Spring 5-22-2020

Evidence-Based Detection of Pancreatic Canc

Rajeshwari Deepak Chandratre

Evidence-Based Detection of Pancreatic Cancer

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

Rajeshwari Deepak Chandratre

May 2020

ABSTRACT

This study is an effort to develop a tool for early detection of pancreatic cancer using evidential reasoning. An evidential reasoning model predicts the likelihood of an individual developing pancreatic cancer by processing the outputs of a Support Vector Classifier, and other input factors such as smoking history, drinking history, sequencing reads, biopsy location, family and personal health history. Certain features of the genomic data along with the mutated gene sequence of pancreatic cancer patients was obtained from the National Cancer Institute (NIH) Genomic Data Commons (GDC). This data was used to train the SVC. A prediction accuracy of ~85% with a ROC AUC of 83.4% was achieved. Synthetic data was assembled in different combinations to evaluate the working of evidential reasoning model. Using this, variations in the belief interval of developing pancreatic cancer are observed. When the model is provided with an input of high smoking history and family history of cancer, an increase in the evidential reasoning interval in belief of pancreatic cancer and support in the machine learning model prediction is observed. Likewise, decrease in the quantity of genetic material and an irregularity in the cellular structure near the pancreas increases support in the machine learning classifier's prediction of having pancreatic cancer. This evidence-based approach is an attempt to diagnose the pancreatic cancer at a premalignant stage. Future work includes using the real sequencing reads as well as accurate habits and real medical and family history of individuals to increase the efficiency of the evidential reasoning model. Next steps also involve trying out different machine learning models to observe their performance on the dataset considered in this study.

Key Words: Pancreatic Cancer, Evidential Reasoning, Next Generation Sequencing, Genetic Mutations, Machine Learning, Support Vector Classifier, Biomarkers.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Pancreatic cancer is belligerent in its own way, since it hardly displays any observable symptoms before metastasis. Due to this, it is difficult to treat it before it spreads beyond control. The lone remedial option is the surgical resection of the tumor (Amin & DiMaio, 2016). The fact that it is fourth in position to cause cancer associated deaths reflects the lethality of this disease (Amin & DiMaio, 2016). Pancreatic Ductal Adenocarcinoma (PDAC) is one of the most aggressive types of pancreatic cancer in which cancerous tumors develop in the ductal cells of the pancreas. Tumors are accountable to certain genetic mutations occurring in the cells nearing pancreas. Several other factors such as smoking and drinking habits, a history of cancer in the family, and other hereditary factors may lead to aberrant mutation of these cells (See Appendix A for more information on these factors). PDAC is not only the most common pancreatic cancer, but also the deadliest one. The rate of a patient surviving beyond 5 years after being diagnosed with PDAC is as low as 7% (Gharibi, et al., 2017). This is because the PDAC is very difficult to detect at an early stage.

The statistics of the pancreatic cancer cases resulting in death is brutal with an estimate of 83% cases leading to death. (Siegel, Miller, & Jemal, 2015). Gender wise, pancreatic cancer is observed to be a little more prevalent in males than in females: 13.9 out of 100,000 for males and 10.9 per 100,000 for females. Ethnicity wise, African Americans are at a higher risk of developing pancreatic cancer with probability of almost 15.8 out of 100,000 than that of Asian Americans which is 9.8 out of 100,000 (Howlader, et al., 2016). More age may mean greater likelihood of developing pancreatic cancer. About 27% of new diagnoses belong to the age group

of 75 and 84 and 9% are in the age group of 45 and 54 years old (Howlader, et al., 2016).

PDAC is usually diagnosed at an advanced stage when the tumor has spread beyond the pancreatic region of the patient. At this point, it is hard to remove the tumor through surgery (Amin & DiMaio, 2016). As per the study by Gharibi, Adamian, and Kelber, due to its aggressive nature, PDAC metastasizes rapidly and the treatment of PDAC becomes extremely challenging (Gharibi, Adamian, & Kelber, 2016). The 5-year survival rate decreases rapidly when PDAC is detected at an advanced stage. Therefore, there is a need to detect the pancreatic cancer at a primitive stage when it is localized to pancreas in order to treat it successfully (Amin & DiMaio, 2016).

