Machine Learning Integration For Predictive Modeling of Medication Associated Cancer Risk

BY:
Mohammad R. Rashid

Harmony School of Innovation Sugar Land

TABLE OF CONTENTS

ABSTRACT

Machine Learning Integration For Predictive Modeling of Medication Associated Cancer Risk

More than 50% of diagnosed cancers are detected at a late stage, resulting in poor prognosis and high fatality rates (WHO, 2022). Current diagnostic methods necessitate in-person visits or specific test requests, leading to undetected cases until patients see the necessity when they experience health decline. Relatively few standard cancer screening protocols exist, and almost none of them are based on medication exposure history. Our solution, Cancer Assessment by Machine Learning (CAML), plans to revolutionize early detection by fixating on three silos of data: medications taken, associated relative risks, and associated cancers. Certain medications, such as Insulin, have an association with certain cancers, such as pancreatic cancer (De Souza et al., 2016). Other medications, such as tamoxifen, directly increase the risks of cancers due to their active ingredients, such as endometrial cancer (ACOG, 2006). Using an initial database from current research, CAML will leverage specific fields of Artificial Intelligence (AI) to map associations and identify patterns through decision tree regression methods. CAML's accuracy is then evaluated through mean squared error, correlation, and variance calculations. By leveraging this new field in medicine, CAML is envisioned to become a critical tool for medical professionals, predicting relative cancer risks and revolutionizing the early detection of cancers while also adapting to new data.

Keywords: Late diagnosis, correlation/association, machine learning, decision tree regression, mean square error, predicting cancer risk.

INTRODUCTION

RATIONALE:

The diagnosis of cancer in its later stages is a major contributor to the high mortality rates observed across different types of cancer, including lung cancer, breast cancer, and colon cancer – the most screened-for cancers in the United States. According to recent statistics by the Office for National Statistics (ONS), detecting cancers in the cervix, lungs, and bowel before reaching stage II significantly increases the survival rate, reaching up to 90%. Conversely, the survival rate drops dramatically to a mere 10% when cancer is detected after progressing to stage III (John, 2019). Several factors contribute to late-stage diagnoses, including a lack of public awareness of early symptoms and the asymptomatic nature of cancer until late metastasis.

According to the Cleveland Clinic, early-stage cancer often presents symptoms that overlap with mild cold or flu symptoms, such as headaches, tiredness, and muscle cramps (Cancer, 2022). This resemblance to common ailments leads to the oversight of cancer in its initial stages by both patients and healthcare providers. Consequently, there is an urgent need for innovative approaches to cancer prediction and early detection.

Through the careful analysis of patient medication history and external risk factors contributing to the risk of developing cancer, CAML can greatly increase the chance of either identifying factors that cause cancer or aid in the diagnosis of early-stage cancer. The integration of machine learning in understanding connections between a range of medications and their carcinogenic risks also enables providers to create a wider database of drugs and treatments that could put their patients at risk of developing certain types of cancer. CAML aims to provide easy-to-use software for patients of any ability to utilize to understand how their course of medication impacts their risk of developing certain types of cancer, among other diseases.

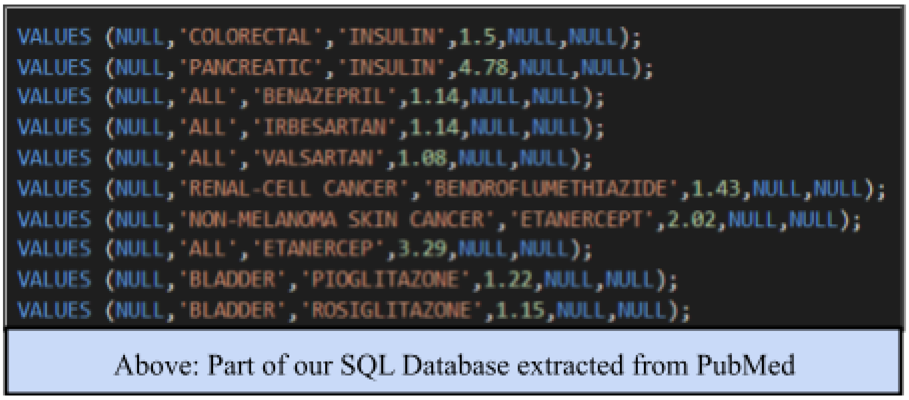
BACKGROUND RESEARCH

To address the challenge of late-stage cancer diagnoses, researchers are exploring unique approaches to predictive modeling of cancer risk. Artificial Intelligence (AI) and Machine Learning (ML) integration in the medical field holds promise due to their accuracy and predictive capabilities (Gibbons et al., 2019). However, to leverage these technologies effectively, researchers must identify trends or associations within data sets that algorithms can analyze (Brooks, 2023).

Recent research has uncovered a potential correlation between certain medications and an increased risk of developing cancer. The formation of nitrosamines, known to cause cancer, is linked to a weakened immune system induced by specific medications, occurring through lesions made in DNA by these nitrosamines, allowing mutations to develop and manifest as cancer. The US Food and Drug Administration (FDA) has identified over 250 medications, including commonly used drugs like epinephrine, sibutramine, and tamoxifen, that pose an increased risk of nitrosamine development during or after treatment cycles. In some cases, medications have been recalled due to dangerous levels of nitrosamines or nitrosamine-causing factors, including Zantac, Metformin, and Rifampin – all medications that treat severe conditions and diseases (FDA, 2023). Even a seemingly mild antifungal medication like voriconazole can increase your risk of melanoma due to phototoxic active ingredients (Miller et al., 2010).

In addition, some other medications are linked to specific cancers without necessarily posing any cancer risk. For example, patients who take medications such as Insulin have been shown to have an increased relative risk of developing pancreatic cancer. However, it is not the medication that poses cancer risk, but rather the disease that it treats: diabetes type II (De Souza et al., 2016). This association can be utilized as a sort of **“medication marker”** for the prediction of cancer risk.

Building on this information, we propose that the medication history of individuals can be correlated with the relative risks of developing specific types of cancer. By merging these two distinct sets of data – medication history and cancer risk – a machine learning algorithm could be developed to predict relative risks. This approach has the potential to enhance existing databases and contribute to a more comprehensive understanding of cancer risk factors.



```
VALUES (NULL,'COLORECTAL','INSULIN',1.5,NULL,NULL);
VALUES (NULL,'PANCREATIC','INSULIN',4.78,NULL,NULL);
VALUES (NULL,'ALL','BENAZEPRIL',1.14,NULL,NULL);
VALUES (NULL,'ALL','IRBESARTAN',1.14,NULL,NULL);
VALUES (NULL,'ALL','VALSARTAN',1.08,NULL,NULL);
VALUES (NULL,'RENAL-CELL CANCER','BENDROFLUMETHIAZIDE',1.43,NULL,NULL);
VALUES (NULL,'NON-MELANOMA SKIN CANCER','ETANERCEPT',2.02,NULL,NULL);
VALUES (NULL,'ALL','ETANERCEP',3.29,NULL,NULL);
VALUES (NULL,'BLADDER','PIOGLITAZONE',1.22,NULL,NULL);
VALUES (NULL,'BLADDER','ROSIGLITAZONE',1.15,NULL,NULL);
```

Above: Part of our SQL Database extracted from PubMed

Using the current research available, we extracted data from PubMed and created a Structured Query Language (SQL) database. This database includes the medication name, the type of cancer associated with it, and the relative risk of developing that type of cancer. With this initial database, we were ready for the development of CAML.

PURPOSE

The purpose of this research is to develop and evaluate a predictive modeling system, termed CAML, aimed at revolutionizing early cancer detection. By analyzing medication exposure history and associated relative risks, CAML aims to predict the likelihood of developing specific types of cancer, thereby enabling healthcare professionals to intervene early and improve patient outcomes.

RESEARCH QUESTION(S), HYPOTHESIS(ES), ENGINEERING GOAL(S), EXPECTED OUTCOMES:

Question:

Is it possible to integrate a machine-learning system with medication records to identify medications that cause cancers or are linked to cancers? Can these types of medications – which I coin as "medication markers" – serve as another way to identify relative cancer risks and save lives?

Hypothesis:

If cancer risks can be determined by mapping patterns with medication history, then an initial database from PubMed can be used to train a machine-learning model that will assess and predict cancer risks. This model will then identify "medication markers" to build up the initial database and become more accurate over time.

