procedure, Doctor-X utilized multiple different algorithms while testing the AI model, allowing comparisons between many machine learning algorithms (see Table 1.0). In training the AI model, a dataset with 800 diseases, 376 symptoms, and over 9000 metrics was used.

Overall, the Bayesian Network performed best, with an accuracy of 70.5% using a dataset with more than 300 cases. In comparison, the Euclidean KNN algorithm scored 38%, and the Manhattan KNN algorithm scored about 34%.

DISCUSSION

By analyzing other AI healthcare systems like IBM Watson and Dr. Google, one can see the advantages and disadvantages of the

Figure 3: Data Visualization System

AI system currently used in Doctor-X and areas of improvement. IBM Watson relies on unsupervised machine learning, meaning it learns by itself by reading medical papers. By analyzing texts, it can break down data into usable forms that it can use later to determine a final answer (Giacaglia, 2019). Doctor-X uses supervised machine learning, using datasets curated by medical professionals to learn and creating feature vectors to represent each data point in the dataset. Ultimately, this means that even though Doctor-X has no self-learning ability, relying on the knowledge of experts can generate accurate predictions. In fact, supervised machine learning may also result in a more accurate and precise model as humans can modify and fix the errors that they observe in supervised models because the underlying mechanisms are well understood more easily. In unsupervised machine learning, since the AI learns by itself, oftentimes, the AI model becomes a black box. Researchers can test how precise and accurate it is but are unable to understand why the AI makes the decisions it makes. Therefore, if something goes wrong in unsupervised machine learning, researchers are often forced to generate a completely new model.

Finally, the model was examined to determine how it could be improved. A big discrepancy in their accuracy can be observed between the two models tested (KNN and Bayesian). This may be due to the lack of available data, as a Bayesian Network can make relatively accurate guesses with a low amount of test data, while the KNN needs a large amount of data to become accurate. Most of the errors in both models come from a lack of clean data, which is one of the greatest struggles of AI in healthcare. For example, one dataset claimed that tuberculosis had a strong correlation with fever, and another claimed that it had a weak correlation. This may or may not be true, depending on the different frames of reference. If one dataset was curated by a doctor who mainly worked with fever-related diseases and with a doctor who doesn't, there would be large differences in how they perceive fever-related correlations. In addition to this, diseases behave differently based on location, age, genetics, and other factors, causing further discrepancies (CDC, 2012). This highlights the importance of creating one unified system vetted by multiple professionals and the necessity of centralizing medical information, as it is impossible for an AI to return sensible results with insufficient data (Mirin, 2021). Therefore, innovations in data collection are necessary, which is one of the reasons why the Doctor-X system was created. It is possible that the AI model would improve drastically with access to more comprehensive and objective data from hospitals and clinics.

CONCLUSION

To conclude, Doctor-X allows patients to store their data in one single, secure location. By quantifying all the data collected using a standardized method, Doctor-X paves the way for other AI innovations in healthcare. Doctors can also utilize Doctor-X to group data, showing trends in the health metrics of their patients, which helps forecast potential problems that may develop. Doctors can use the built-in AI recommender system to help them make a faster and more accurate diagnosis. The AI system could also become much more accurate given more data, metrics, and

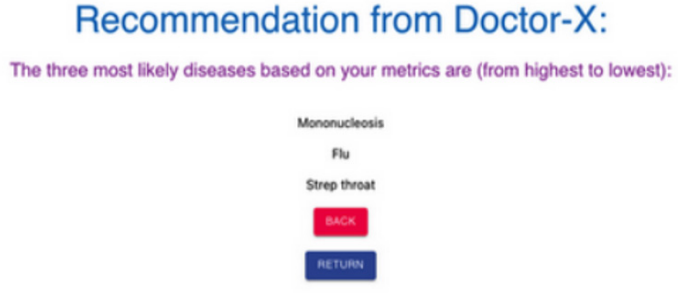

Figure 4: AI Results

Table 1: Advanatages and Disadvantages of Multiple AI systems in Doctor-X Ecosystem