The limitation with existing methods to detect PDAC are that they fail in detecting the disease at a premalignant stage. Since pancreatic cancer causes genetic alterations, monitoring the presence of biomarkers in a tumor specimen remains a popular technique. But the attempts made to detect the presence of pre-malignant tumors by observing the genetic mutations in the patient's tissue focus on limited genetic material. Research has found that there are about 63 different genes that get altered by pancreatic cancer (Gharibi, et al., 2017). Each of these genes may undergo multiple kinds of mutations, but an effective technique to predict the likelihood of pancreatic cancer by considering all genetic alterations as well as other factors such as smoking and drinking history, medical history, cellular structure, and others does not exist. Hence, there is a need to develop a technique that will predict the likelihood of a person developing pancreatic cancer at an early stage using various mentioned factors which may impact the development of pancreatic tumor to ensure timely treatment.

BACKGROUND

Limitations of the existing methods in detecting pancreatic cancer at an early stage exist because of lack of symptoms. Pain in abdomen, jaundice, loss of weight are common early symptoms of pancreatic cancer. Some of the other symptoms are back pain, anorexia, heartburn, and dysgeusia (Risch, Yu, Lu, & Kidd, 2005). It is often the case that symptoms of pancreatic cancer can be confused at later stages with other ailments. Medical imaging is used to detect the presence of pancreatic cancer which results in discovery of masses. The noticeable masses are generally detected at a later stage. At this point, the survival rate of pancreatic cancer drastically decreases. Thus, lack of imaging technology to detect pancreatic cancer at a premalignant stage calls for the need to find newer methods which may aid in earlier detection of this disease. (See Appendix B for imaging technology details)

Use of Biomarkers in Detection

Presence of biomarkers is one of the important techniques in early detection of pancreatic cancer. Biomarkers are characterized by high presence of cellular molecules such as proteins, antigens, etc. in pancreatic cancer patients. They can be found using tissue biopsy of liquid biopsy (Qi, et al., 2018). Samples collected for studying biomarkers before the diagnosis are very less, but they are the ones which are preferred over the samples of patients already diagnosed with pancreatic cancer (O'Brien, et al., 2015).

Carbohydrate antigen 19-9 (CA19-9) is one of the most common biomarkers used presently, with a sensitivity between 69% - 98% and a specificity between 46% - 98%. O’Brien et al. observed that there was a detectable surge in CA19-9 levels 3

years prior to PDAC diagnosis (O'Brien, et al., 2015). However, CA19-9 is not specifically related to pancreatic cancer as it is also detected in gastrointestinal tumors and hence cannot be sufficient enough to be considered alone as a biomarker in identification of pancreatic cancer.

Existing Research in Genomic Biomarkers

There has been a significant body of research in detection of pancreatic cancer-causing cell mutations. Various literature sources study the presence of specific biomarkers which may lead to pancreatic cancer and there is a considerable development in this research suggesting the presence of certain genetic material or biomarkers in the tissue specimen of patients is indicative of presence of malignant tumors in their pancreas. Experiments help to identify characteristics of these biomarkers and how they can potentially cause PDAC. Traditional biomarker detection techniques use tissue biopsy method to detect the presence of malignant tumors. This involves sectioning the patient's tissue using surgical or needle biopsy (Qi, et al., 2018). In contrast to the invasive traditional biopsy techniques, Z. Qi et al. suggest that inference based on presence of a single biomarker, or study of a single biopsy sample may provide only limited information as the tumors and their genetic composition is found to be heterogenic. They argue that due to the limited true positive and true negative rates provided by biomarkers such as carcinoembryonic antigen (CEA) and carbohydrate antigen, there is a need for a better technique that can monitor new biomarkers. This study further suggests that liquid biopsy is a promising technique because of its non-invasiveness and effectiveness in detecting circulating tumor cells (CTCs) and cell-free circulating nucleic acids (cfNAs) which keep circulating in the body fluids such as blood.

However, one of the major limitations associated with liquid biopsy technique is that it relies more on body fluids rather than cancerous tumors to detect the biomarkers, which may produce misleading conclusions (Qi, et al., 2018).