Goals/Development Criteria:

When developing CAML, criteria were set into place that would define our development process:

1. CAML must collect data beyond the initial research/databases provided
2. The solution needs to continuously analyze data imputed to draw conclusions – without human intervention – about relative cancer risk
3. The user interface (UI) has to be understandable and should display all relevant data

Expected Outcomes:

We expect that CAML will demonstrate high accuracy and reliability in predicting relative cancer risks based on medication exposure history due to the use of accurate data from PubMed research. Additionally, we anticipate that CAML will provide healthcare professionals and patients with valuable insights into cancer risk factors, enabling early intervention and improved patient outcomes.

MATERIALS and RESEARCH METHODOLOGY

Materials

1. Laptop/Desktop with capable CPU and GPU
2. Visual Studio Code
3. GitHub
4. Excel
5. PubMed
6. YouTube
7. Graphing Calculator

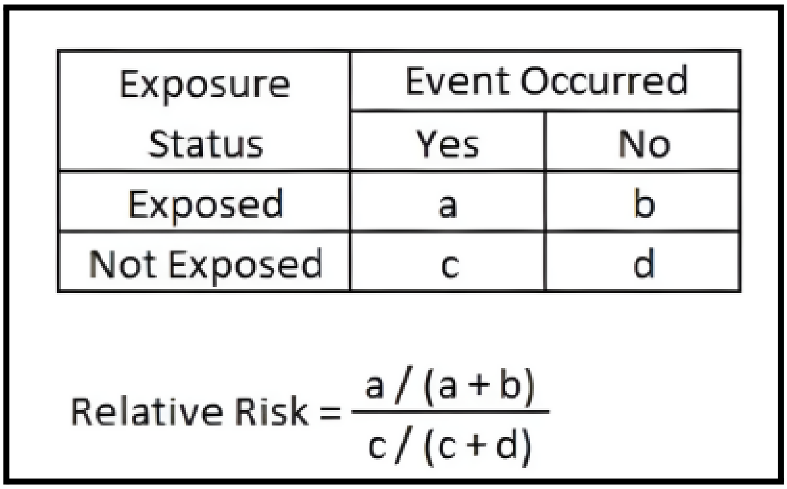
Location and Time Research

The development of the CAML project took place primarily at home and in a school computer lab equipped with powerful RTX 3060 GPUs. The development spanned over several months, from data collection to model training and testing.

Research Design

Section I - Expanding the Initial Database

Starting with the first criterion, the initial SQL database consists of only 39 entries and 10 medication categories, which is insufficient to cover all the medications a doctor/user may input. In addition, if we were to train our ML model on this data, it would be highly inaccurate with a large mean squared error in the regression model. This is because regression models are more accurate with greater amounts of data (Apte et al, 2022).



Exposure Status	Event Occurred	
	Yes	No
Exposed	a	b
Not Exposed	c	d

$$\text{Relative Risk} = \frac{a / (a + b)}{c / (c + d)}$$

To address the challenge at hand, CAML not only has the task of predicting relative risk but also needs to gather data on current cancer cases to establish the initial SQL database. Utilizing information from both cancer and non-cancer patients, CAML constructs the initial database using the formula for relative risk. This way, CAML is able to build up the initial database by adding more values over increased usage of the software. However, it's important to note that these relative risk calculations, while valuable for database construction, fall short when it comes to predicting the relative risk (RR) for individual patients. This limitation arises from the static nature of the calculations, preventing CAML from adapting its predictions to account for multiple associated cancers and relative risks for one medication. Furthermore, these static calculations are not able to draw patterns and learn from the data, and cannot discover new “medication markers”. The static nature of the calculations means that CAML cannot incorporate multiple contexts and create advanced prediction models. In complex scenarios like cancer prediction, where multiple factors come into play, this adaptability is crucial. This leads us to the second criterion for the development of CAML.

Section II - Machine Learning With Decision Tree Regression

CAML's second phase of development is aimed at solving this criterion and revolves around the concept of AI. Specifically, the field of ML was the most promising due to the possibility of the use of decision tree regression (DTR) models. DTRs take X values (medication name and associated cancer) and Y values (Associated RR values) from the initial database. Then a decision tree is automatically constructed with the designated training part of the database (80% of the initial database). DTRs, unlike