Algorithm Comparison		
Algorithm	**Description**	**Observations gathered from the Doctor-X system**
KNN: Cosine Similarity	Measures the angle between two vectors to determine if they are pointing in a similar direction. The smaller the cosine of the angle, the more similar the two vectors are.	Fails when two vectors are collinear. Also, all vectors must be non-zero, which may not always be the case.
KNN: Manhattan Distance	Finds the kth nearest vectors projected in multidimensional space. Distance is calculated using the absolute value difference between coordinates.	More accurate than Cosine Similarity, but impacted by Curse of Dimensionality. As a result, it needs much more trainig data to be accurate. Therefore, this algorithm will also become very slow.
KNN: Euclidean Distance	Finds the kth nearest vectors projected in multidimensional space. Distance is calculated using the formula for Euclidean distance with Cartesian Coordinates in n dimensional space.	More accurate than Cosine Similarity, but impacted by Curse of Dimensionality. As a result, it needs much more trainig data to be accurate. Therefore, this algorithm will also become very slow.
Bayesian Network	A Bayesian network is a probabilistic graphical model in the form of a DAG (Directed Acyclic Graph) where each node in the graph is an event and each edge represents the probability that an event is true based on the conditions of a parent event. The likelihood of an event happening is then calculated using Bayes' Theorem.	Requires minimal amount of training data before it becomes accurate, allowing it to spend less time on analysing data. As a result, it is much faster than the KNN algorithm. Can provide the specific probability that someone is suffering of a certain disease. It is also very scalable, and can easily adapt to new training data.