A. Gharibi et al. study the presence of other biomarkers which may indicate the presence of ductal tumors. They analyzed and found that integrin alpha 1 (ITGA1), acts as an ideal biomarker for diagnosis and therapeutic technique for PDAC. Presence of ITGA 1 is detected in high quantity in 42% of the tumor tissue of PDAC patients, whereas it does not occur in the normal pancreatic ductal epithelial cells (Gharibi, et al., 2017). Thus, it can be used effectively in the PDAC detection.
In another study, A. Gharibi, Y. Adamian, and J.A. Kelber examine various genetic mutations and suggest that Kirsten rat sarcoma (KRas) is the most frequently observed gene that is mutated in PDAC related cases (Gharibi, Adamian, & Kelber, 2016). It is followed by other genes including p53, p16, SMAD4, PUC1, and SRC. These genetic materials after undergoing mutations, lead to inappropriate proliferation of cells in pancreas. Amongst these, p53, p16, and SMAD4 are tumor suppressing transcription factors, which normally arrest the progression of cell cycle; however, their mutated presence in abnormal quantities or their absence, leads to cell proliferation (Gharibi, Adamian, & Kelber, 2016). Hence, these are some of the important biomarkers in PDAC treatment.

Similarly, J. A. Kelber et al. found that Pseudopodium-Enriched Atypical Kinase One (PEAK1) helps in regulation of cell migration and proliferation. It is a biomarker with therapeutic target in PDAC, which regulates the necessary cell migration characteristics such as shape change of cells (Kelber, et al., 2012). Thus, this research paper concludes that PEAK1 plays a major role in limiting cancer cell migration as well as growth.

However, one major limitation of these studies is that they consider only a few genetic material or biomarkers which might prove unreliable given that the tumors may contain a variety of genetic mutations, but the biomarkers under consideration are limited. Moreover, it is observed that barring some genes, not all the biomarkers can aid in detecting of pancreatic cancer early.

Circulating Tumor Cells

Circulating Tumor Cells (CTCs) are the tumor shed cells which circulate in the blood stream of the cancer patients. A study found out that out of 12 pancreatic cancer subjects under consideration, 11 had KRAS gene mutations in their CTCs (Court, et al., 2016). Court et al. concluded that at least 10 CTCs should be present in order to detect KRAS mutations. Moreover, 7.5 mL of blood sample contains only about 1-50 CTCs along with more than a million white blood cells. Hence to detect CTCs in blood sample, a test with a high sensitivity and specificity is needed because of lesser number of CTCs in bloodstream (Court, et al., 2016). CTCs are believed to be a good prognostic biomarker, but out of all the CTCs detaching from tumors, only 0.01% develop into metastases. Newer techniques are being formulated to detect CTCs as they are great potential biomarkers in cancer development.

Sequencing

Classification of gene mutations is achieved using various sequencing technologies. Study by Lawrence, et al., mentions methods such as Mutation Significance of Covariance (MutSigCV) which are widely used to classify gene mutations from the tumor tissue of an affected individual. Lawrence et al. used MutSigCV to discover

abnormal variation in mutation frequency and spectrum observed in tumors for different cancer types by applying the technique on exome sequences from 3,083 tumor–normal pairs. Cancer related genes or mutations truly found in cancerous tumors are found out by MutSigCV by introducing the analysis with mutational heterogeneity. (Lawrence, et al., 2013). Using this sequencing approach, genes such as KRas, TP53, CDKN2A, Smad4, BCLAF1, IRF6, FLG, AXIN1, GLI3 and PIK3CA were found to be mutated expressively. Pancreatic tumor cells were separated by microdissection approach from the microenvironment in 109 affected patients using surgical resection. Various other sequencing techniques also aided in discovering novel mutations in these tumorous cells which underwent whole-exome sequencing technique (Witkiewicz, et al., 2015) (See Appendix C for more details on sequencing details).