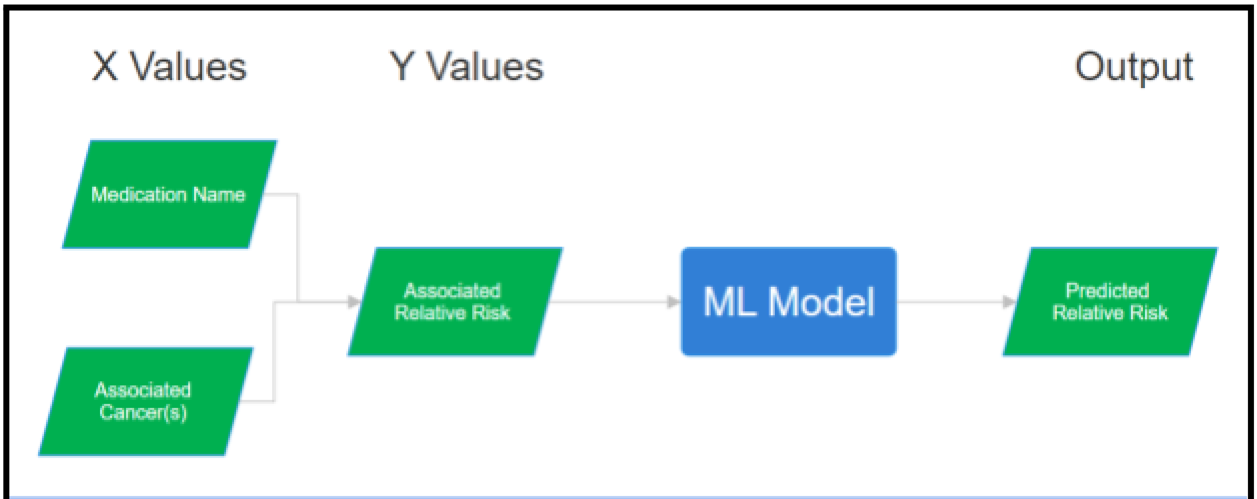


Figure 1: Abstraction of CAML's ML Algorithim

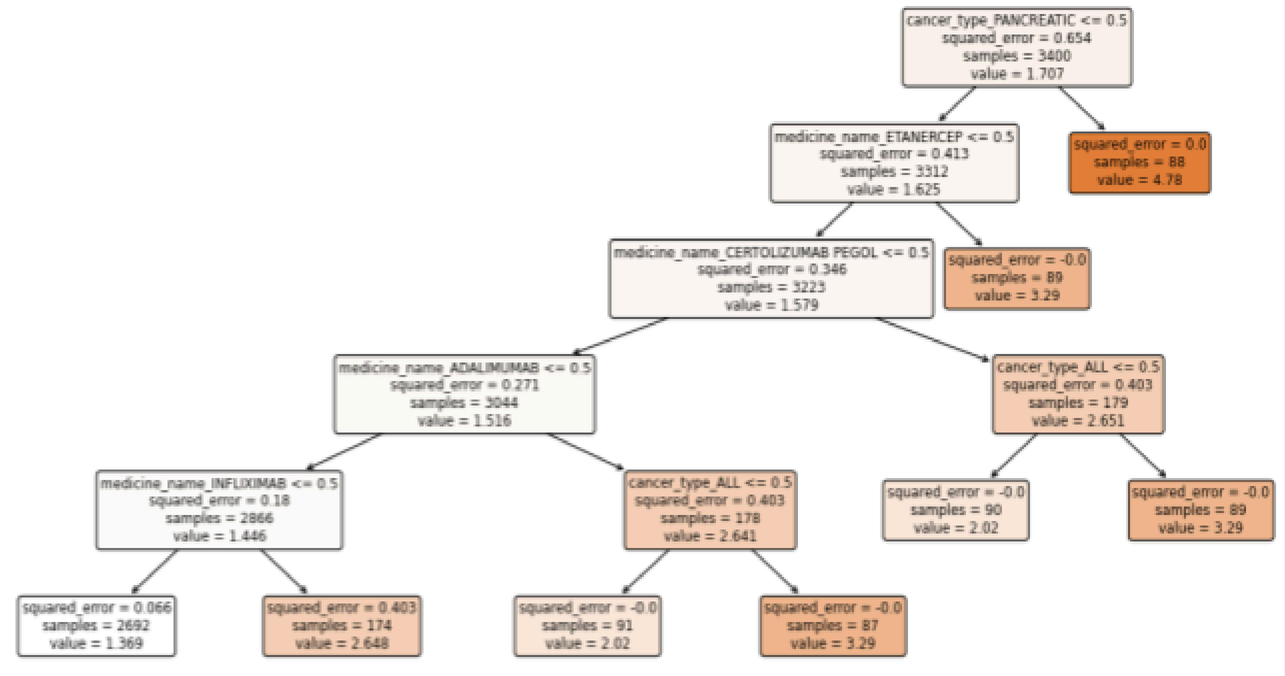


Figure 2: ML generated decision tree with value (RR) prediction

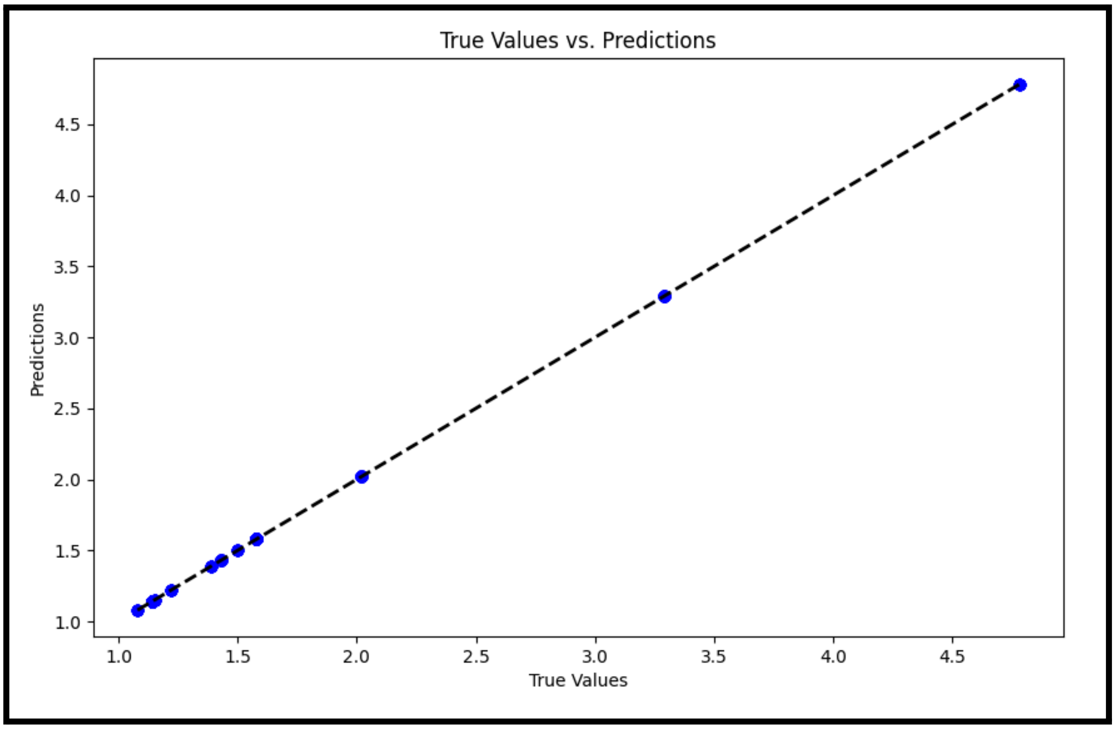
classic decision trees, focus on predicting a number with relational operators and boolean evaluation. The algorithm looks at the features and determines the optimal splits to make predictions based on the target variable (See Figure 2, which is a machine learning-generated tree that was limited to 5 nodes for representation purposes). Once trained, the decision tree acts as a predictive model. It can take new instances with features similar to those it has seen during training and predict the target variable based on the learned patterns.

After this is done, the model's accuracy is scored with the designated testing part of the database (20% of the initial database). This is done through the calculation of mean squared error (MSE), where the closer the value is to zero, the more accurate the model is to the initial database.

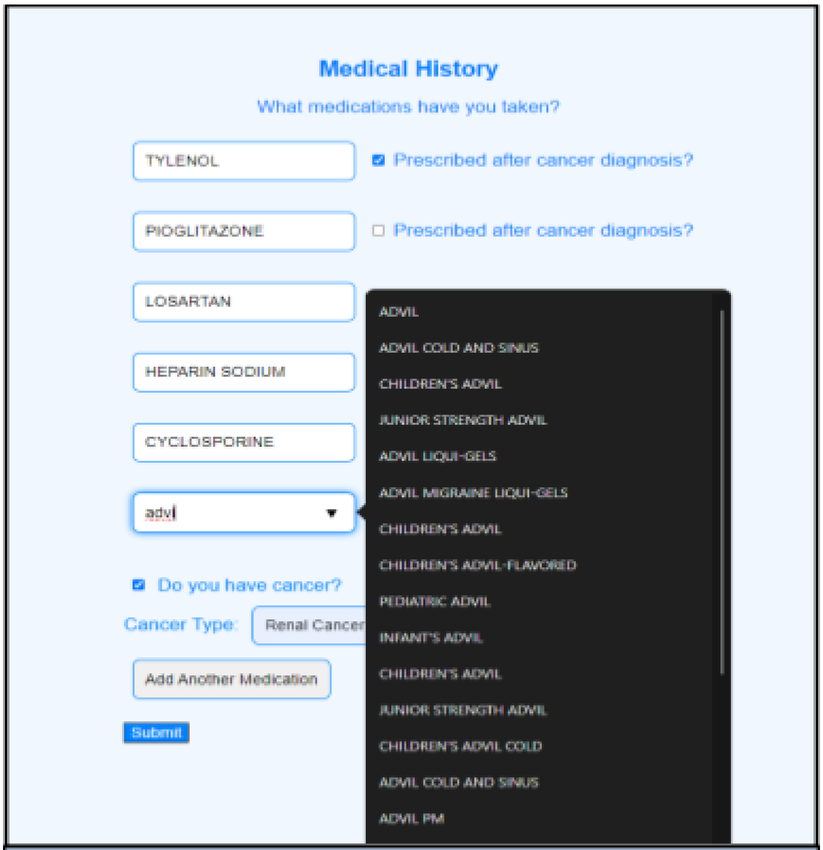
In the case where the user inputs a new medication unknown to the initial database or to the appended values from collected data (described in Section I), then CAML finds the mean value of all the nodes to provide an estimate of the relative risk of this new medication.

Section III - UI Development and ML Visualization

For the UI to be user-friendly while also being informative, CAML had to focus on simplicity. The first page the user is greeted by is very straightforward, asking several questions depending on the user's medication history. Our team made it easier for users by even adding an autocomplete drop-down list to prevent the hassle of spelling out individual medication names. This was done by extracting a 50-thousand-entry list of medications from the Federal Drug Administration (FDA). All this information is taken anonymously without the collection of identification data (such as location or name). After all the medication is imputed, it is asked if they have ever been diagnosed with cancer. If so, the name of the cancer is collected, and the



initial database collects these values to build upon the existing data. When the user presses submit, the ML model runs on the inputs and produces results almost instantly.



Medical History

What medications have you taken?

TYLENOL ☑ Prescribed after cancer diagnosis?

PIOGLITAZONE □ Prescribed after cancer diagnosis?

LOSARTAN

HEPARIN SODIUM

CYCLOSPORINE

advi

☑ Do you have cancer?

