Fall 2019

Toward Early Detection Of Pancreatic Cancer: An Evidence-Based Approach

Omid Sharagi
San Jose State University

Computer
Science

Omid Sharghi

has passed the defense for the project

TOWARD EARLY DETECTION OF PANCREATIC CANCER: AN EVIDENTIAL BASED APPROACH

Advisor's Signature Leonard P. Wesley — Date 12/9/2019

Committee Member's Signature Skyler Payne — Date 12/9/2019

Committee Member's Signature Philip Heller — Date 12/9/19

NOTE: The advisor should send the final report to the Graduate Coordinator so that the student can be cleared for graduation

San José State
UNIVERSITY

Toward Early Detection Of Pancreatic Cancer: An Evidence-Based Approach

A project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfilment

Of the Requirements for the

Degree Master of Science

By

Omid Sharghi

December 2019

The Designated Project Committee Approves the Project Titled

Toward Early Detection Of Pancreatic Cancer: An Evidence-Based Approach
by
Omid Sharagi

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE
San José State University
December 2019

Dr. Leonard P. Wesley, Department of Computer Science
Dr. Philip Heller, Department of Computer Science
Dr. Skyler Payne, Linkedin

ABSTRACT

This study observes how an evidential reasoning approach can be used as a diagnostic tool for early detection of pancreatic cancer. The evidential reasoning model combines the output of a linear Support Vector Classifier (SVC) with factors such as smoking history, health history, biopsy location, NGS technology used, and more to predict the likelihood of the disease. The SVC was trained using genomic data of pancreatic cancer patients derived from the National Cancer Institute (NIH) Genomic Data Commons (GDC). To test the evidential reasoning model, a variety of synthetic data was compiled to test the impact of combinations of different factors. Through experimentation, we monitored how the evidential interval for pancreatic cancer fluctuated based on the inputs that were provided. We observed how the pancreatic cancer evidential interval increased and the machine learning prediction of pancreatic cancer was supported when the input changed from a non-smoker and non-drinker to an individual with a highly active smoking and drinking history. Similarly, we observed how the evidential interval for pancreatic cancer increased significantly when the machine learning prediction for pancreatic cancer was maintained as high and the input of the quality of the sequencing read was changed from a high quantity of cytosine guanine content and homopolymer regions to a moderate quantity of cytosine guanine content and low homopolymer regions; indicating that there was initially a higher likelihood of error in the sequencing reads, resulting in a more inaccurate machine learning output. This experiment shows that an evidence-based approach has the potential to contribute as a diagnostic tool for screening for high-risk groups. Future work should focus on improving the machine learning model by using a larger pancreatic cancer genomic database. Next steps will involve programmatically analyzing real sequencing reads for irregular guanine cytosine content and high homopolymer regions.

Key Words: Next Generation Sequencing, Mutations, Homopolymer Regions, Pancreatic Cancer, Evidential Reasoning, Support Vector Classifier.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

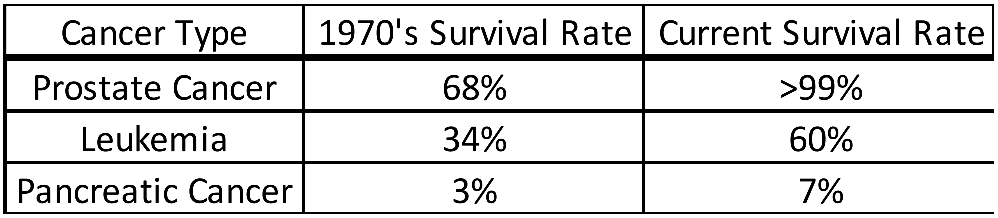
Pancreatic cancer is an aggressive form of cancer that begins localized to the pancreas, but due to lack of symptoms and limited detection options, the disease goes undetected and spreads to other organs. Because tumors spread while remaining undetected and treatment options are limited, pancreatic cancer has the lowest survival rate of all cancers, with a 5-year survival rate of ~7% (Gharibi, Adamian, & Kelber, 2016). Of all the types of pancreatic cancer, pancreatic ductal adenocarcinoma (PDAC) is not only the most common, but also the most aggressive and ranks as fourth in most cancer related deaths (Gharibi, Adamian, & Kelber, 2016).

PDAC is a complex form of pancreatic cancer with an average of 63 mutations per tumor (Amin & DiMaio, 2016). PDAC forms as precursor lesions, known as pancreatic intraepithelial neoplasia (PanIN). The source of PDAC is the cells of the pancreatic duct (Stark & Eibl, 2015). Genetic mutations in the ductal epithelium are the source of the development of the precursor lesions. The most scrutinized source is the oncogenic KRAS gene. When the tumor suppressor genes CDKN2A, TP53, and SMAD4 are deactivated, the ductal epithelium undergoes drastic transformations causing lesions to worsen to grades 2 (G2) and 3 (G3); PDAC grades G1 means the cancer looks similar to the healthy surrounding pancreatic tissue, G3 means the cancer tissue looks very abnormal, and G2 lies somewhere between G1 and G3 (Isaji, et al., 2018) (Amin & DiMaio, 2016).

Among new cases of PDAC, 51% and 49% are estimated to impact men and women (Gharibi, Adamian, & Kelber, 2016). It is estimated that 83% of PDAC cases end in death (Siegel, Miller, & Jemal, 2015). Age and ethnicity also play a hand in the likelihood of developing PDAC. Nearly 27% of all new diagnoses are in the age range of 75 and 84 and 9% are in the range of 45 and 54 years old (Howlader, et al., 2016). Among ethnicities, African Americans are at highest risk of developing PDAC with a probability of 15.7 out of 100,000, while Asian-Americans have the lowest probability of 9.8 out of 100,000 (Howlader, et al., 2016).

Unfortunately, while the development and morbidity rates of other cancers decline, pancreatic cancer continues to rise. Detecting pancreatic cancer early, while the tumors have not metastasized, dramatically increases the likelihood of survival. Unfortunately, only 9% of cases are detected during the early stage (Howlader, et al., 2016). Even when diagnosed early, the 5-year survival rate is merely 26%, while most other cancers have a much higher 5-year survival rate (Bachmann, Michalski, Martignoni, Büchler, & Friess, 2006). For example, the 5-year survival rate for breast and prostate cancer is 99%, 92% for the kidney, 90% for colon and rectum, 83% for oral cavity and pharynx, 64% for stomach, and 54% for lung and bronchus (Gharibi, Adamian, & Kelber, 2016). Once pancreatic spreads, the 5-year survival rate drops down to 10%, however, tumor detection rises to 28%. Once the tumors reach more distant organs, the 5-year survival rate drops down to 2%, while the likelihood of detection rises to 53% (Howlader, et al., 2016). Over the past 40 years, 5-year survival rates for a variety of cancers have increased significantly, while pancreatic cancer has remained nearly the same, as shown in the Table I (Gharibi, Adamian, & Kelber, 2016).

Table I
Past Vs. Present Survival Rates of Various Cancers



Cancer Type	1970's Survival Rate	Current Survival Rate
Prostate Cancer	68%	>99%
Leukemia	34%	60%
Pancreatic Cancer	3%	7%

When tumors are still localized to the pancreas, surgery still provides the highest chance of survival. In 2016 the reported median survival rate was 14-20 months and the 5-year survival rate was up to 25% post-resection of the pancreas (Gharibi, Adamian, & Kelber, 2016). Because of the unassuming nature of symptoms associated with pancreatic cancer, tumors metastasize by the time they are detected through imaging. Currently, diagnostic biomarkers are the focus of researchers as a means of detecting the cancer in its early stages. The CA19-9 antigen is considered the best pancreatic cancer biomarker; however, CA19-9 is not specific to pancreatic cancer, and is known to result in false positives because it is detected in both benign and malignant tumors (Gharibi, Adamian, & Kelber, 2016). Also, 10% of the population cannot produce the protein (Gharibi, Adamian, & Kelber, 2016). Finding a novel biomarker that is unique to pancreatic cancer would be a significant advancement in early detection.

BACKGROUND

The difficulty in detecting pancreatic cancer is primarily due to the lack of symptoms. Even in later stages, symptoms can be unassuming and easily confused with other illnesses. Early symptoms of pancreatic cancer include weight loss, a lack of appetite, abdominal pain, and jaundice (see Appendix A for symptom details). When pancreatic cancer is suspected, medical imaging is conducted, leading to discovery of masses. Unfortunately, cancerous masses that are detectable via imaging are most often at a late stage, stressing the importance of new methods that result in earlier detection (see Appendix B for imaging technology details).

One strategy in detecting pancreatic cancer earlier is the use of biomarkers associated with the disease (see Appendix C for additional biomarker details). A significant hurdle in studying biomarkers for early detection is the lack of samples collected before diagnosis. While samples from healthy controls or patients diagnosed with pancreatic cancer are helpful, samples collected from those prior to a pancreatic cancer diagnosis are preferred (O'Brien, et al., 2015).