Machine Learning

Machine learning has proved to be a promising technique to detect pancreatic cancer-causing genetic mutations. In an interesting study, G. P. Way et al. developed a machine learning model to detect abnormal Ras activation in the cancer tumors using the knowledge of Kirsten Rat Sarcoma (KRAS), Neuroblastoma Rat Sarcoma (NRAS) and KRAS, also known as transforming protein p21. The mutations in Ras pathway genes are known to drive PDAC. A classification model was developed for distinguishing the aberrant Ras pathway activity in tumors using features such as RNA-seq, copy number, and mutation data. A logistic regression classifier is used to train a model by combining these features from 33 types of cancers from The Cancer Genome Atlas (TCGA) PanCanAtlas database. The model learned a combination of weights and gene related important scores that separate aberrant patterns (Way G. P., et al., 2018). The results of this classification model by the

authors were observed in terms of area above 84% covered by the receiver operating characteristic (AUROC) curve and above 63% covered by the precision recall (AUPR) curve (Way G. P., et al., 2018).

In other research, J. Jeon et al., considered mutated genetic data as the dataset from the Catalogue of Somatic Mutations in Cancer (COSMIC) database. In general terms, they computed the probability of 15,663 proteins of being an appropriate drug target in case of treatment of cancers such as pancreatic cancer, breast cancer, and ovarian cancer. Using features such as mRNA expression intensity score, gene essentiality score, DNA copy number and mutation occurrence, 3 classifiers were developed with the help of support vector machine (SVM) and RBF kernel. SVM perform exceptionally well in inferring gene-disease correlation and in identifying the drugs with the target disease (Jeon, et al., 2014). It was further found that of the three diseases they considered the SVM could predict 43 of the 69 drug targets considered. Out of all the three classifiers, the combined accuracy obtained was 91.69%, meanwhile the specificity was 91.91%. A major drawback of this study is that these features provide little throughput screening and relatively low coverage of human genomic data which confines their use in generalized identification of genome-wide drug targets (Jeon, et al., 2014)

Thus, using machine learning in pancreatic cancer detection is a novel technique that can aid in early diagnosis of the disease. It can also create a breakthrough in the early treatment of this disease which is the goal. But it still has a scope for improvement as currently, only limited genomic data has been considered presently.

Existing Research in Evidence Based Approach

The work done by Sharghi which has been extended in this study is based on a similar evidential reasoning approach (Sharghi, 2019). This dataset gathered for this

study consisted of mutations from the Cancer Genome Atlas – Pancreatic Adenocarcinoma (TCGA-PAAD) project. Sharghi used these mutations to find the cases from other cancer projects that shared the mutations from TCGA-PAAD project (Sharghi, 2019). A permutation factor was made use of, to find the mutation entries for every individual case. This dataset had an imbalance of classification labels as only a certain percent of data was queried for each case. The Support Vector Classifier was trained on the dataset of unique gene mutation permutation and its pancreatic cancer label. An average precision score of 92% was obtained along with a ROC AUC of 92%. As per Sharghi, even though the results of the classifier were promising, its performance was still questionable owing to the limited size of 185 cases. Sharghi expressed a need for utilizing a larger dataset to find more common gene mutation combinations (Sharghi, 2019). The present study tries to overcome this limitation by considering a larger volume of dataset with mutations observed in every project in the GDC portal along with additional features such as impact of the variant on protein and the project in which the mutation occurred. Further, various experiments were carried out considering different scenarios regarding the smoking and drinking history of individual, family and personal history along with the machine learning prediction, biopsy site, and even the sequencing reads and NGS technologies. These scenarios were input to an evidential reasoning model which produced promising results.

Addressing the Technological Gaps

The present diagnostic methods fail to detect pancreatic cancer in the early stage resulting in lowering the rate of survival of this disease. According to Gharibi et al., pancreatic cancer starts metastasis 10 years prior to showing symptoms in the body of the patient (Gharibi, et al., 2017). Thus, if the mutations in the genes leading to

pancreatic cancer are detected early, the cancer can be removed effectively before it spreads beyond the pancreas, after which it becomes very difficult to treat the disease.