Cancer Type: Renal Cancer

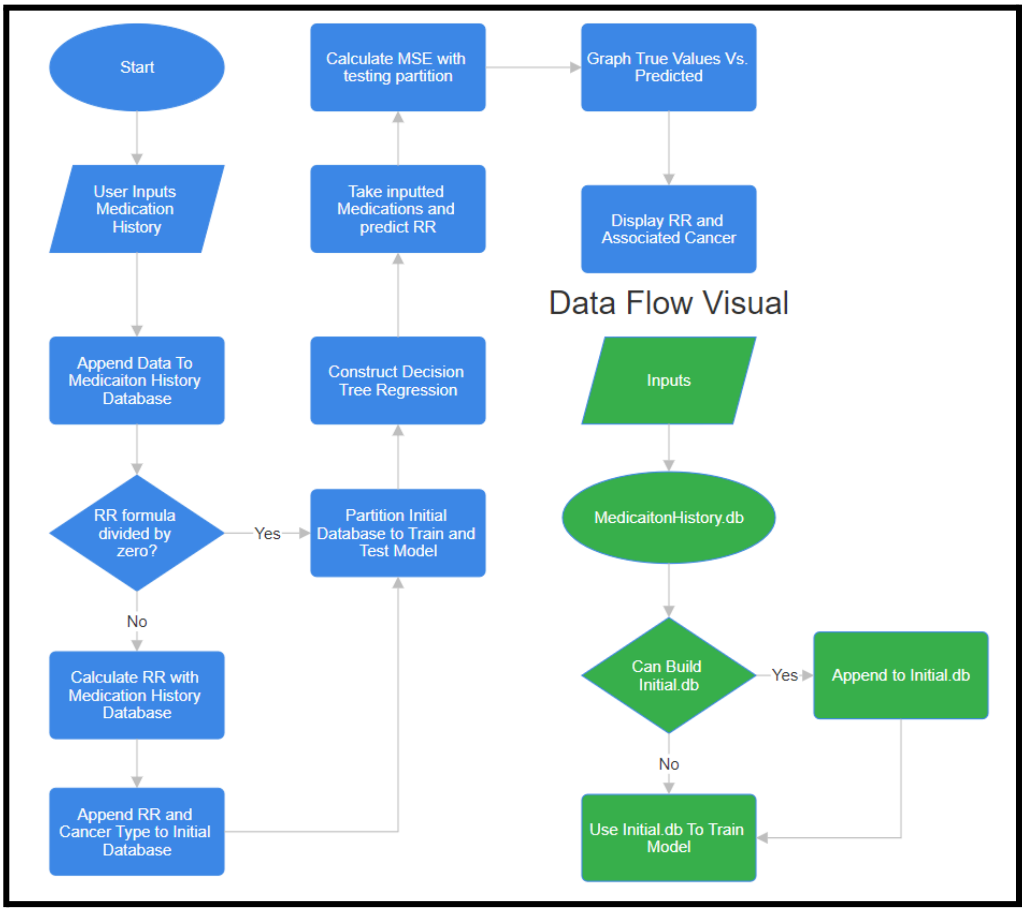
Example UI from CAML with autofill

The user is then greeted by a submission page that displays the predicted relative risk, the name of the medication with that risk, and the type of cancer(s) that may be associated with this medicine.

To allow a healthcare professional or user to visualize the accuracy of the model, the page will also display a graph. This graph displays True Values vs. Predictions and draws a line of best fit to allow users to determine the strength of correlation. The MSE is also displayed under the graph to further assist the user in determining the chance of error. If the model is deemed accurate, a healthcare professional can then determine the proper course of action with these predictions.

Abstracted Flow Chart Representation

w


Start
User Inputs Medication History
Append Data To Medicaiton History Database
RR formula divided by zero?
Yes
No
Calculate RR with Medication History Database
Append RR and Cancer Type to Initial Database
Partition Initial Database to Train and Test Model
Construct Decision Tree Regression
Take inputted Medications and predict RR
Calculate MSE with testing partition
Graph True Values Vs. Predicted
Display RR and Associated Cancer
Data Flow Visual
Inputs
MedicaitonHistory.db
Can Build Initial.db
Yes
Append to Initial.db
No
Use Initial.db To Train Model

Testing Methods

One of the most crucial factors in the development of a machine-learning algorithm includes prediction accuracy. The greater the accuracy a model presents, the greater the impact it has on its users. In CAML's case, accuracy means more accurate cancer predictions, which can save lives. In addition, variation between predictions – with similar data set sizes – has to be low to ensure data consistency.

Section IV - Coefficient of Determination and MSE Testing Method

To determine current accuracy and under which circumstances the accuracy of the model improves, our team designed a test. At an abstracted level, the test compares the results of our model to the existing research. For example, our team would input the medication Insulin into our model, record the relative risk, and then reference it to the relative risk for Insulin in current research (control). From there, our team would graph our data with a linear regression model to determine the coefficient of determination (R^2 coefficient). The team would also collect the MSE provided by the program from the True Values vs. Predictions graph. The MSE value will be used to cross-reference our findings with the R^2 coefficient.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

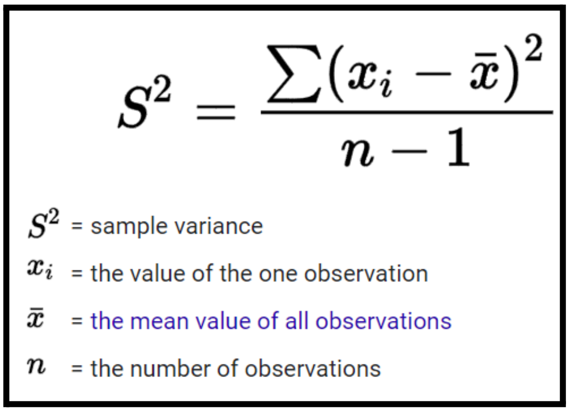
Y_i = observed values

$\hat{Y}_i$ = predicted values

Some variables that will be kept the same (control variable) include the medications being tested, database size, and number of trials for each medication. The independent variable will be the number of data entries in the initial database (either from PubMed or crowdsourcing) and the dependent variable will be the MSE and R^2.

Section V - Variance Testing Method

Consistency is a crucial factor in verifying the reliability of our model. Although variance is a natural phenomenon in machine learning, our model should not produce widely differing results under similar or identical conditions. To statistically measure this value, we adapted our program to calculate variance automatically. We will run this test with multiple data set sizes and multiple trials for each. Then, we will compare the variance values and determine if it is within an optimal range. The independent variable is the same as the first test – being the number of data entries – and the dependent variable is the variance values.



$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

S^2 = sample variance

x_i = the value of the one observation

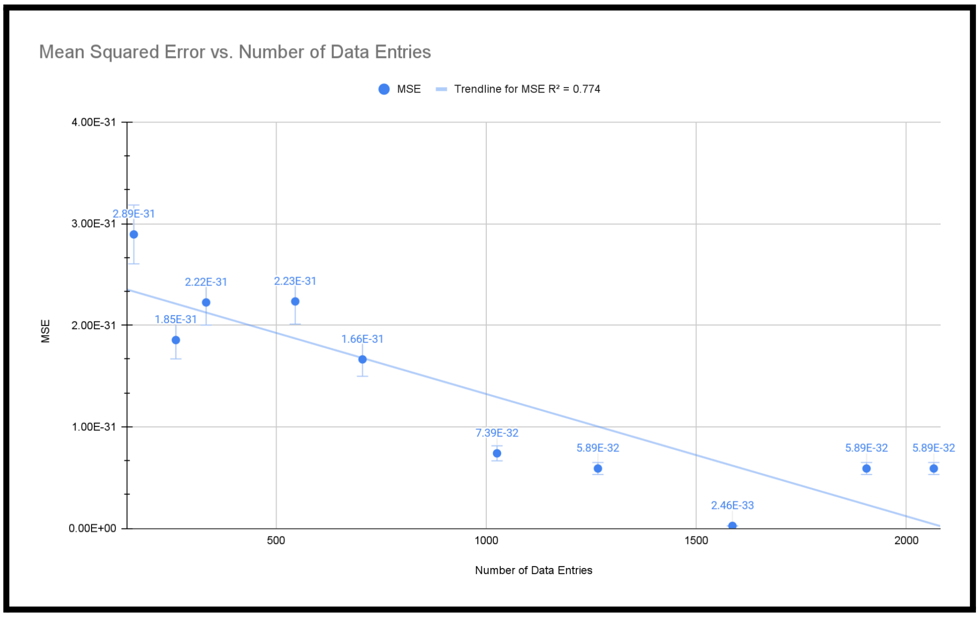
$\bar{x}$ = the mean value of all observations

n = the number of observations

Data and Results

Section VI - Coefficient of Determination and MSE Data and Results

During this testing phase, our team set out to collect data on the independent variable (Number of Data Entries) and the dependent variable (the value of the MSE).