At the moment, the most commonly used biomarker associated with pancreatic cancer is carbohydrate antigen 19-9 (CA19-9). CA19-9 has a sensitivity that varies between 69%-98% and a specificity between 46%-98%. In a study using the serum samples of women participating in an ovarian cancer study, but who were later diagnosed with pancreatic cancer, O'Brien et al. found strong evidence that CA19-9, as well as the other proteins CA125, CEACAM1, and REG3A could play a significant part in early pancreatic cancer diagnosis. In the study, serum samples that were collected < 12 months before diagnosis, CA19-9 was found to have a median level of 43.2 U/mL compared to the control sample median of 3.1 U/mL (O'Brien, et al., 2015). O'Brien et al. grouped their data sets by time to diagnosis and found

CA19-9 to be significantly higher in PDAC cases when compared to the control in the 0-0.5, 1-2, and 2-3, and 3+ year groups (O'Brien, et al., 2015). While most CA19-9 levels were most significant within 12 months of diagnosis, O’Brien et al. found two cases with adequate longitudinal samples where an increase in CA19-9 levels was detectable 3 years prior to diagnosis of PDAC (O'Brien, et al., 2015).

One of the issues with CA19-9 is that it can be found in benign and malignant gastrointestinal tumors, making it not specific to pancreatic cancer. Another issue with CA 19-9 is that nearly 10% of the population has trouble producing the protein (Gharibi, Adamian, & Kelber, 2016). However, a certain level of CA 19-9 detected in a blood sample could aid in the diagnosis of pancreatic cancer. Because CA 19-9 is not tumor specific, it is not enough of a tool to be used alone to confirm diagnosis of pancreatic cancer. CA 19-9 is predominantly used in the management of pancreatic cancer post diagnosis, due to its correlation to the disease, however research in the levels of CA 19-9 in the years leading up to the disease show promise (O'Brien, et al., 2015).

Although progress is being made in diagnosing pancreatic cancer a year or two in advance, it is believed that it takes well over a decade before pancreatic cancer fully develops (Gharibi, Adamian, & Kelber, 2016), providing a large window for early detection. Finding prevalent and precise biomarkers is a challenge and many researchers are studying new ways to detect these proteins. Because CA19-9 alone is not a strong enough indicator of pancreatic cancer, detection could be improved by coupling CA19-9 with other proteins that could be obtained through serum samples.

Discovering New Biomarkers

Proteomics is a methodology used to find new biomarkers because of its ability to measure protein and post-translational changes in protein levels. Gharibi et al. cites a study where protein levels were derived from cells extracted from primary and metastatic tumors sites. Using proteomics, the study found 547 proteins that were unique to the primary tumor site and 487 unique to the metastatic tumors. Of these proteins, 134 were found to have significantly higher levels between the two sites. Focusing on these proteins may provide insight on metastasis and could be useful in research in developing new therapies. One of the challenges of proteomics is separating the cancerous tissue from the surrounding microenvironment.

Secretomics, a division of proteomics, is the analysis of protein secretions of cells. Gharibi et al. discusses one study that analyzed the secretome of pancreatic cells and found 145 secreted proteins that were 1.5 times more upregulated in pancreatic cancer cells than healthy cells. Another secretomic study used mass spectrometry to find that glypican-1 (GPC1), a cell-surface protein, is significantly more abundant on pancreatic cancer derived exosomes than those of a healthy control. The study compared GPC-1 to CA 19-9 as a biomarker and found that CA19-9 had high levels in both benign pancreatic related diseases and PDAC, while GPC-1 exosome levels were high in only PDAC cases (Gharibi, Adamian, & Kelber, 2016) (See Appendix D for additional details regarding discovering new biomarkers).

Circulating Tumor Cells

Circulating tumor cells are another potential biomarker that have garnered attention from researchers. One study aimed to identify the KRAS mutation in CTCs derived from pancreatic cancer patients. Researchers were able to identify mutations of the

KRAS gene in CTCs from 11 out of 12 pancreatic cancer patients, while not finding any mutant KRAS sequences in the hematopoietic cells from the same set of patients (Court, et al., 2016). Court et al. came to the conclusion that a minimum of 10 CTCs is necessary to conclude KRAS mutations are present. When <10 CTCs were acquired, they found a significant drop in the detection of KRAS mutations. Using Sanger sequencing to emulate clinical sequencing, Court et al. found that WGA (whole genome amplification) was responsible for most sequencing cases that resulted in failure and found an ADO (allele drop-out) rate of 85% (Court, et al., 2016). However, detecting CTCs in the bloodstream is challenging because of the low level of CTCs in blood samples. It is believed that a range of 1-50 CTCs is contained in a 7.5 mL blood sample. A 7.5 mL blood sample is believed to have more than a million white blood cells, requiring tests to have very high sensitivity and specificity in order to recognize CTCs (Court, et al., 2016). Studies have found one CTC cell per 10^9 hemocytes in a blood sample of a cancer patient (Qi, et al., 2018). Of all the CTCs that break away from the primary tumor, it is believed that .01% form into metastases. However, CTCs still have the potential of being excellent biomarkers of cancer progression and novel techniques are being developed to detect these cells (see Appendix E for additional CTC details).

Genomic Biomarkers

While detecting CTCs successfully depends on the level of CTCs in a sample, serum samples can potentially give researchers access to a battery of different cancerous cellular samples such as cell-free nucleic acid (cfNA), which includes cell-free DNA (cfDNA) and cell-free RNA (cfRNA). These cfNAs are rich with useful information regarding somatic mutations, cancer associated DNA methylation, and more. A study in 1998 found that cfDNA circulating in blood varied between healthy controls and those diagnosed with pancreatic cancer (Giacona, et al., 1998). The concentration of cfDNA was found to vary between different types of pancreatic cancer and varied at different stages of the same cancer (Qi, et al., 2018). Mutations of genes associated with pancreatic cancer could also be seen in the cfDNA. One study analyzed 54 genes that were common in PDAC patients and found that 90% of mutations associated with the genes found in the biopsies of tumors were also seen in the cfDNA, leaving researchers to believe that cfDNA could be detected with high specificity and sensitivity (Zill, et al., 2015). Detection of KRAS mutations have been found in the plasma of nearly 50% of PDAC patients, while almost no mutations were found in the plasma of healthy donors, leading researchers to believe KRAS mutations in ctDNA to be a potential viable biomarker (Qi, et al., 2018). One type of cfNA, miRNA (microRNA) has the specificity and sensitivity to be a pancreatic cancer biomarker. One study found that by combining miR-196a and miR-196b, a sensitivity of 100% and specificity of 90% were achieved in detecting pancreatic cancer. Another study found the saliva sample of pancreatic cancer patients to have specific miRNA that were notably upregulated while another study found stool samples of PDAC patients to have higher levels of specific miRNA compared to those of a normal control (Qi, et al., 2018) (see Appendix F for additional details regarding genomic biomarkers).

Sequencing

Sequencing technology is commonly utilized to classify gene mutations. Following sequencing, techniques such as Mutation Significance of Covariance (MutSigCV) are performed to discover mutations (Gharibi, Adamian, & Kelber, 2016). Analyzing pancreatic tumor cells with MutSigCV resulted in 24 notably mutated genes that were found in >3.5% of cases. While some of these mutations were already known to be associated with PDAC, several novel genes were also identified (Gharibi, Adamian, & Kelber, 2016). While highly useful, the technology used to sequence DNA can be imprecise if the DNA being analyzed consists of homopolymer regions and irregular guanine/cytosine (GC) content (Yeo, et al., 2012) (Benjamini & Speed, 2012) (see Appendix G for additional sequencing details).

Machine Learning Applications

The advances in computing power and access to cancer related genomic data have led to new studies involving data analysis of genomic data and other biomarkers to aid in the classification of different types of cancer. Way et al. investigated a machine learning approach attempting to classify genes involved in the activation of the Ras pathway (Way, et al., 2018). The classifier used in their experiment had an area under the receiver operating characteristic curve (AUROC) greater than 84% and an area under the precision recall curve (AUPR) greater than 63%. When testing the classifier on non-training data, the predictions resulted in a AUROC of 75.2% and a AUPR of 24.7%, indicating that the classifier was able to classify Ras activation signals without being exposed to the associated tissues during training (Way, et al., 2018). Way et al. also attempted to classify Ras mutations by applying their classifier to the RNA sequences of 737 cell lines obtained from the Cancer Cell Line Encyclopedia. Their classifier resulted in 357 out of 393 cell lines being correctly classified as wild-type RAS and 153 out of 344 cell lines being correctly classified as mutated Ras (See Appendix H for additional details).

Remaining Technical Gaps

The capability to predict a disease as complicated as pancreatic cancer may depend on additional factors besides a precise and reliant biomarker. These additional factors could be diverse and derived from various sources. Such information is likely to be inaccurate, incomplete and imprecise to varying degrees. Statistical data by nature can be flawed due to biases or subjective focus of those in charge of collecting the data. Currently, demographics of those at highest risk are based on age and ethnicity, resulting in a broad population of individuals, where only a small subset will eventually be diagnosed. Besides age and ethnicity, statistical data does not provide precision to diagnose a patient in such a wide population. Also, because the disease is diagnosed late, there is limited statistical data of patients collected prior to diagnosis, leaving insufficient useful data to work with. As a result, a sophisticated mathematical calculus is needed to represent and reason from imperfect information. The belief function (BF) and evidential reasoning calculi facilitates representing and reasoning from imperfect and diverse information in a way that is more flexible and less limiting than traditional probabilistic and statistical methods (Lowrance, et al., 1991). Belief functions and evidential reasoning can be regarded as a generalization of traditional probabilistic methods. Information that has the potential of being imperfect yet still useful for a pancreatic cancer prediction could include biopsy location, sequencing technology used, an individual’s habits such as diet and smoking, the amount of genetic material collected and more. The work described in this report begins to bridge this technical gap by demonstrating the viability of an evidence-based approach toward representing and reasoning from diverse and imperfect pancreatic cancer related data and information.