Further, diagnosing a lethal disease like pancreatic cancer using a few biomarkers or analyzing limited genomic data using machine learning may prove unreliable and inaccurate. The reason behind this being that there are many factors responsible for causing the aberrant mutation of cells in the pancreatic tissues. These factors may be environmental (tobacco intake, alcohol consumption, etc.) or even hereditary (genetic mutations occurring due to inherited conditions) (See Appendix A for more details regarding the causes of pancreatic cancer). All these factors including the statistical data could be vague, inaccurate, diverse, imprecise, or even insufficient to certain extent. Moreover, if age and ethnicity wise data of people at higher risk is considered, there is a chance that only a subset of a wider population being diagnosed eventually (Sharghi, 2019).

The Belief Function (BF) and evidential reason calculi is a part of a sophisticated mathematical approach used to formulate reasonable and precise results regarding likelihood of developing pancreatic cancer. It does this in a more flexible and less restrictive manner than the traditional statistical and probabilistic approaches (Lowrance, et al., 1991).

A useful knowledge source can be developed using imperfect and diverse information such as the genomic structure, alcohol consumption habits, smoking habits, health history, biopsy location, etc. This tactic can aid in early cancer diagnosis if supplied with real time data, by using belief functions and evidence-based reasoning approach. The work described in this report attempts analyzing the feasibility of an evidence-based approach to represent reasonable information from complex and miscellaneous dataset.

APPROACH AND METHOD

Even though machine learning approach is establishing a hold in the diagnosis world, presently, there are no techniques that could consider a plethora of significant factors such as medical and personal history, smoking or drinking history by screening the patient and estimate the likelihood of developing the pancreatic cancer before its metastasis (Sharghi, 2019). Data analysis techniques depend on limited genomic data supplied to machine learning classifiers predicting an outcome. This can be exemplified in the study by Way et al. where a machine learning classifier detected altered Ras activity with promising results, but the data considered was limited to genomic evidence and transcriptome (Way G. P., et al., 2018). To overcome these shortcomings, a sophisticated and powerful mathematical calculus is needed to represent, combine, and draw inferences from such types of data and information. With the aid of available genetic material such as the mutations undergone by the genes, and other factors affecting the likelihood of developing pancreatic cancer, a screening option could be developed to detect whether an individual stands a greater risk of suffering from this cancer.

Multiple environmental and hereditary factors affect the pancreatic cancer detection. Evidential Reasoning (ER) approach draws information from these diverse factors and helps combine it to produce the likelihood of developing pancreatic cancer. Machine learning model is one of the diverse yet useful sources which analyzes genetic data to predict if a subject has developed pancreatic cancer or not. The machine learning (ML) classifier is based purely on analysis of genomic data, it lacks considering the impact of other factors responsible for genetic alterations such as the medical history or smoking and alcohol habits. Evidential reasoning allows us to incorporate all these factors along with the ML prediction, without knowing its

distribution even though some of the data may be imprecise, estimated, and incomplete. ML prediction is also affected by factors such as the quality of sequencing reads and the type of technology used to sequence the genomic data. Moreover, ML prediction is impacted by the amount of genetic material considered during the biopsy and its site and structure. Even though the ML classifier does not use these factors as its features, the ER model allows to amalgamate ML prediction results along with all of these factors and forms a consensus on the prediction of developing pancreatic cancer.

Heavy tobacco use constitutes around one fourth of cancer related deaths with a 5-6 fold increased risk of developing pancreatic cancer due to smoking a pack of cigarettes in a day (Pandol, Apte, Wilson, Gukovskaya, & Edderkaoui, 2012), and an elevated risk of about 1.5 to 6-fold due to drinking alcohol (Gupta, Wang, Holly, & Bracci, 2010). Thus, evidences of smoking and drinking histories are features that can be helpful in balancing the results of an otherwise unbiassed machine learning classifier, towards a higher belief of having pancreatic cancer (Sharghi, 2019). Moreover, DNA Single Nucleotide Polymorphisms (SNP) comprising of homopolymer region may lead to sequencing errors negatively affecting the accuracy of machine learning classifier using the genomic data. Thus, errors occurring while sequencing genomic data can be considered as another factor for orienting the machine learning results (Sharghi, 2019). In this way, the evidential reasoning approach can be used to make a more precise prediction about the pancreatic cancer diagnosis using mathematical calculus by looking past the traditional statistical and probabilistic approach (Sharghi, 2019).Thus, rather than relying on a single source of data, this approach takes into account various factors which may be diverse and imprecise. Each of these factors are assigned a belief and

are weighed together to make a final decision which measures the support in favor of pancreatic cancer prediction as well as no pancreatic cancer prediction.