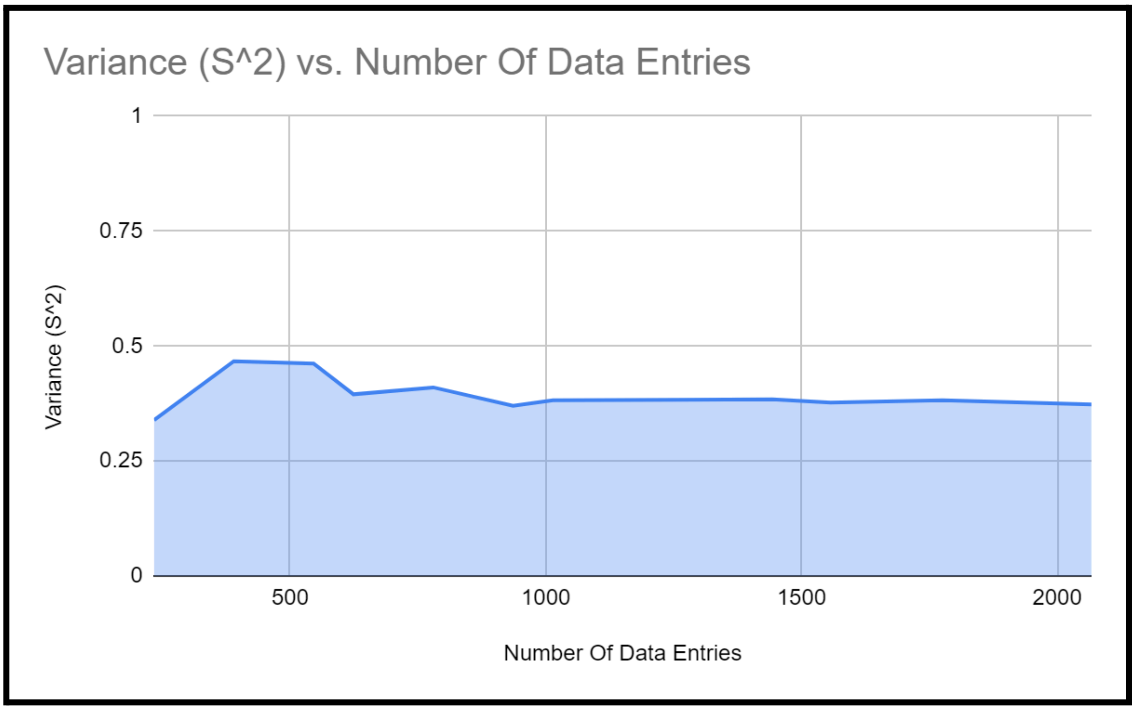
We started testing at 162 entries and concluded testing at

2065 entries for one medication, Benazepril. We performed multiple trials and aggregated the results by using mean averages to create one graph. From there, we used a spreadsheet to automatically calculate the strength of correlation, or R^2.

Analyzing the data, it can be inferred that a medium-strong correlation exists. However, it is also noted that the MSE starts to increase with smaller data sizes. This phenomenon is expected, as machine learning models typically require larger sets of data to achieve greater accuracy.

Section VII - Variance Data and Results

Determining the variance (S^2) of the model followed an identical procedure to the R^2 Coefficient test. However, unlike the previous test, instead of only examining the value S^2, we also need to examine the consistency over the trials, with each trial containing more data entries.



After nine trials, we graphed our data. In the first four trials with data entries under 1000, the variance nearly reached values of 0.5. In addition, the first four trials had measurable inconsistencies between the variance values.

From there, our team decided to continue increasing the size of the data entries. After entering above 1000 entries, it was found that variance values stayed under 0.40. The variance values between the rest of the trials were also similar in magnitude.

Research Procedure

Data Collection:

- Gathered initial data from PubMed and created an SQL database containing medication names, associated cancers, and relative risks.
- Expanded the initial database by incorporating data from cancer and non-cancer patients to establish a foundation for CAML.

Development of CAML:

- Set criteria for CAML development, including the collection of extensive data, continuous analysis for relative cancer risk without human intervention, and a user-friendly interface.
- Recognized the need to expand the initial database for accurate machine learning predictions, considering the limitations of regression models with smaller datasets.
- Formulated a strategy for CAML to autonomously analyze and draw conclusions from inputted data to predict relative cancer risks.

Machine Learning with Decision Tree Regression:

- Employed decision tree regression (DTR) models for machine learning, which utilized X values (medication name and associated cancer) and Y values (associated relative risk).

- Split the initial database into training (80%) and testing (20%) sets for the DTR model.
- Trained the DTR model to predict relative cancer risks based on learned patterns from the training set.
- Evaluated the model's accuracy using mean squared error (MSE) on the testing set.

UI Development and ML Visualization:

- Designed a user-friendly interface for CAML, focusing on simplicity.
- Implemented an autocomplete feature for medication names using an extensive list from the FDA to enhance user experience.
- Collected data on users' medication history and cancer diagnosis, which was then used to build and refine the database.

Testing Methods:

- Developed testing methods to assess CAML's prediction accuracy, consistency, and variance.
- Utilized the coefficient of determination (R^2) and MSE to evaluate the accuracy of the model concerning existing research data.
- Conducted variance testing to assess the consistency of results under similar conditions and data set sizes.

Data and Results Analysis:

- Conducted testing with different data set sizes for a specific medication (Benazepril) to observe the correlation between the number of data entries and model accuracy.
- Analyzed the collected data to infer trends and correlations, acknowledging the expected behavior of machine learning models requiring larger datasets for greater accuracy.
- Examined variance data to ensure consistency in results under different conditions and data set sizes.

Drawing Conclusions:

- Concluded that CAML exhibited proportionality in terms of accuracy and the number of data entries.
- Recognized that larger data sets led to better model accuracy, a common characteristic in machine learning models.

Suggestions for Future Tests:

- Identified the need for long-term research studies to assess the efficacy of CAML in predicting cancer risks and its impact on patient outcomes.
- Highlighted the importance of testing the scalability of CAML to ensure efficient performance with larger datasets.
- Emphasized the necessity of security testing to safeguard user data and prevent potential vulnerabilities.

Discussion and Measuring Impact:

- Discussed potential benefits and consequences of CAML, including the importance of measuring its impact through partnerships with healthcare providers.
- Proposed future improvements based on impact measurement research, including incorporating more factors and transitioning to a Deep Learning algorithm.

GitHub Repository Link:

- Provided a link to the GitHub repository for CAML, allowing for further exploration and collaboration through open source development.

Bibliography:

- Cited relevant sources and research studies supporting the rationale and development of CAML, including studies on machine learning in medicine, cancer therapy-related hypertension, and the influence of diagnostic delay on survival rates for colorectal cancer.

Data Analysis

Section VIII - Analyzing the Data

Analyzing the results from the two tests, one major trend was found: CAML exhibits a proportionality in terms of accuracy and the number of data entries. This trend was more profound in the variance test, but also notable in the variance test.

These behaviors are expected and displayed in many current machine-learning models. One notable example that corresponds with our topic of interest includes models for medical imaging. These models require large amounts of data to accurately complete their respective tasks, and improve with more data inputs until a certain threshold (Rana et al., 2022).

Overall, CAML’s model presented results that proved its ability to predict cancer risks as it was able to replicate results similar to current PubMed research. Furthermore, CAML proves very reliable with the testing set due to little variance.

Risk and Safety

Potential risks associated with the CAML project include data privacy concerns, as the project involves collecting and analyzing sensitive medical information. To mitigate these risks, stringent measures were implemented to ensure data anonymization and security. Additionally, ethical considerations were taken into account to ensure the responsible and ethical use of patient data in the development and testing of the CAML model. Furthermore, safety precautions were observed while working with powerful GPUs in the computer lab to prevent accidents or damage to equipment.

RESULTS and DISCUSSION

Results

Preliminary findings indicate that CAML demonstrates a high level of accuracy in predicting relative cancer risks associated with medications when cross-referenced with PubMed data. The R^2 coefficient analysis revealed a strong correlation between CAML predictions and actual data, suggesting that the model effectively captures underlying trends and associations.

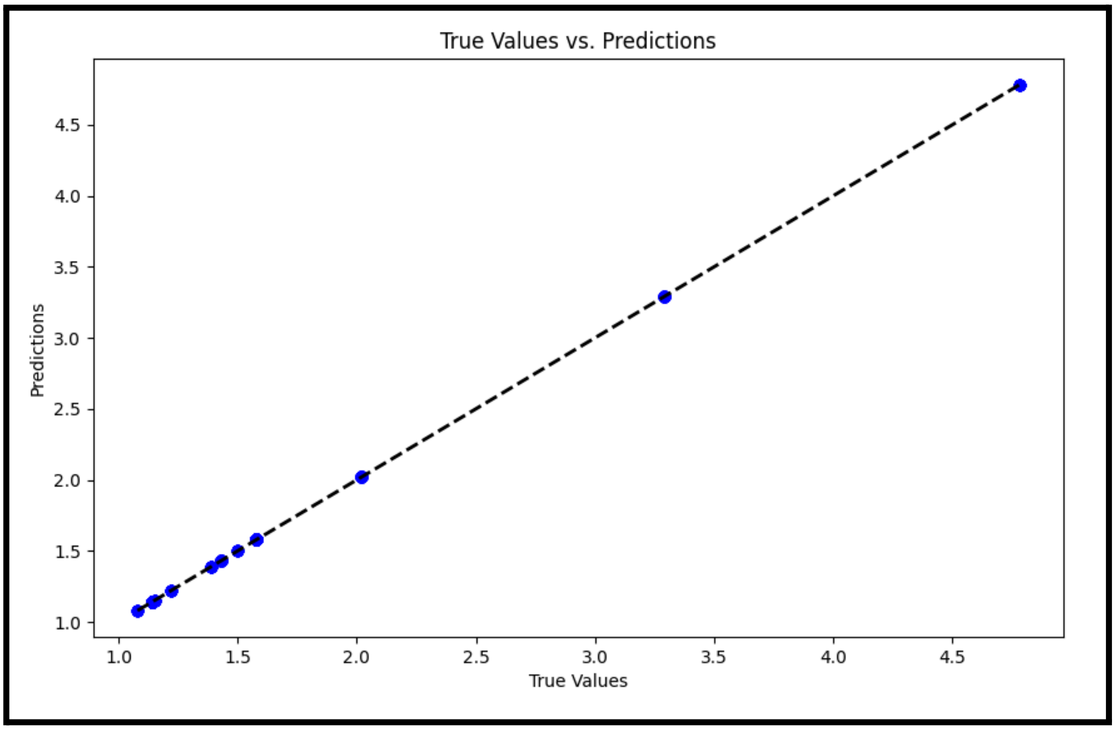
Furthermore, the MSE analysis demonstrated low error rates, indicating that CAML's predictions closely align with the actual relative risks documented in PubMed.