APPROACH

As reported by Gharibi et al., it is believed that pancreatic cancer develops in the body for nearly 10 years before symptoms present themselves. However, current diagnostic methods are not designed for early detection that might result in prompt treatment that increases the probability of survival. Even though the genes and mutations correlated with the disease are known, there currently are no serum based diagnostic methods that detect these mutations with sufficient diagnostic fidelity. The capability to precisely detect mutation combinations early might provide sufficient time for effective treatments to be applied. Individuals who are at higher risk for developing pancreatic cancer, whether through genetic disposition or through risk factors such as smoking or drinking, currently have no options for screening to detect the disease before metastasis. With more cancer related genomic data becoming available, data analysis techniques of predicting cancer is a growing area of interest. However, most genomic data analysis approaches rely solely on the output of a machine learning algorithm to make a final classification or prediction. The experiment conducted by Way et al. had fruitful results but did not consider other crucial evidence that could help support or challenge results (Way, et al., 2018). The approach undertaken in this project is to use the output of a machine learning model as one input to be combined with other factors, such as family history and health-risk habits, before making a final decision. This approach of combining evidence is especially useful when using imperfect data that is often irregularly distributed, consists of uneven class cases, and incomplete, which is often the case with medical related data. One factor that can be used is an individual's smoking history. An individual with a long smoking history is five to six times more likely to develop pancreatic cancer (Pandol, Apte, WIlson, Gukovskaya, & Edderkaoui, 2012). Evidence with such a strong correlation to pancreatic cancer could be used to tilt an impartial machine learning result towards a higher likelihood of pancreatic cancer. Besides smoking history, we can also consider drinking history, which similarly has been reported to result in a 1.5 to 6-fold increased risk in pancreatic cancer based on dosage and extent (Gupta, Wang, Holly, & Bracci, 2010). Because we are using genomic data, the error rates of the sequencing technology used to obtain the genomic information can also be used as a factor. As mentioned earlier, if an individual's DNA SNP consists of homopolymer regions, there is an increased likelihood in an error during the sequencing process that could impact the accuracy of the data used to train the machine learning model. This experiment uses a mathematical calculus that can combine diverse factors, like those described above, and relate them to one another to help us look beyond the traditional statistical and probabilistic techniques and potentially develop a new diagnostic method that is both precise and timely. A variety of experiments will be conducted to help us evaluate a null hypothesis.

METHOD

The mathematical calculus used in this study to combine and weigh diverse and imperfect factors before making a final decision is known as evidential reasoning. The evidential reasoning approach does not depend on a single source of data, but rather assigns belief to different factors to make a classification on the likelihood of pancreatic cancer. This approach is valuable because the statistical data needed to make a classification is often imprecise and scarce. This experiment utilizes factors such as an individual's family history, smoking history, drinking history, results of a machine learning classifier, sequencing read, the type of NGS technology used to obtain an individual's genomic data, biopsy location, and amount of genetic material as input to an evidential reasoning model that could output a prediction regarding an individual's likelihood of pancreatic cancer based on the assigned beliefs to the factors. The pancreatic cancer evidential reasoning prediction model used in this study is based on a network of relationships between inputs that narrow down to a final pancreatic cancer prediction, as shown in Figure 1. Factors, or random variables, such as an individual's smoking history are defined as frames consisting of propositions that are meant to delimit all possible situations in which only one can be true at a time. These propositions could be discrete or continuous values. For instance, a frame representing smoking history could have the following propositions: low, medium, or high usage. Our smoking history frame should contain all these possibilities, along with the possibility of not knowing what an individual's smoking history is at all, which would be represented as the disjunction of all these possibilities (Low $\vee$ Medium $\vee$ High). We would build similar frames for each random variable (ML prediction, drinking history, etc.) we intend to use as input into our evidential reasoning model. The propositions used in this experiment are synthesized based on data found in the NIH GDC dataset. In evidential reasoning, a knowledge source assigns probabilities to propositions, contained in frames. The assigned probabilities express the truth of the statement. The collection of frames is called a gallery and describes what is possible. To jointly consider proposition statements from two distinct frames, a compatibility relation, which declares which propositions from two frames can be true at the same time, must be defined. The compatibility relation consists of a subset of the cross product between two frames. Using Dempster's Rule of Combination, frames are fused together to form a new body of evidence, which could be fused with other bodies of evidence (Yager, Liu, Dempster, & Shafter, 2008). The result of fusing two frames is propagated through the model, all the way to the final frame (see Appendix I for more evidential reasoning details). Because one of our frames is a machine learning prediction of pancreatic cancer based on an individual's genomic data, we tested whether a machine learning algorithm could be trained to accurately and consistently classify pancreatic cancer using a dataset of genes and mutations associated with the disease. Genomic data was acquired through the NIH GDC API and manipulated before training a linear support vector classifier (SVC) (see Appendix J for additional details regarding how the machine learning model was built).

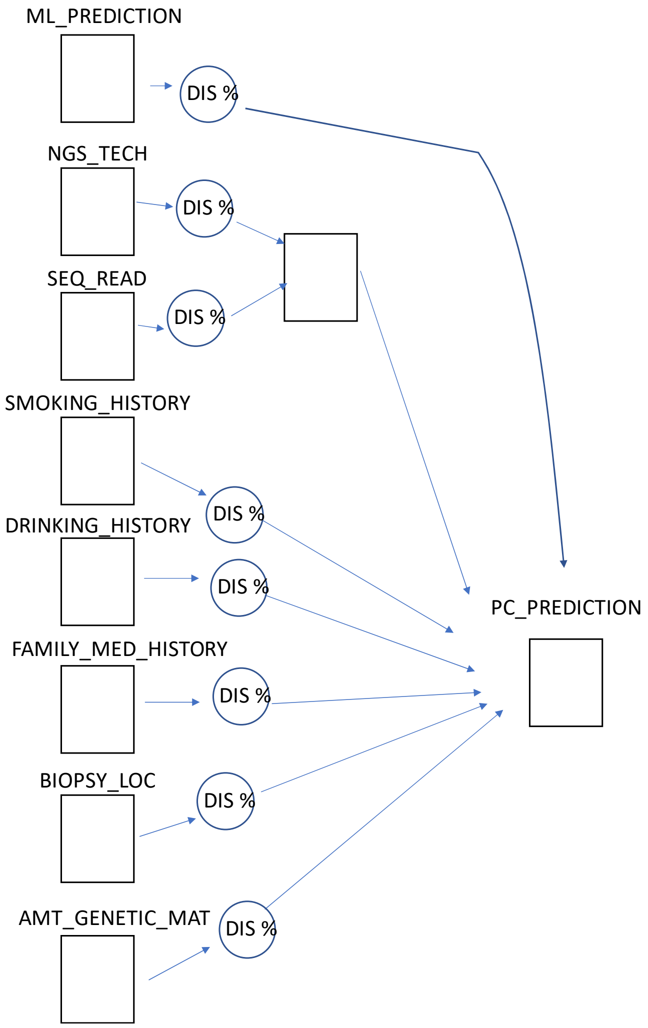


Figure 1. Evidential Reasoning Model

EXPERIMENTS

The primary source of data used for this experiment was the National Cancer Institute (NIH) Genomic Data Commons (GDC). The pancreatic project in the GDC consists of 185 cases of which 100 cases have been confirmed as deceased. The average number of days to death following diagnosis was 459, with a standard deviation of 362. The longest an individual lived following diagnosis was 2182 days and the shortest was 12 days. The experimentation in this project uses the following hypotheses:

H_o: Will not be able to detect pancreatic cancer significantly earlier than currently possible

H_A: Will be able to detect pancreatic cancer significantly earlier than currently possible

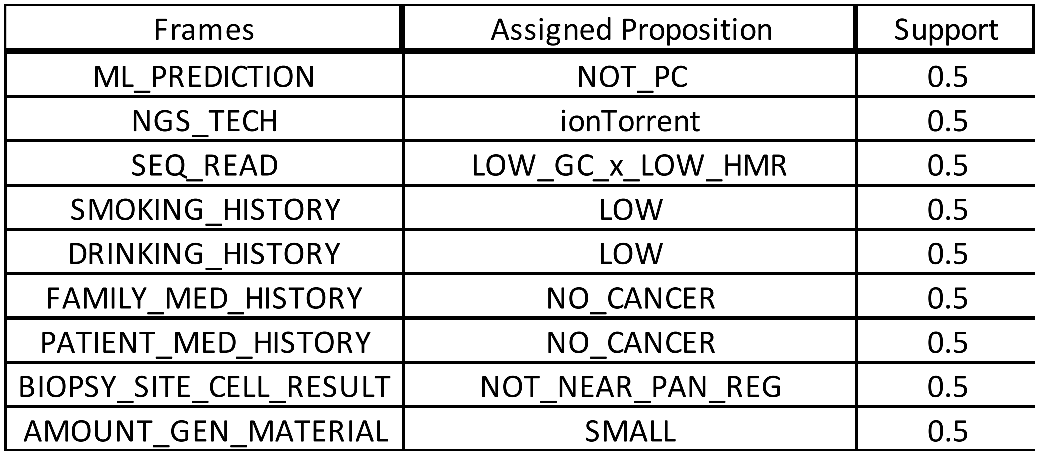
Using the NIH GDC dataset consisting of cases that did not survive the disease, an estimated sample size can be calculated with the assumption that the survival rate can be doubled. Using a significance level (α) of .05 and a power of .80, we calculate that a sample size of 6.96 will be necessary to test H_o:

$$n = [(z_\alpha + z_\beta)(\sigma) / (\mu_1 - \mu_2)]^2$$

$$n = [(1.645 + 0.84)(362) / (-459 +800)]^2$$

Our evidential reasoning model will be tested by observing how the results vary as the propositions and their corresponding masses are changed. Depending on the selected propositions and assigned masses, some combinations will build support for a prediction of pancreatic cancer, while a combination of other frames will decrease the likelihood of a pancreatic cancer prediction. For instance, combinations such as a highly active smoking and drinking history and a history of cancer in one's personal and family medical history are expected to increase the likelihood of a pancreatic cancer diagnosis. Another example would be a biopsy with a low amount of genetic material taken from a region far from the pancreas is expected to lower the likelihood of a pancreatic cancer diagnosis. The output of the evidential reasoning experiment will be judged based on how the evidential interval supports the expected results of such scenarios. To observe how the output of the evidential reasoning model changes, separate experiments are conducted to demonstrate the impact of the combination of specific inputs. All experiments will begin with the baseline input and discount rates shown in Table II and Table III.

Table II
Baseline Propositions and Corresponding Support



Frames	Assigned Proposition	Support
ML_PREDICTION	NOT_PC	0.5
NGS_TECH	ionTorrent	0.5
SEQ_READ	LOW_GC_x_LOW_HMR	0.5
SMOKING_HISTORY	LOW	0.5
DRINKING_HISTORY	LOW	0.5
FAMILY_MED_HISTORY	NO_CANCER	0.5
PATIENT_MED_HISTORY	NO_CANCER	0.5
BIOPSY_SITE_CELL_RESULT	NOT_NEAR_PAN_REG	0.5
AMOUNT_GEN_MATERIAL	SMALL	0.5

Table III
Discount Rates and Corresponding Frames

Frames	Discount Rate
ML_PREDICTION	0.1
NGS_TECH	0.1
SEQ_READ	0.1
SMOKING_HISTORY	0.3
DRINKING_HISTORY	0.3
FAMILY_MED_HISTORY	0.2
PATIENT_MED_HISTORY	0.1
BIOPSY_SITE_CELL_RESULT	0.2
AMOUNT_GEN_MATERIAL	0.1

A discount rate is applied to each frame to either reduce the impact of that frame or to express a lack of credibility. For instance, the NGS technology used or the amount of genetic material are more conclusive compared to an individual's smoking and drinking history, which are dependent on the patient's credibility. Therefore, a higher discount rate is applied to the smoking and drinking history frames. The combination of our discount rate and initial inputs results in the following baseline output:

```
Belief Of Having Pancreatic Cancer Lies Between:        (0.064, 0.142) (0)|**--------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.857, 0.935) (0)|--------**|(1)
```

Such an output shows a very high likelihood of the individual not having pancreatic cancer. The evidential reasoning experiments are designed to test how the following combination of random variables impact the prediction of pancreatic cancer:

- Machine learning prediction, smoking history and drinking history
- Machine learning prediction, family health history, and personal health history

- Machine learning prediction, biopsy location, and amount of genetic material
- Machine learning prediction, sequencing technology utilized, and quality of sequencing read

EVALUATION OF RESULTS

First, the performance of the SVC classifier was determined. To gauge variance, the SVC classifier underwent a cross validation of 5 and resulted in an accuracy of 91%. The classifier had an average precision score of 92% and a ROC AUC score of 92% as shown in Figure 2. and Figure 3.

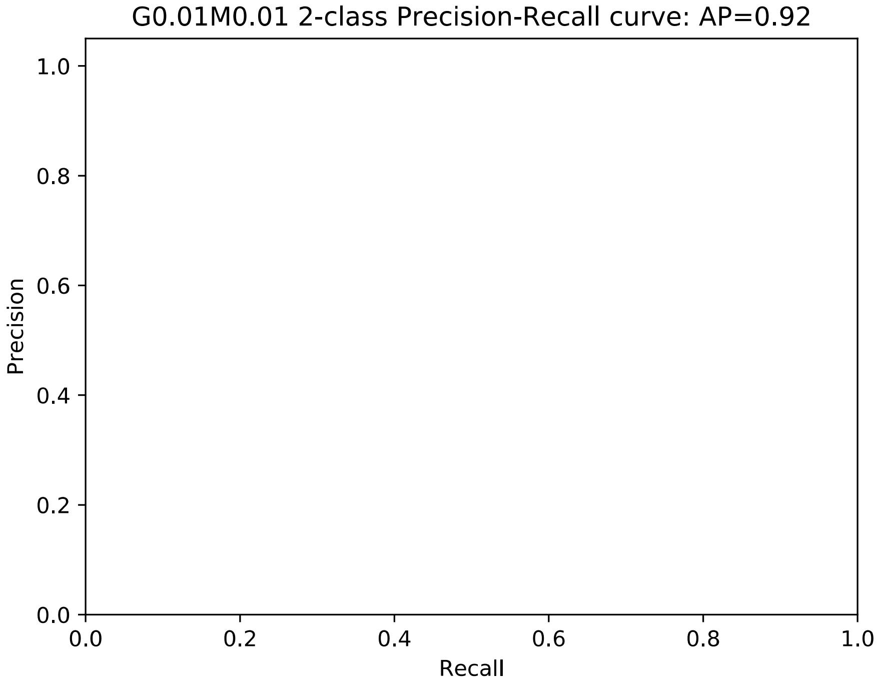


Figure 2. Support Vector Classifier Precision-Recall

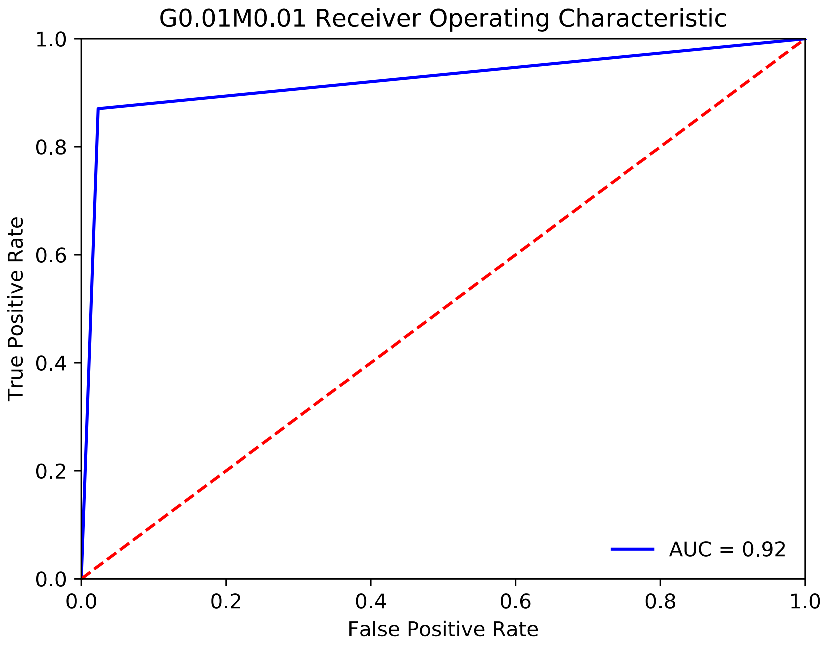


Figure 3. Support Vector Classifier ROC

While our machine learning model indicates good performance, the true performance of this classifier remains questionable due to the limited dataset size of 185 cases. When plotting the frequency of the gene mutation combinations labeled to be pancreatic cancer, we can see a significant amount of gene mutation combinations that only appear once and then a significant drop off, as demonstrated in Figure 4. This stresses the importance of utilizing a larger dataset so that more common gene mutation combinations can be found.