The approach used in evidential reasoning model considers as input, factors such as results of a machine learning classifier, patient's medical history, family history, drinking and smoking history, the type of NGS technology used to obtain patient's genomic data, biopsy site and cell structure, sequencing read, and amount of genetic material. Every input factor is known as a frame and a collection of such frames is called a gallery.

Every frame comprises of propositions defining all the possible scenarios. No two scenarios can be true together at the same time. A probability signifying truth of the statement is assigned to each of these propositions. A discount value applied to these frames indicates the reduction of belief impact in that frame. Such a grid of interconnected propositions ultimately produces a single resulting prediction of likelihood of pancreatic cancer as shown in the Figure 1. Propositions for a given frame can either be discrete or continuous in nature. For instance, the frame 'Drinking History' may have propositions such as 'low', 'medium', or 'high', whereas the frame 'ML Prediction' may have propositions such as 'PC' (pancreatic cancer) or 'NOT_PC' (no pancreatic cancer). All the features (sequencing read, family medical history, smoking history, etc.) are represented as similar frames and are input to the evidential reasoning model. A compatibility relation is the subset of cross product between two frames meaning that it is the joint possibility distribution of the frames under consideration. The generalization of the compatibility relation is defined as 'the multiplication of possibility distributions' (Yager, Liu, Dempster, & Shafter, 2008). It is defined for every frame in the evidential model. A new body of evidence forms as a result of merging of frames into one another as per Dempster's Rule of Combination evidence (Yager, Liu, Dempster, & Shafter, 2008).

The propositions in this project are synthesized using NIH GDC dataset (see Appendix F for more evidential reasoning details).