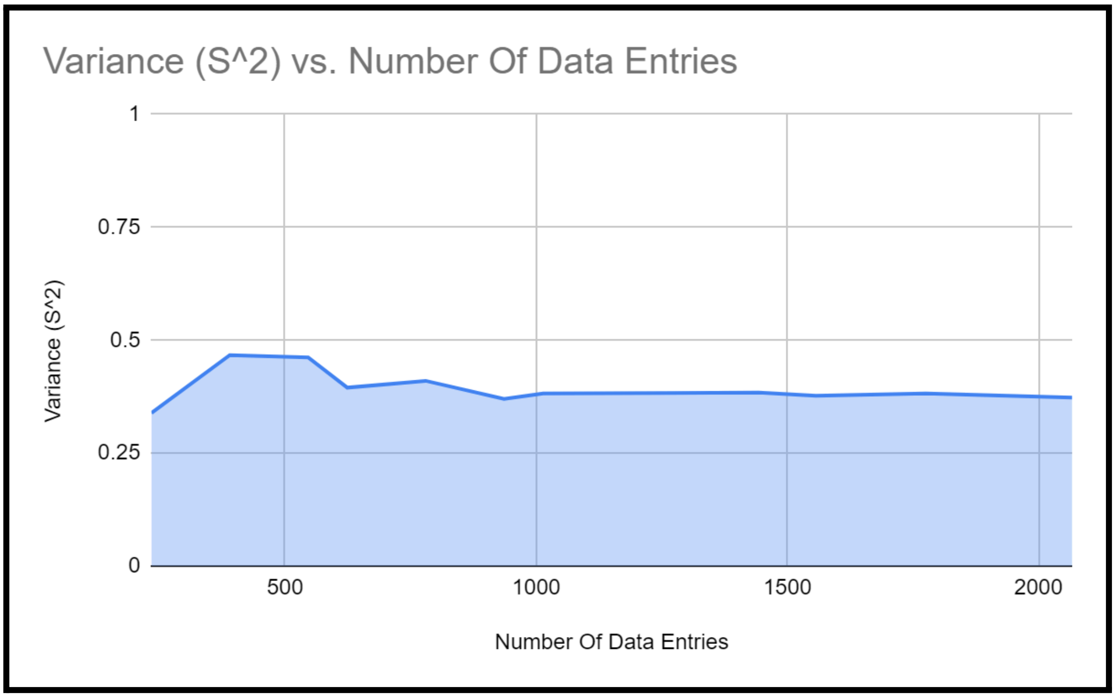
The small margin of error underscores the reliability and precision of CAML in predicting cancer risks based on medication history.

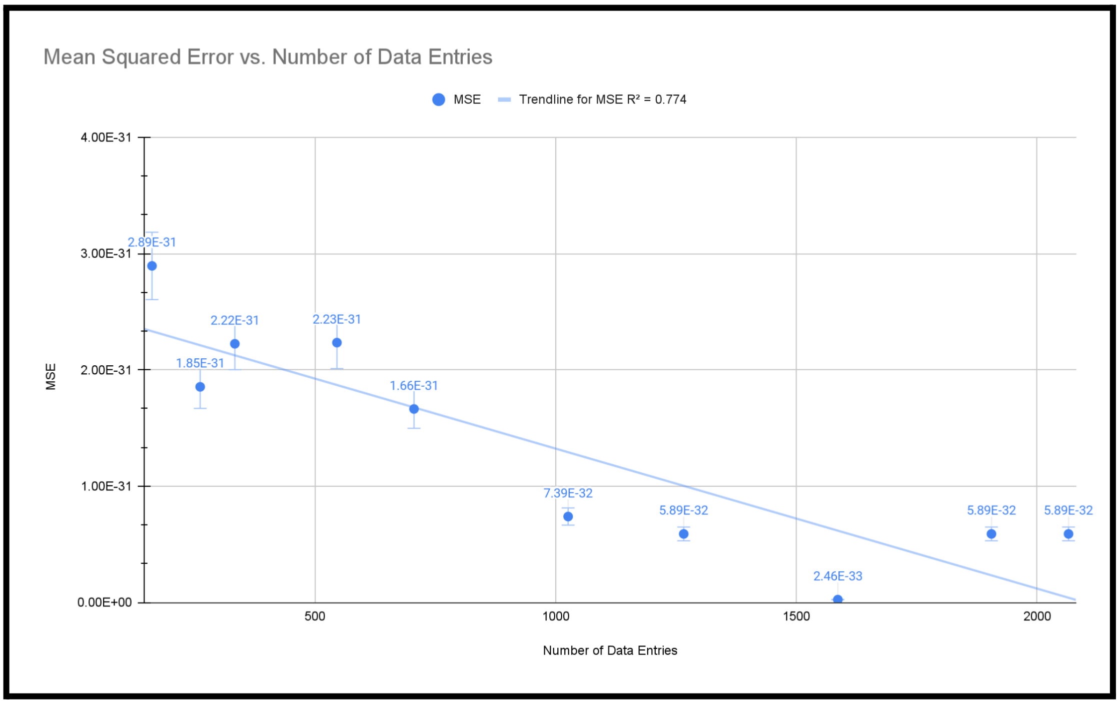
Specifically, CAML's predictions exhibited consistent alignment with known associations between medications and cancer risks, as documented in PubMed. The model accurately identified medications with elevated cancer risks and provided reliable estimates of the relative magnitudes of these risks.

Overall, the results suggest that CAML holds promise as a robust tool for predicting cancer risks associated with medication usage. Its ability to generate accurate predictions based on machine learning algorithms offers valuable insights for healthcare providers and patients in understanding and mitigating cancer risks associated with specific medications. However, further validation and refinement of CAML through larger-scale studies and real-world applications are warranted to confirm its efficacy and reliability in clinical settings.