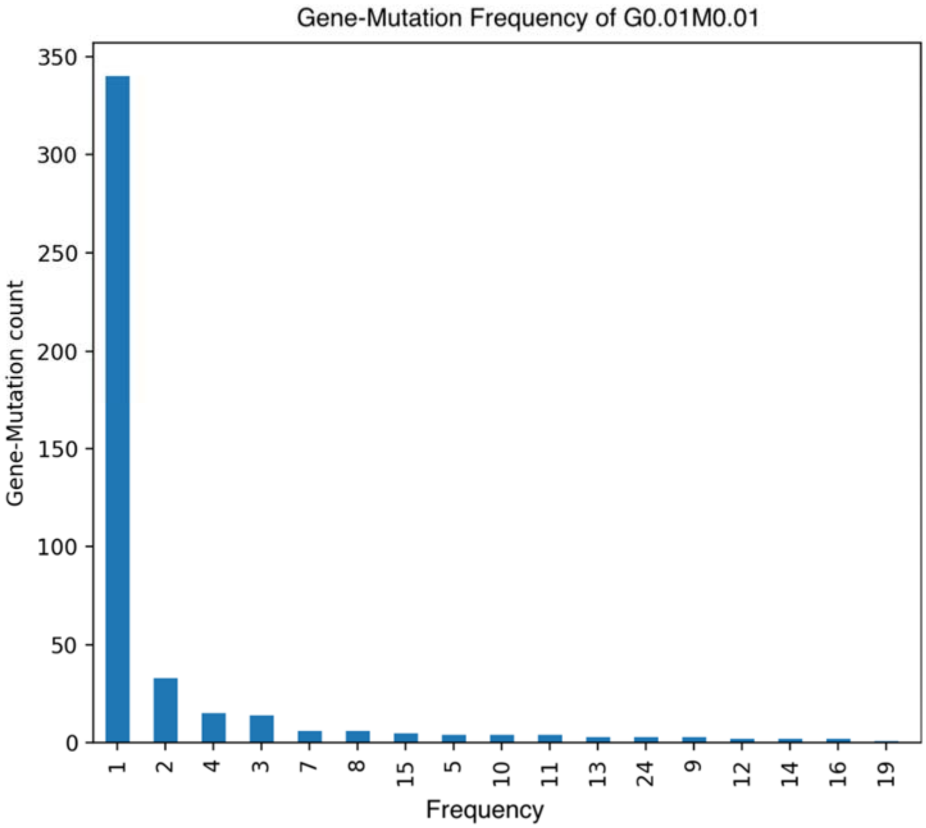


Figure 4. Dataset Gene-Mutation Frequency

ER Experiment 1

The first experiment focuses on the ML (machine learning) prediction and drinking and smoking history frames. The frames were adjusted as follows:

- Change ML Prediction from NOT_PC to PC with a mass of 0.5

```
Belief Of Having Pancreatic Cancer Lies Between:        (0.26, 0.375) (0)|--**------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.625, 0.739) (0)|------**--|(1)
```

- Increase ML Prediction mass to 0.9

```
Belief Of Having Pancreatic Cancer Lies Between:        (0.568, 0.639) (0)|-----**---|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.361, 0.431) (0)|---**-----|(1)
```

- Change Smoking History frame proposition to HIGH with a mass of 0.5

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.764, 0.822) (0)|-------**-|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.178, 0.235) (0)|-**-------|(1)
```

- Change Drinking History frame proposition to HIGH with a mass of 0.5

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.888, 0.929) (0)|--------**|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.07, 0.111) (0)|**--------|(1)
```

- Change Smoking and Drinking History frame proposition to LOW with a mass of 0.9

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.297, 0.333) (0)|--**------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.666, 0.702) (0)|------**--|(1)
```

- Decrease ML Prediction frame proposition mass to 0.2:

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.056, 0.103) (0)|**--------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.896, 0.943) (0)|--------**|(1)
```

This experiment demonstrates that an increased belief in an individual having a highly active smoking and drinking history supports the ML prediction of pancreatic cancer and increases the evidential interval of belief in an individual having pancreatic cancer. A sharp decline in the pancreatic cancer evidential interval can be seen as the smoking and drinking history frame propositions are changed from HIGH to LOW with a belief of 0.9, reducing support for the ML prediction.

ER Experiment 2

The second experiment aims to demonstrate how the family and personal health history frames are correlated to the ML prediction frame. The experiment begins with the original baseline values and makes the follow changes:

- Baseline inputs

  ```
  Belief Of Having Pancreatic Cancer Lies Between:         (0.064, 0.142) (0)|**--------|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between:  (0.857, 0.935) (0)|--------**|(1)
  ```

- Keep ML prediction frame proposition as NOT_PC with a mass of 0.5 and change family and personal history frame propositions to CANCER with a mass of 0.8

  ```
  Belief Of Having Pancreatic Cancer Lies Between:      (0.361, 0.45) (0)|---**-----|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between:  (0.549, 0.638) (0)|-----**---|(1)
  ```

- Change ML prediction frame proposition to PC with a mass of 0.5

  ```
  Belief Of Having Pancreatic Cancer Lies Between:        (0.674, 0.758) (0)|------**--|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between:  (0.242, 0.325) (0)|--**------|(1)
  ```

- Change mass in patient and family health history frames to 0.2

  ```
  Belief Of Having Pancreatic Cancer Lies Between:      (0.427, 0.573) (0)|----**----|(1)

  Belief Of Not Having Pancreatic Cancer Lies Between:  (0.427, 0.572) (0)|----**----|(1)
  ```

- Change propositions for patient and family history frames to NO CANCER with a mass of 0.8

  ```
  Belief Of Having Pancreatic Cancer Lies Between:      (0.174, 0.252) (0)|-**-------|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between:  (0.748, 0.825) (0)|-------**-|(1)
  ```

As expected, changing the frames of personal and family health history to CANCER with an increased mass increases the likelihood of cancer. However, the likelihood of cancer is still low with the ML prediction proposition set to NO PC, even when the patient has a personal and family history of CANCER with a high mass of 0.8. As the ML prediction frame is adjusted to support the health history frames, the pancreatic cancer evidential interval increases from (0.361, 0.45) to (0.674, 0.758). Similarly, as the mass for the health history frames is reduced, our pancreatic cancer evidential interval also decreases. Once the personal and family health history frames are both set to NO CANCER with a mass of 0.8, the pancreatic cancer evidential interval decreases even more.

ER Experiment 3

In experiment 3 attention is shifted to the biopsy site and amount of genetic material frames, which impact the quality of the ML prediction. If the amount of genetic material is low, then there is a greater likelihood that there is not enough DNA material to detect the mutations that are needed to make a credible machine learning decision. Also, if the biopsy location originates from a region distant from the pancreas, there might not be enough pancreas related DNA material for the SVC to classify. The experiment begins with the same baseline as the other experiments:

- Baseline input

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.063, 0.147) (0)|**--------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.853, 0.936) (0)|--------**|(1)
```

- Change amount of genetic material to LARGE with a mass of 0.5 and change biopsy site and cell result to NEAR PANCREAS and REGULAR with a mass of 0.5 and ML Prediction to PC with a mass of 0.5

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.26, 0.374) (0)|--**------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.626, 0.739) (0)|------**--|(1)
```

- Change biopsy site and cell result to NEAR PANCREAS and IRREGULAR with a mass of 0.5

```
Belief Of Having Pancreatic Cancer Lies Between:      (0.534, 0.653) (0)|-----**---|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.346, 0.465) (0)|---**-----|(1)
```

- Change biopsy site and cell result mass to 0.8 and amount of genetic material to SMALL with a mass of 0.9

```
Belief Of Having Pancreatic Cancer Lies Between:      (0.675, 0.758) (0)|------**--|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.241, 0.324) (0)|--**------|(1)
```

- Change biopsy site and cell result to NOT NEAR PANCREAS and REGULAR with a mass of 0.9

```
Belief Of Having Pancreatic Cancer Lies Between:      (0.141, 0.202) (0)|-**-------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.797, 0.858) (0)|-------**-|(1)
```

This experiment demonstrates that having moderate mass for a LARGE quantity of genetic material and a biopsy location NEAR the pancreas with REGULAR cell results does not strongly support the ML prediction of PC. However, when the cell result is changed to IRREGULR with a belief of 0.5, we can see the pancreatic cancer evidential interval increase from (0.26, 0.374) to (0.534, 0.653). Furthermore, when the amount of genetic material is changed to SMALL with a high mass of 0.9, while the biopsy location remains NEAR the pancreas and the cell result remains IRREGULAR with an increased mass of 0.8, the evidential interval increased further to (0.675, 0.758), indicating that the amount of genetic material is not as significant of a factor when compared to the biopsy location and cell result. Once the biopsy location is changed back to NOT NEAR and cell result is changed to REGULAR, with a high mass of 0.9, we can see the pancreatic cancer evidential interval drop significantly.

ER Experiment 4

The final experiment tests how the NGS technology and the quality of the sequencing read impact the ML prediction. The expectation is that as the sequencing reads become more error prone by having a high guanine cytosine (GC) content and high homopolymer regions, the ML prediction becomes less reliable, resulting in a lower evidential interval for pancreatic cancer.

The experiment begins with the same baseline:

- Baseline inputs

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.063, 0.147) (0)|**--------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.853, 0.936) (0)|--------**|(1)
```

- Change ML prediction to PC with a mass of 0.7

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.373, 0.471) (0)|---**-----|(1)
Belief Of Not Having Pancreatic Cancer Lies Between:  (0.529, 0.626) (0)|-----**---|(1)
```

- Change sequencing read to HIGH GC HIGH HMR with a mass of 0.9

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.206, 0.329) (0)|--**------|(1)

Belief Of Not Having Pancreatic Cancer Lies Between:  (0.671, 0.793) (0)|------**--|(1)
```

- Change NGS tech mass to 0.9

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.206, 0.328) (0)|--**------|(1)

Belief Of Not Having Pancreatic Cancer Lies Between:  (0.671, 0.793) (0)|------**--|(1)
```

- Change NGS tech to ILLUMINA

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.205, 0.329) (0)|--**------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.67, 0.794) (0)|------**--|(1)
```

- Change ML prediction mass to 0.8

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.282, 0.394) (0)|--**------|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.606, 0.717) (0)|------**--|(1)
```

- Change sequencing read to MOD GC LOW HMR with a mass of 0.9

```
Belief Of Having Pancreatic Cancer Lies Between:     (0.732, 0.777) (0)|-------*--|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.223, 0.267) (0)|--*-------|(1)
```

After the belief in the ML prediction input of PC is increased from 0.5 to 0.7, resulting in an increased pancreatic cancer evidential interval, the change of the sequencing read frame from LOW GC LOW HMR to HIGH GC HIGH HMR causes the evidential interval to decrease from (0.373, 0.471) to (0.206, 0.329). This is because a HIGH GC and HIGH HMR increases the likelihood of errors during our sequencing, resulting in a ML prediction that is not as reliable. Little changes when the NGS tech is changed from ION TORRENT to ILLUMINA or when the ML prediction is increased from 0.7 to 0.8. Once the sequencing read is changed to MOD GC LOW HMR with a higher belief of 0.9, our ML prediction becomes significantly more reliable, resulting in an increase in the evidential interval to (0.732, 0.777).