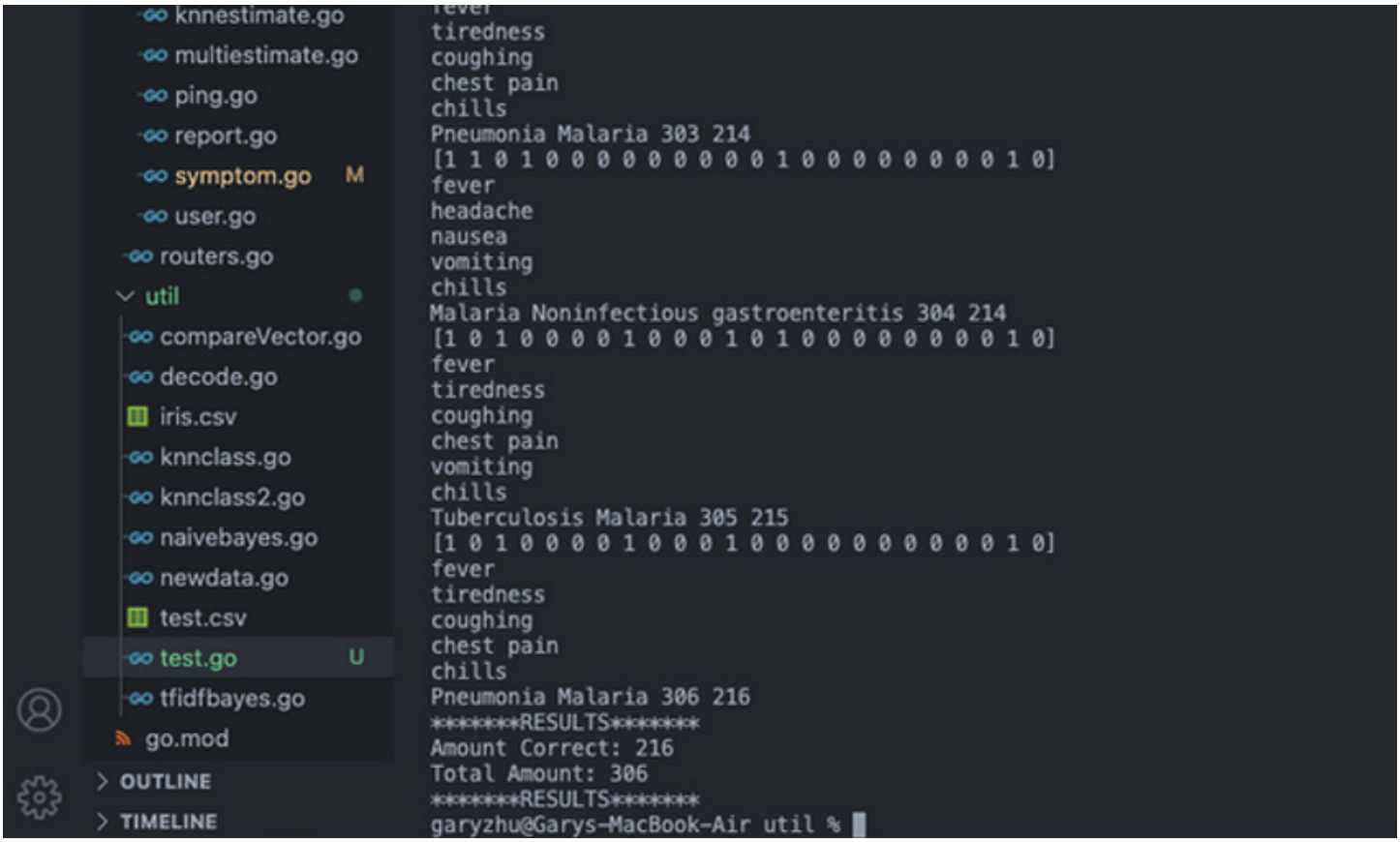

Figure 5: Results of Testing

different models. Even so, it still achieved a respectable accuracy of 70.5%. Doctor-X also comes with an alert system that can automatically email doctors when something is wrong, prompting them to check the Doctor-X software. Doctor-X shows that it works in both the prevention and early detection of diseases, potentially saving hundreds of millions of dollars for our healthcare system and many lives.

ACKNOWLEDGEMENTS

A very special thanks goes out to my parents, who have supported me every step of the way. This would not have been possible without your encouragement!

Another thank you goes to Dr. Charles Ling, a Professor from the University of Western Ontario specializing in Machine Learning, who generously took the time to provide feedback on my project.

Thanks to Jeff Regan and Dr. Tanner Tessier for providing guidance and helping me polish Doctor-X. I also would like to thank my editor, Luka Zivkovic for his invaluable feedback on my paper. Thank you all so much for helping me take this project to the next level!

I would also like to thank Dr. Kelly-Ann MacAlpine and Susan Lindsay for guiding me through the CWSF process. You were always there to answer any questions I had!

Finally, I would like to give a big thank you to both CWSF and TVSEF for giving me the opportunity to participate in the science fair, and to the CSFJ for the opportunity to publish my research and findings.

REFERENCES

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance Analysis and comparison. *Computers in Biology and Medicine,* Vol. 136, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Barkved, K. (2022). How to know if your machine learning model has good performance. *Obviously AI, Inc.*© Retrieved from https://www.obviously.ai/post/machine-learning-model-performance/

CDC Web Archive. (2012). Principles of Epidemiology in Public Health Practice, Third Edition: An Introduction to Applied Epidemiology and Biostatistics. *CDC*. Retrieved from https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section6.html

Giacaglia, G. (2019). How IBM Watson works [Blog]. *Medium*. Retrieved from https://medium.com/@giacaglia/how-ibm-watson-works-40d8d5185ac8

Langarizadeh, M., & Moghbeli, F. (2016). Applying naive Bayesian networks to disease prediction: A systematic review. *Acta Informatica Medica*, 24(5), 364-369. https://doi.org/10.5455/aim.2016.24.364-369

Liu, S., McGree, J., Ge, Z., & Xie, Y. (2016). Classification methods (Ch. 2). *Computational and Statistical Methods for Analysing Big Data with Applications,* pp. 7–28. https://doi.org/10.1016/b978-0-12-803732-4.00002-7

Mirin, K. (2021). Implementation of AI in Healthcare: Challenges and Potential [Blog]. *Postindustria, Inc.*© Retrieved from https://postindustria.com/implementation-of-ai-in-healthcare-challenges-and-potential/

Norori et al. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8515002/

Reisman, M. (2017). EHRs: The Challenge of Making Electronic Data Usable and Interoperable. *P & T* . Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5565131/

ABOUT THE AUTHOR - GARY ZHU

My name is Gary Zhu and I am currently a Grade 12 Student at Sir Frederick Banting Secondary School in London, Ontario. I'm very interested in both computer science and biology. I enjoy finding solutions to novel problems using computer algorithms. I love to challenge myself and learn more about the world around us. In the future, I wish to use my knowledge to help others. In my spare time, I like to hike, play the piano, and read novels.