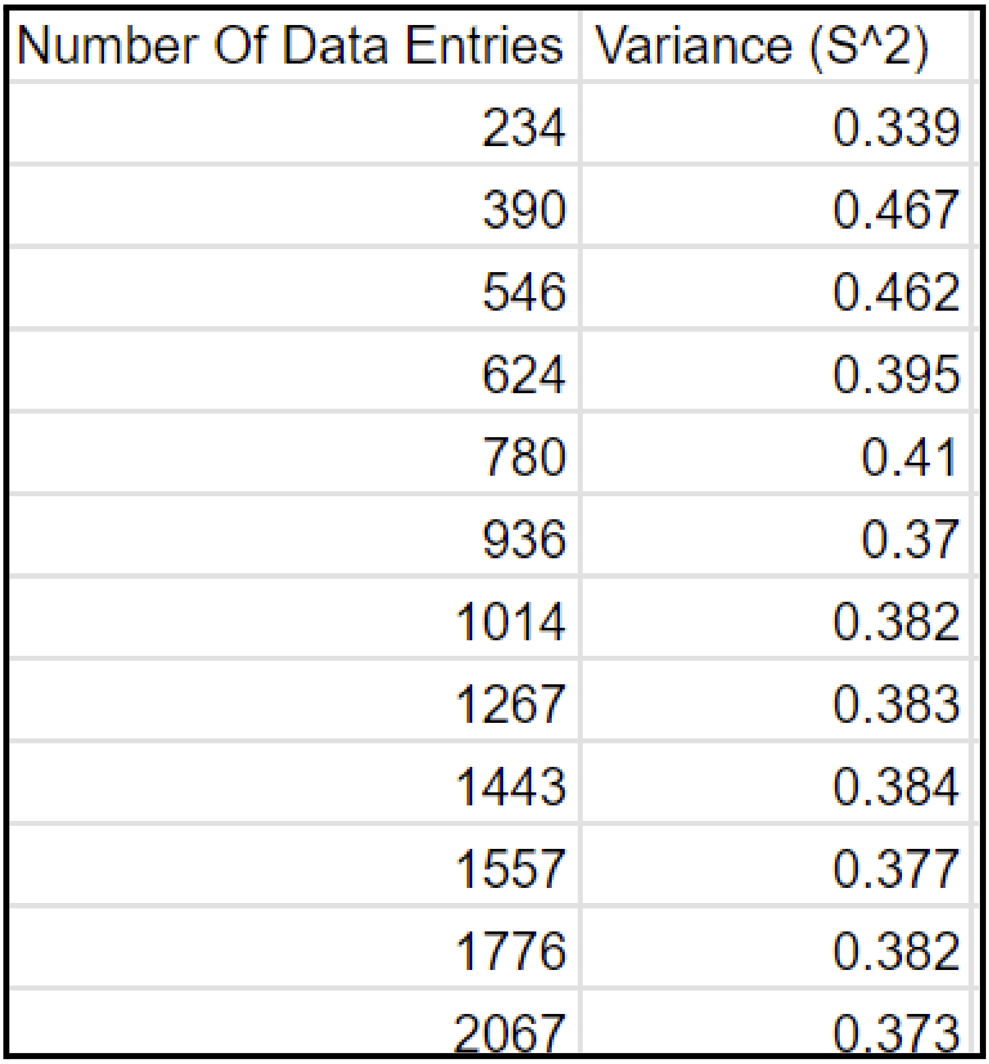
Data Tables, Graphs



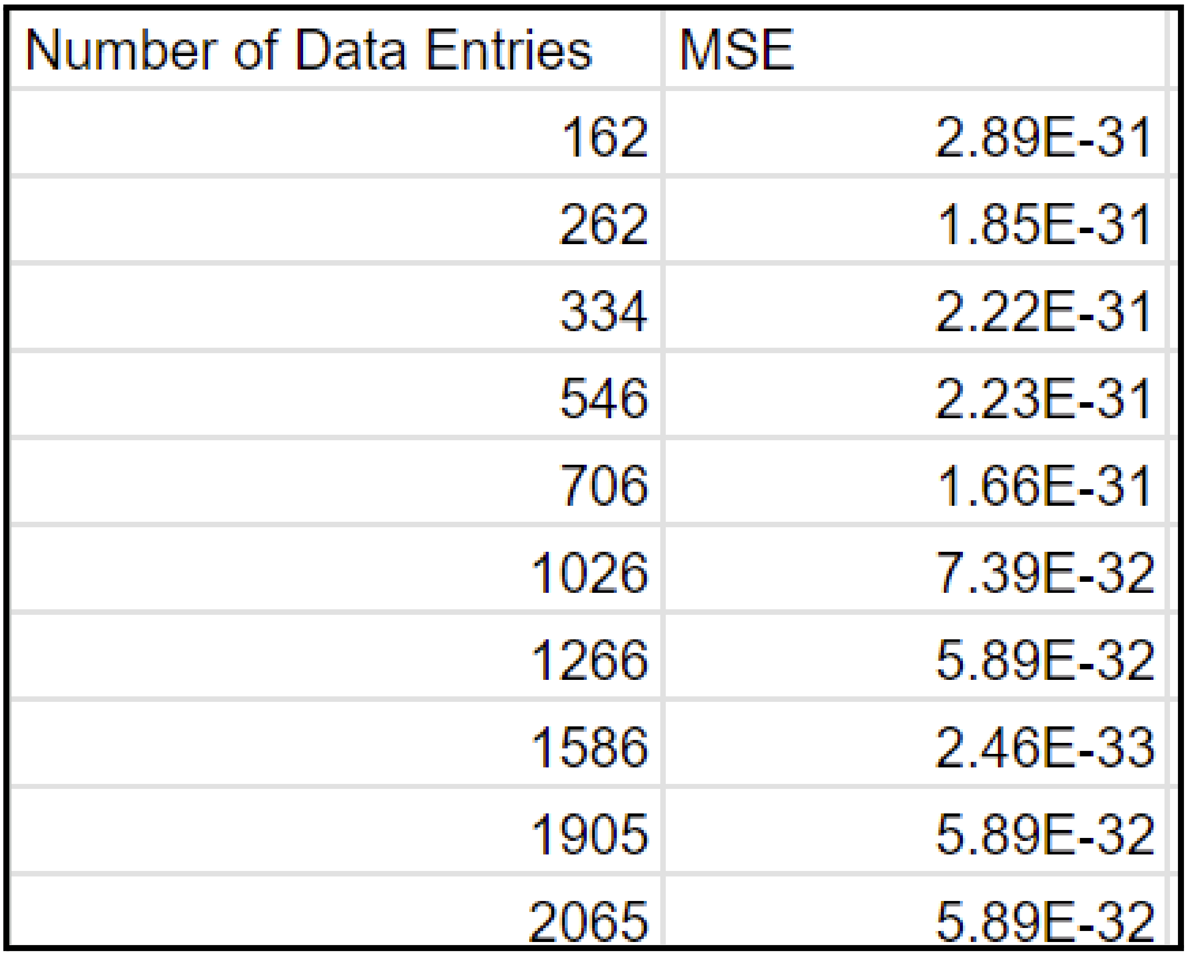




Note: Tables below are the average of three sets of tests with multiple trials. Calculations done by CAML and validated by our calculations done by hand.



Number Of Data Entries	Variance (S^2)
234	0.339
390	0.467
546	0.462
624	0.395
780	0.41
936	0.37
1014	0.382
1267	0.383
1443	0.384
1557	0.377
1776	0.382
2067	0.373



Number of Data Entries	MSE
162	2.89E-31
262	1.85E-31
334	2.22E-31
546	2.23E-31
706	1.66E-31
1026	7.39E-32
1266	5.89E-32
1586	2.46E-33
1905	5.89E-32
2065	5.89E-32

Discussions

Measuring Impact and Future Development

When CAML is implemented in the field, measuring the overall impact will be a critical aspect in determining future potential. If CAML is to serve as a prototype – or even a precedent – for the future development of machine-learning cancer assessments, it is crucial to evaluate the impact of this approach not only in simulated tests but also in the healthcare sector.

One of how the potential of CAML can be measured is through partnerships with local hospitals or medical clinics. With this help, we can develop a more in-depth study that might analyze true predictions versus total predictions or any other data that is relevant to success. We can also measure the overall percent impact by taking the number of true predictions over the total predictions multiplied by 100. These are some examples, but it is important to note that this paper is a very superficial report on CAML, and more research is necessary for implementation.

CONCLUSIONS and FURTHER RESEARCH

Conclusion

Consequences

Every piece of technology has potential benefits and consequences associated with it. CAML holds immense promise in the field of wider patient care and increases the quality of care provided by healthcare professionals. Through the careful analysis of patient medication history and external risk factors contributing to the risk of developing cancer, CAML can greatly increase the chance of either identifying factors that cause cancer or aid in the diagnosis of early-stage cancer. The integration of machine learning in understanding connections between a range of medications and their carcinogenic risks also enables providers to create a wider database of drugs and treatments that could put their patients at risk of developing certain types of cancer. CAML aims to provide easy-to-use software for patients of any ability to utilize to understand how their course of medication impacts their risk of developing certain types of cancer, among other diseases. With an easy-to-understand UI, and the implementation of the largest list of compiled carcinogenic medications to date, CAML provides an extremely reliable source of risk evaluation that can be easily understood by both medical professionals and users. With the use of this software, patients can develop increased awareness of the risks associated with certain medications and can discuss plans with their doctors to prevent dangerous side effects. A patient using this software would be able to understand the risks of carcinogenic exposure or associated risks through each medication, and can also see the type of cancer that may present risk. Once a patient obtains information through this software, they can make informed decisions and discuss with their healthcare provider to gain a deeper knowledge of their medications and pursue alternative treatments or follow up with their provider for cancer tests. Finally, the collection of data from a large user base would allow our machine learning patterns to continue to grow, as it would be provided with real-time developments and information about current users and their experiences with medications.