We can verify the results of our experiments based on our knowledge of the relationships between the chosen inputs. In all four experiments the evidential reasoning model resulted in outputs that were near to what was expected.

We can hypothesize the possibility of our evidential reasoning model predicting pancreatic cancer twice as early under the assumption that 28 out of the 185 cases in the NIH GDC dataset had at least 2 of the prevalent mutations in the dataset, resulting in a positive machine learning prediction of pancreatic cancer. Using a significance level (α) of .05%, a sample size of 28 and the standard deviation (362) and mean (459) of the days to death after diagnosis from the original NIH GDC dataset, we can calculate a z-score of 4.97 and a statistically significant p-value of .00001. Because our p-value is less than our significance level, we can reject the null hypothesis under such a scenario.

CONCLUSION AND DISCUSSION

The output of the evidential reasoning model used in this experiment changed in a manner that was expected, based on the inputs. This experiment shows that use of an evidential reasoning model as a diagnostic tool is not out of reach. However, the viability of this approach depends on advances in obtaining high quality pancreatic serum samples, accessibility to powerful and accurate NGS technology, and accurate personal medical history. Such an approach could help narrow down a wide demographic, into a manageable population that could be observed closely and screened annually. As machine learning continues to expand into medical diagnostics, the integration of an evidential reasoning approach could be utilized for other applications to improve prediction results.

FUTURE WORK

The next phase of this project should have a focus on genomic mutations that are classified as having a deleterious or lethal impact. According to the National Cancer Institute, these are disease causing mutations because they increase an individual's predisposition to a disease and are most often inherited. The GDC database that was utilized to obtain the test data for this experiment includes an impact label for each mutation that is set to either significant, moderate, or deleterious. Mutations with a deleterious impact could be queried and analyzed using the GDC API. This will result in a more focused dataset but runs the risk of being sparse. The pancreatic cancer genomic data used in this project was gathered through The Cancer Genome Atlas Program (TCGA). TCGA may be the most reputable source for genomic data, however, it is limited, with only 185 cases in the pancreatic cancer project. In order to improve the machine learning model, it will be necessary to seek other sources of pancreatic cancer genomic data. Also, sampling techniques should be considered to handle imbalanced data. To continue building and improving the evidential reasoning model, tests should be conducted using real individual health and family history associated with the cases in the pancreatic cancer dataset, along with information regarding the verified type of NGS technology used to obtain the genomic data. To increase the precision of the sequencing reads input, genomic samples should be programmatically analyzed to detect low/high GC counts and quantity of homopolymer regions.

REFERENCES

Allard, J. W., Jeri, J., Miller, C. M., Repollet, M., Connelly, M. C., Rao, C., . . . Terstappen, L. W. (2004). Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 6897-904.

Amin, S., & DiMaio, C. J. (2016). Pancreatic Adenocarcinoma. In M. S. Wagh, & P. V. Draganov (Eds.), *Pancreatic Masses* (pp. 11-20). Switzerland: Springer, Cham.

Bachmann, J., Michalski, C. W., Martignoni, M. E., Büchler, M. W., & Friess, H. (2006). Pancreatic resection for pancreatic cancer . *HPB*, 346–351.

Beger, H. G., Nakao, A., & Neoptolemos, J. P. (Eds.). (2015). *Pancreatic Cancer, Cystic Neoplasms and Endocrine Tumors : Diagnosis and Management.* John Wiley & Sons, Incorporated.

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing . *Nucleic Acids Research*.

Cho, S., Kim, H., Oh, S., Kim, K., & Park, T. (2009). Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proceedings*.

Court, C. M., Ankeny, J. S., Sho, S., Hou, S., Li, Q., Hsieh, C., . . . Tomlinson, J. S. (2016). Reality of Single Circulating Tumor Cell Sequencing for Molecular Diagnostics in Pancreatic Cancer. *The Journal of molecular diagnostics : JMD, 18*(5), 688–696.

De La Cruz, M. S., Young, A. P., & Ruffin, M. T. (2014, April 15). Diagnosis and Management of Pancreatic Cancer. *American Family Physician*.

Gharibi, A., Adamian, Y., & Kelber, J. A. (2016, April). Cellular and Molecular Aspects of Pancreatic Cancer. *Acta Histochemica, 118*(3), 305-316.

Giacona, M. B., Ruben, G. C., Iczkowski, K. A., Roos, T. B., Porter, D. M., & Sorenson, G. D. (1998). Cell-Free DNA in Human Blood Plasma: Length Measurements in Patients with Pancreatic Cancer and Healthy Controls. *Pancreas*, 89-97.

Gupta, S., Wang, F., Holly, E. A., & Bracci, P. M. (2010). Risk of pancreatic cancer by alcohol dose, duration, and pattern of consumption, including binge drinking: a population-based study. *Cancer Causes Control*, 1047-1059.

Heydari, M., Miclotte, G., Van de Peer, Y., & Fostier, J. (2019, June 3). Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. *BMC Bioinformatics, 20*(1).

Hijioka, S., Yamao, K., Mizuno, N., Imaoka, H., Bhatia, V., & Hara, K. (2017). Early Diagnosis of Pancreatic Cancer Using Endoscopic Ultrasound. In H. Yamaue (Ed.), *Innovation of Diagnosis and Treatment for Pancreatic Cancer* (pp. 3-11). Singapore: Springer.

Howlader, N., Krapcho, M., Garshell, J., Neyman, N., Altekruse, S., Kosary, C., . . . Cronin, K. (2016). *SEER Cancer Stat. Rev.* Bethesda: National Cancer Institute.

Isaji, S., Mizuno, S., Windsor, J. A., Bassi, C., Fernández-del Castillo, C., Hackert, T., . . . Wolfgang, C. L. (2018). International consensus on definition and criteria of borderline resectable pancreatic ductal adenocarcinoma 2017. *Pancreatology*, 2-11.

Lowrance, J. D., Strat, T., Wesley, L. P., Garvey, T. D., Ruspini, E., & Wilkins, D. (1991). *The Theory, Implementation, and Practice of Evidential Reasoning.*

Man, Y., Wang, Q., & Kemmner, W. (2011). Currently Used Markers for CTC Isolation - Advantages, Limitations and Impact on Cancer Prognosis. *Journal of Clinical & Experimental Pathology*.

Molina, J. R., & Adjei, A. A. (2006). The Ras/Raf/MAPK Pathway. *Journal of Thoracic Oncology*, 7-9.

O'Brien, D. P., Sandanayake, N. S., Jenkinson, C., Gentry-Maharaj, A., Apostolidou, S., Fourkala, E.-O., . . . Timms, J. F. (2015, February). Serum CA19-9 is significantly upregulated up to 2 years before diagnosis with pancreatic cancer: Implications for early disease detection. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 21*(3), 622-631.

Pandol, S. J., Apte, M. V., WIlson, J. S., Gukovskaya, A. S., & Edderkaoui, M. (2012). The Burning Question: Why is Smoking a Risk Factor for Pancreatic Cancer? *Pancreatology*, 344-349.

Qi, Z.-H., Xu, H.-X., Zhang, S.-R., Xu, J.-Z., Li, S., Gao, H.-L., . . . Liu, L. (2018). The Significance of Liquid Biopsy in Pancreatic Cancer. *Journal of Cancer*, 3417–3426.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012, July 24). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*, 341.

Risch, H., Yu, H., Lu, L., & Kidd, M. (2015, July 1). Detectable Symptomatology Preceding the Diagnosis of Pancreatic Cancer and Absolute Risk of Pancreatic Cancer Diagnosis. *American journal of epidemiology, 182*(1), 26–34.

Siegel, R. L., Miller, K. D., & Jemal, A. (2015). Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians*.

Site-specific cancer series : Pancreatic and hepatobiliary cancer. (2012). In C. M. Handy, & D. O'Dea (Eds.). Oncology Nursing Society.

Stark, A., & Eibl, G. (2015, May 23). Pancreatic Ductal Adenocarcinoma. *Pancreapedia: Exocrine Pancreas Knowledge Base*.

Way, G. P., Sanchez-Vega, F., La, K., Armenia, J., Chatila, W., Luna, A., . . . Greene, C. S. (2018). Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell reports*, 172-180.

Wesley, L. e. (2019, January). MRI: Acquisition Of A Hybrid Computer/GPU Node And PB-Storage For STEM R&D And Education, Proposal to the National Science Foundation-Major Research Initiative program.

Wright, E. S., & Vetsigian, K. H. (2016, November 4). Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics, 17*(1).

Yager, R., Liu, L., Dempster, A. P., & Shafter, G. (2008). *Class Works of the Dempster-Shafter Theory of Belief Functions.* New York: Springer-Verlag Berlin Heidelberg.

Yeo, Z. X., Chan, M., Yap, Y. S., Ang, P., Rozen, S., & Lee, A. S. (2012, September 19). Improving Indel Detection Specificity of the Ion Torrent PGM Benchtop Sequencer. *PLoS ONE, 7*(9).

Zill, O. A., Greene, C., Sebisanovic, D., Siew, L. M., Leng, J., Vu, M., . . . Tala. (2015). Cell-Free DNA Next-Generation Sequencing in Pancreatobiliary Carcinomas. *Cancer discovery*, 1040-8.

APPENDIX A

Among other symptoms, intolerance to glucose is another symptom some pancreatic cancer patients encounter (Site-specific cancer series : Pancreatic and hepatobiliary cancer., 2012). Additional symptoms include back pain, anorexia, nausea, epigastric bloating, heartburn, pruritus, and dysgeusia (Risch, Yu, Lu, & Kidd, 2015). Individuals who aggressively develop atypical type 2 diabetes mellitus, who are also thin and over the age of 50, are suspect to have pancreatic cancer (De La Cruz, Young, & Ruffin, 2014).

APPENDIX B

A first step for patients presenting the common symptoms associated with pancreatic cancer is ultrasonography imaging of the abdomen. Pancreatic cancer is most often diagnosed after detection through some visualization method such as computed tomography (CT) and/or magnetic resonance imaging (MRI) coupled with magnetic resonance cholangiopancreatography (MRCP), or endoscopic ultrasound (EUS). Multi-detector (MD) row CT is widely utilized in evaluating pancreatic cancer, however, it has a low rate of detection. Conversely, EUS has the ability to detect small pancreatic masses with high sensitivity. If a pancreatic mass is discovered, the patient usually undergoes EUS and fine-needle aspiration (FNA) biopsy of the mass (De La Cruz, Young, & Ruffin, 2014). EUS is viewed as one of the highest accuracy methods for detecting focal lesions and tumors that have a size of <=2 cm. It is believed that EUS can detect tumors that are less than 10mm. One report found EUS to have a sensitivity of 84% when detecting 25 small pancreatic tumors less than 10mm each (Hijioka, et al., 2017). Another study using EUS-FNA to detect pancreatic cancer masses with a size less than 10mm in 23 patients resulted in an accuracy of 96% (Hijioka, et al., 2017). One issue with EUS is its tendency to overlook a pancreatic mass in individuals suffering from other pancreas related issues such as chronic pancreatitis (Hijioka, et al., 2017). While visualization has proven as an effective aide in the detection of pancreatic cancer, in order to increase the survival rate, detection needs to occur before pancreatic masses can be visualized.

APPENDIX C

Biomarkers are proteins, antigens, and other cellular molecules that are expressed in higher levels in the presence of a disease. Biomarkers can be obtained in a variety of ways, the most common being biopsy of tissue or liquid serum. Traditional biopsies of tumor tissue are common but have several limitations and most often take place too late. Accurate liquid biopsies are a subject of interest because they are easily repeatable and result in real time detection. The content of liquid biopsies, such as circulating tumor cells (CTCs), cell-free nucleic acid, and exosomes, all have potential to contribute to the detection of pancreatic cancer (Qi, et al., 2018).

APPENDIX D

With proteomics, one approach is to compare the healthy tissue and tumor tissue of individuals diagnosed with pancreatic cancer, in order to find differences in global protein levels. A higher or lower level of a protein could potentially indicate whether the protein plays a part in tumor growth or suppression. Another approach is to compare the protein levels in primary and metastatic tumors, in hopes of finding proteins that play a role in metastasis (Gharibi, Adamian, & Kelber, 2016). Secretomics focuses on secreted proteins that are involved in cell signaling and the dissemination of tumor cells. Secretome samples can be obtained via cancerous pancreatic cells. Secretome proteins can be identified using a mass spectrometry approach known as stable isotope labeling with amino acids in cell culture (SILAC) (Gharibi, Adamian, & Kelber, 2016). Another option is to examine the exosomes derived from tumor tissue. Exosomes are vesicles that store the nucleic acid and proteins that are secreted by cells. Inspecting the contents of cancer exosomes could reveal information about the proteins that play a role in metastasis.

APPENDIX E

Circulating tumors cells (CTCs) are promising traceable components of cancer. CTCs are cells that originate from a primary tumor but break away and circulate in the bloodstream. These cells are a subject of interest because they can enter the bloodstream during early stages, as tumors form. It is believed that the number of CTCs in a blood sample is correlated to the stage of tumor development (Qi, et al., 2018). One approach to detecting CTCs is to use antibodies against antigens that are located on the surface of the CTCs. CTCs can be detected by using epithelial cell adhesion molecule (EpCAM) as a marker (Gharibi, Adamian, & Kelber, 2016). Because carcinoma and epithelial cells commonly express EpCAM, while healthy blood cells do not, the presence of EpCAM can separate CTCs from otherwise healthy blood cells (Man, Wang, & Kemmner, 2011). CellSearch (Veridex), an FDA approved method, utilizes ferrofluids that are developed with EpCAM to catch CTCs. A study using this method found that CTCs related to prostate and breast cancer were found in higher quantity compared to CTCs related to pancreatic cancer. The CTCs that were found to be associated with pancreatic cancer had levels that were similar to CTCs of nonmalignant diseases (Allard, et al., 2004). As a result, the low sensitivity and EpCAM dependence makes CTC detection for pancreatic cancer difficult. Researchers have been able to fluorescently tag and track pancreatic cancer related CTCs in mouse models and observe the progression of cancer (Gharibi, Adamian, & Kelber, 2016). Researchers have also found ways to deal with the low levels of CTCs. Molecular analysis and immunocytochemistry (ICC) are methods used to determine whether a CTC originates from a tumor (Court, et al., 2016). However, a small number of CTCs means limited DNA is available. ICC is coupled with whole genome amplification (WGA) to make up for the lack of DNA microgram levels required, when only pictogram levels are available (Court, et al., 2016). However, one of the limitations of WGA is allele dropout (ADO) as a result of amplification bias causing certain alleles to not be adequately amplified. Because WGA does not provide adequate coverage of certain genes, a mutant allele could be completely missed. These are the types of potential errors that must be taken into account when deciding on sequencing methods (Court, et al., 2016).

APPENDIX F

Circulating tumor DNA (ctDNA) is a unique DNA type that enters the bloodstream through apoptosis and necrosis of tumor cells. ctDNA is also found in high concentration in blood samples and is packed with useful information that can be helpful in detecting tumors. A study found a high concentration of cfNA in the blood of pancreatic cancer patients that was different from those of a healthy control (Qi, et al., 2018). Exosomes are another type of potential biomarker that can be found in a blood sample. Exosomes are extracellular vesicles that are secreted into circulation by healthy and cancerous cells. Exosomes are involved in the communication of cancer cells and their environment and can be found in blood samples earlier than cfNA's (Qi, et al., 2018). Exosomes can be found in significantly higher levels in pancreatic cancer patients than healthy individuals. High levels of exosomes are due to the extensive exosome secretion during carcinogenesis. Because exosomes are the byproduct of the secretion of living cells, they can be found in blood when tumors are at an early stage (Qi, et al., 2018). It is believed that serum-exosome protein could be a viable marker for diagnosing pancreatic cancer. In one experiment, a panel of cancer initiating proteins and miRNA that were found to be highly expressed in the exosomes of pancreatic cancer were selected (Qi, et al., 2018). The panel of proteins and miRNA were found to be highly regulated in the exosomes derived from pancreatic patients, but not found in healthy patients, hinting that a combined panel of proteins and miRNA could be used as a diagnostic tool with high sensitivity. Because exosomes contain proteins and RNA, a study was able to confirm that pancreatic cancer exosomes contained genomic DNA. The study found that exosomes hold more than 10kb portions of double stranded DNA (Qi, et al., 2018). Using genomic DNA from pancreatic cancer exosomes, KRAS and P53 gene mutations were able to be identified, indicating that genomic data from exosomes obtained by serum could be used for diagnosing pancreatic cancer. Another study found that the GPC1 protein, located in the exosome membranes, was found at a higher concentration in 190 pancreatic cancer patients compared to healthy donors (Qi, et al., 2018). These GPC1 proteins could be discovered before masses could be detected through an MRI. Receiver operating curves indicate that GPC1 proteins are an ideal biomarker with 100% specificity and sensitivity (Qi, et al., 2018).