While models such as CAML provide an abundance of potential benefits, there is also a range of limitations and considerations to be noted before mass-releasing such models. One of the main concerns with CAML would be real-time accuracy and risk assessment. As our algorithm continues to grow off on patient data, it depends entirely on the reliability of patient inputs and professional intervention upon false information. Proper data and readings may not always be available during the early stages of development as a result of these inaccuracies and could pose complications during patient use of CAML. Additionally, the use of such a model to gain an understanding of one's cancer risk may cause misinterpretation. It is important to note that some medications are associated with cancer rather than causing the cancer. For example, insulin is associated with pancreatic cancer due to the condition of diabetes. Essentially, diabetes is what increases the risk of pancreatic cancer, not insulin. Although this is the concept that allows CAML to draw conclusions and link medications to cancers, it may slightly confuse the user. This may unintentionally imply to the user that insulin is a cancer-causing medication, but in reality, it is only associated with it.

Despite these potential consequences of the use of CAML, there are still a great deal of benefits to the development of this software. It would allow future developers to understand the importance of early identification of cancer, and allow patients to get intervention at an appropriate time.

Further Suggestions

Future improvements and development of CAML would be based on the results of the impact measurement research. Future versions of CAML can also improve accuracy by taking in more factors and constructing a more complex decision tree. Additionally, if provided enough funding and proven worthwhile, switching to a Deep Learning algorithm will increase the scalability of CAML allowing it to become a global network of medication marker identification.

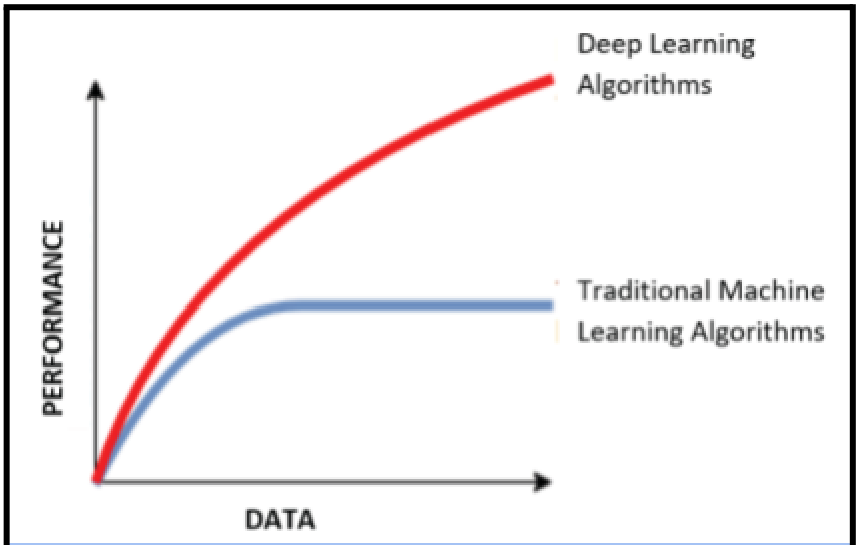


Figure 3: Deep Learning Versus Machine Learning Performance (Lamtougui et al, 2020)

REFERENCES

A, J., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4

Apte, P. P., & Spanos, C. J. (2022, July 13). *Adding More Data Isn't the Only Way to Improve AI*. Harvard Business Review. https://hbr.org/2022/07/adding-more-data-isnt-the-only-way-to-improve-ai

Brooks, D. M. (2023, March 10). *Why AI Is Different Than Any Other Invention or Technology in Human History*. Https://Www.drmikebrooks.com/. https://www.drmikebrooks.com/why-ai-is-different-than-any-other-invention-or-technology-in-human-history/

Cohen, J. B., Brown, N. J., Brown, S.-A., Dent, S., van Dorst, D. C. H., Herrmann, S. M., Lang, N. N., Oudit, G. Y., & Touyz, R. M. (2023). Cancer Therapy–Related Hypertension: A Scientific Statement From the American Heart Association. *Hypertension*, *80*(3). https://doi.org/10.1161/hyp.0000000000000224

De Souza, A., Irfan, K., Masud, F., & Saif, M. W. (2016). Diabetes Type 2 and Pancreatic Cancer: A History Unfolding. *JOP : Journal of the Pancreas*, *17*(2), 144–148. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860818/#:~:text=Pancreatic%20Cancer%20is%20the%20fourth

Information about Nitrosamine Impurities in Medications. (2020). *FDA*. https://www.fda.gov/drugs/drug-safety-and-availability/information-about-nitrosamine-impurities-medications

María Padilla-Ruiz, María Morales-Suárez-Varela, Rivas-Ruiz, F., J. Alcaide-García, Varela‑Moreno, E., Zarcos‑Pedrinaci, I., Téllez, T., Nerea Fernández‑de Larrea‑Baz, Baré, M., Bilbao, A., Sarasqueta, C., Urko Aguirre Larracoechea, Quintana, J. M., & Redondo, M. (2022). Influence of Diagnostic Delay on Survival Rates for Patients with Colorectal Cancer. *International Journal of Environmental Research and Public Health*, *19*(6), 3626–3626. https://doi.org/10.3390/ijerph19063626

Miller, D. D., Cowen, E. W., Nguyen, J. C., McCalmont, T. H., & Fox, L. P. (2010). Melanoma Associated With Long-term Voriconazole Therapy. *Archives of Dermatology*, *146*(3). https://doi.org/10.1001/archdermatol.2009.362

Rana, M., & Bhushan, M. (2022). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-14305-w

Tamoxifen and Uterine Cancer. (2006). Www.acog.org. https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2014/06/tamoxifen-and-uterine-cancer#:~:text=Most%20studies%20have%20found%20that

Why is early cancer diagnosis important? (2023, March 30). Cancer Research UK. https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important/1000

World Health Organization. (2022, February 3). *Cancer*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer

ACKNOWLEDGEMENTS

Since most of our work consisted of online development on Visual Studio Code and PubMed, we worked from home and in the school computer labs. Our school had especially powerful i7-12th generation Intel processors and RTX 3060 graphics processors that we utilized to train our machine-learning model. Although our laptops were able to run the programs, our school computers allowed us to complete work at a faster pace. Our teacher (also our mentor), Mrs. Farha Ahmed, worked to ensure we had all the necessary resources for our development. She provided us with tools, such as educational access to Excel and GitHub. Furthermore, she supervised our testing and experimentation process. Although she is not well-versed in machine learning development, her role as a science teacher allowed her to point out some notable mistakes that we fixed, such as the lack of cross-referencing our tests with the machine's tests and negligence of two control variables: The computer used and the order of entries in the initial database. Although we were not sure what effect these control variables might have, it would not be incorrect to follow her suggestions to prevent any faulty testing.

To receive constructive and critical feedback, we talked to Ibtesam Jamal, a student at the University of Texas A&M. Jamal is pursuing an Honors Mechanical Engineering degree. Jamal has multiple Python certifications, including one in machine learning. He expressed his appreciation for such an innovative approach to predicting cancer diagnosis, but he did recommend certain future improvements. The two main improvements included the security of the website and scalability management.

Next, we talked to a NASA Goddard intern and Purdue University Ph.D. student, Jehan Shalabi, for further recommendations. Her experience in developing and

programming autonomous vehicle systems for an MIT research internship gave her a deeper understanding of our project. She was excited about our idea and also recommended some improvements. One of these improvements was allowing doctors to visualize the MSE and True vs. Prediction chart in order to allow them to figure out if the predictions were accurate. She also highlighted the innovative concept of a "medication marker" our machine learning used.

Finally, we talked to Dr. Murad Alam, a professor at Northwestern University with over 17 years of experience. Dr. Alam has extensive research into many skin cancers, with a h-index of 71. He attained his undergraduate and medical degrees at Yale and Columbia. In addition, he was a member of the Board of Directors of the American Academy of Dermatology. We selected Dr. Alam as our primary source of feedback for these reasons and greatly learned on his criticism. Dr. Alam appreciated the idea and the innovative approach. He also provided us with some improvements/future recommendations. For example, he talked about integrating this idea into the current healthcare sector with medication records and collecting data on different demographics. Dr. Alam also raised concerns over confounding variables, which will need specific real-world testing to solve. He gave us the most feedback out of all the individuals who we asked in a long email.

It is also important to note that we talked to many other people within the fields of computer science, computer engineering, and medicine, both casually and professionally. These people were able to give us much-needed feedback. However, in this essay, we listed the people who gave us the greatest help or had much-needed experience that we did not.

Overall, our team is thankful for all the help we received, and we hope to see the fruition of our idea as this system is implemented in the near future. We are also ready for the multitudes of anticipated challenges in the next steps of implementation, and will persevere with one primary goal in mind: To save lives and give back to society.