APPENDIX G

Sequencing is a methodology used to observe the order of nucleotide bases of the genome or exome, enabling researchers to compare genomic and transcriptomic variations between healthy and cancerous tissue. Knowledge of the nucleotide base order makes classification of gene mutations possible. By identifying mutations, therapies can be developed to target specific mutations while limiting the impact on other cells. When testing for biomarkers, the type of NGS technology utilized along with the type of proteomic material analyzed plays a large role. For analysis of ctDNA/ctRNA, high coverage sequencing is necessary. In order to reduce error rates of sequencing associated with redundancy, molecular barcodes are utilized. NGS approaches for detecting CTCs tend to be less sensitive than digital polymerase chain reaction (PCR) method and other approaches. However, NGS has the benefit of checking a high volume of chromosome loci. When considering whole exome sequencing (WES) instead of whole genome sequencing (WGS), one major disadvantage is the inability to recognize noncoding variants and rearrangements that have the potential to have a major impact on gene regulation. This makes a notable difference when dealing with mutations that have the potential to be biomarkers and are part of the regulatory and/or promoter regions. Also, WGS related methods may obstruct detection of copy number alteration (CNA). Lastly, the quality of each NGS method, for instance the depth of coverage, read length, and other parameters affects the accuracy, precision, and thoroughness of sequence data (Wesley, 2019). While NGS technology continues to advance, errors are still common and vary with the type of NGS technology used. Illumina's sequencing technology is highly utilized for its ability to generate sequences with high accuracy and throughput. It is believed that 90% of sequencing is done using Illumina. Although Illumina is widely adopted, it is not without flaws. Illumina sequencing is known to have an error rate of 1-2%. Biases are also common with this kind of sequencing. These biases include a higher concentration of errors occurring near the end of DNA reads. Substitution errors tend to be more common than insertion/deletion errors. Causes of error in Illumina include crosstalk, phasing, fading, and T accumulation (Heydari, Miclotte, Van de Peer, & Fostier, 2019). Substitution errors occur when a base is incorrectly identified and tend to take place near the end of a sequence. Homopolymer errors occur in regions where the same nucleotide repeats consecutively. This is also known to occur in regions where true polymorphism occurs. Reducing these errors comes at a cost of reducing the sensitivity of the technology (Yeo, et al., 2012). In one study that compared the different sequencing technologies, researchers found that Illumina would result in errors when analyzing long homopolymers that were > 20 bases. Ion Torrent, another sequencing technology, had trouble reading homopolymer regions that were > 14 bases, nor could it accurately predict the bases in homopolymers > 8 bases (Quail, et al., 2012). Homopolymer errors occur in regions where the same nucleotide repeats consecutively. This is also known to occur in regions where true polymorphism occurs. For SNP calling, a type of analysis of NGS data, researchers found PacBio technology to be the most problematic and resulted in the least accuracy compared to the other sequencing technologies. The accuracy of PacBio for SNP detection was found to be 70% with nearly double the false positive rate of the other sequencing technologies (Quail, et al., 2012).

APPENDIX H

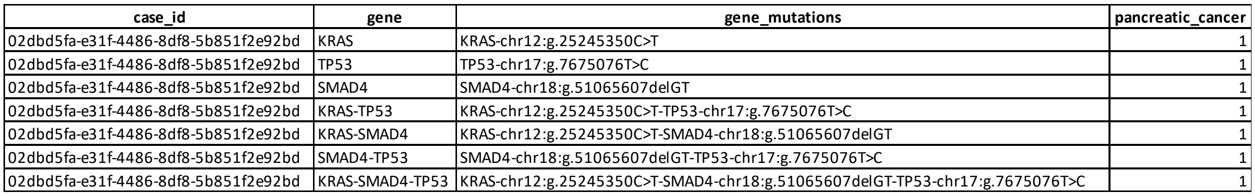
Some of the most common pan-cancer mutated genes are those associated with the Ras pathway. Mutations in the Ras pathway drive tumorigenesis and are tied to drug resistance and low survival odds. Developing therapies for the mis-regulation of the Ras pathway is a major goal of National Cancer Institute (Way, et al., 2018). The Ras pathway is a route used to signal and activate genes for cell growth, division, migration, regulation and more (Molina & Adjei, 2006). In precision oncology, an individual's genomic data is used to find therapies that are best suited to their genomic make up. However, not all patients are able to be paired with a therapy. Way et al. identifies these patients as "hidden responders" and believes their transcriptomes could shed light on therapies they could be responsive to. To improve the matching of these "hidden responders" with suitable therapies, Way et al. believes classification of aberrant pathways, especially in Ras, could be a solution. To classify, Way et al. developed a machine learning classifier that utilizes RNA-seq, copy number variation, and the mutations from over 30 different types of cancer in order to detect abnormal downstream gene expressions associated with aberrant Ras pathway behavior (Way, et al., 2018). With these features, Way et al. is able to not only recognize the activation of Ras, but also identify phenocopying variants and predict the response to MEK inhibitors, which are used to target the Ras pathway. Their classifier of choice was an elastic net penalized logistic regression model that learned changes in pathways from gene expression acquired from biopsies of tumors from various cancer types. Way et al. states that by using the elastic net regularization penalty they are causing sparsity, resulting in mostly a selection of genes associated with activation of the RAS pathway (Way, et al., 2018). Incorporating regularization in analysis of genomic data is not uncommon. In a study to classify SNPs associated with rheumatoid arthritis, Cho et al. applied elastic net regularization to address highly correlated features (Cho, Kim, Oh, Kim, & Park, 2009) in their model. Cho et al. states that elastic net regularization helped their experiment by providing both automatic feature selection and applying continuously shrinking coefficients and is especially helpful when the number of correlated features outnumber the sample size.

APPENDIX I

For each random variable in our evidential reasoning model, a frame is constructed which declares the frame name, the data type (discrete or continuous), the possible propositions, the parent frames that the current frame originates from, the resulting frame that the current frame plugs into, and the compatibility relations of the frame. Frames are defined in a text file referred to as the gallery. A mass distribution file is created containing the masses assigned to the propositions in each frame. For instance, if we are 90% certain an individual had a history as a highly active smoker, a mass of 0.9 would be assigned to the HIGH smoking proposition. Another file assigns a discount rate to each frame with the purpose of either reducing the impact of a frame or to imply a notion of inaccuracy or lack of credibility. These text files are used as input into the program Capri that handles the relationships between frames and the fusing, using Dempster's Rule. The output is an evidential interval that indicates the minimum belief and maximum justifiable belief in the propositions listed in the final output frame.

APPENDIX J

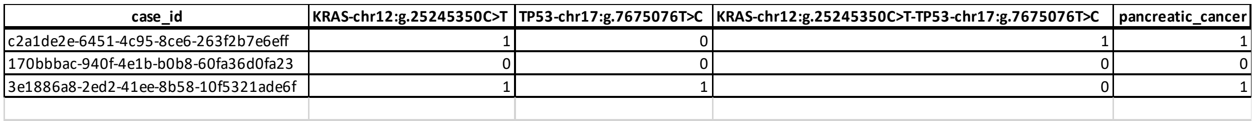
To obtain data to train our machine learning model, a custom program was created to query the desired percentage of the most prevalent genes and associated mutations from the TCGA-PAAD pancreatic project in the GDC database. This program retrieved all the cases from the GDC pancreatic project, along with cases from other cancer projects that shared mutations found in the pancreatic project. The queried GDC data was transformed by applying a permutation function to obtain all the mutations associated with each individual case, resulting in an entry for each possible gene mutation combination per case. If an individual in the dataset had 5 mutations, the transformation would result in 2^5 different entries of mutation combinations, as shown in Figure 5.



case_id	gene	gene_mutations	pancreatic_cancer
02dbd5fa-e31f-4486-8df8-5b851f2e92bd	KRAS	KRAS-chr12:g.25245350C>T	1
02dbd5fa-e31f-4486-8df8-5b851f2e92bd	TP53	TP53-chr17:g.7675076T>C	1
02dbd5fa-e31f-4486-8df8-5b851f2e92bd	SMAD4	SMAD4-chr18:g.51065607delGT	1
02dbd5fa-e31f-4486-8df8-5b851f2e92bd	KRAS-TP53	KRAS-chr12:g.25245350C>T-TP53-chr17:g.7675076T>C	1
02dbd5fa-e31f-4486-8df8-5b851f2e92bd	KRAS-SMAD4	KRAS-chr12:g.25245350C>T-SMAD4-chr18:g.51065607delGT	1
02dbd5fa-e31f-4486-8df8-5b851f2e92bd	SMAD4-TP53	SMAD4-chr18:g.51065607delGT-TP53-chr17:g.7675076T>C	1
02dbd5fa-e31f-4486-8df8-5b851f2e92bd	KRAS-SMAD4-TP53	KRAS-chr12:g.25245350C>T-SMAD4-chr18:g.51065607delGT-TP53-chr17:g.7675076T>C	1

Figure 5. Example of powerset of mutations in dataset

Depending on the percentages queried, the dataset often resulted in an imbalance of classification labels. Before training the SVC, the dataset was balanced by removing entries associated with a label that had a significantly higher count. A final dataset was created consisting of columns for each unique gene mutation permutation and the pancreatic cancer label. Each row in this dataset consisted of a unique case with a 1 for each gene mutation column associated with the case, as shown in Figure 6.



case_id	KRAS-chr12:g.25245350C>T	TP53-chr17:g.7675076T>C	KRAS-chr12:g.25245350C>T-TP53-chr17:g.7675076T>C	pancreatic_cancer
c2a1de2e-6451-4c95-8ce6-263f2b7e6eff	1	0	1	1
170bbbac-940f-4e1b-b0b8-60fa36d0fa23	0	0	0	0
3e1886a8-2ed2-41ee-8b58-10f5321ade6f	1	1	0	1

Figure 6. Example of final dataset used to train SVC

Using the dataset shown above, the correlation of each column with the pancreatic cancer label was generated. Gene mutation columns that fell below a certain correlation were dropped from the data set. To eliminate the possibility of feature correlation, an L1 penalty or Lasso regularization was applied.