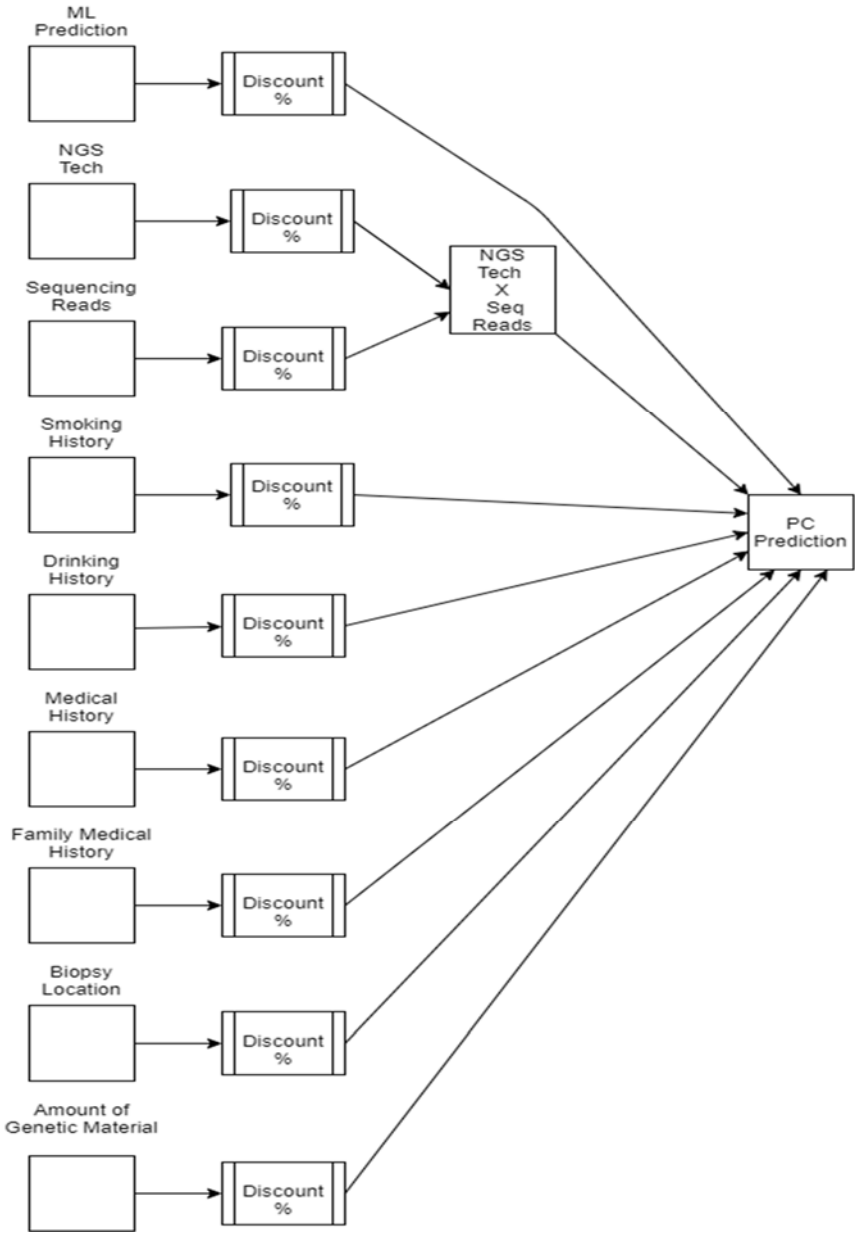

Figure 1. Evidential Reasoning Model

A machine learning model was developed to determine the classification whether a subject has or does not have pancreatic cancer. The results of the classification obtained from the developed machine learning model is one of the frames which are

provided to the evidential reasoning model. (For more information on how the classifier was built, please refer Appendix E)

EXPERIMENTAL METHOD

This study 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

In order to detect the cancer at an early stage, it is necessary to examine the type of mutations undergone by genetic data which may help in predicting the potential of developing the pancreatic cancer. A vast amount of genetic data from cases related to different kinds of cancer can be analyzed to infer the type of mutations the genes undergo that lead to pancreatic cancer. For this purpose, data from the National Cancer Institute (NIH) Genomic Data Commons (GDC) is taken into consideration. The NIH GDC portal contains information related to 22,872 genes in total, with 64 different projects which are different types of cancer. There are 3,142,246 total number of mutations associated with these genes and projects in the NIH GDC data. The experiment consists of data specific to all the cancer related projects. The GDC portal also associates the impacts with every genetic mutation. One of the important features of this experiment is the consideration of the impact caused by the genetic mutation (For more details related to the 'impact' field, see Appendix D). The machine learning classifier trained on the dataset considers the impact associated with every mutation record. To sum up the working of the machine learning classifier, features such as genes, and mutations along with their lethality and the disease associated it were considered to classification model predicting whether a subject suffers from pancreatic cancer or not. Since the features used are dependent variables that cannot be considered as separate features, they were combined into a single string feature. This was a binary classification model built using support vector machine classifier. The kernel 'Radial Basis Function' (RBF) was used as it

gives better cross validation results than linear or polynomial kernels. A cross validation with 10 splits and 3 repeats was applied on the training dataset. (Refer Appendix E to understand more details on how the machine learning model was built). This model produced ~85% average accuracy with a mean Receival Operator Characteristic – Area Under Curve (ROC) of ~0.834. It showed a decent class separability with a fairly good accuracy.

The method for evidential reasoning model was followed as discussed in this section. The evidential reasoning model will predict the belief of developing pancreatic cancer as well as not developing it based on the combination of propositions and their respective masses provided as input. the 'mass' of a specific proposition is a basic probability number assigned to it (Yager, Liu, Dempster, & Shafter, 2008). It can be interpreted as the amount of belief that one has in the proposition. (See Appendix F to learn more about the evidential model engineering). As the propositions and their masses change, so will the prediction results. While some combinations like high smoking history with a personal history of cancer will lead to increase the likelihood of diagnosing pancreatic cancer, other combinations like biopsy site away from pancreatic region with regular cell shape along with machine learning prediction of not developing pancreatic cancer is expected to produce a lower likelihood of pancreatic cancer development. Thus, the model will be tested based on the computed belief value in support of pancreatic cancer for the given scenario.

The baseline input and discount rates displayed in Table I and Table II are used to initialize the evidential reasoning experiments. Each of the frame is assigned a certain discount value indicating a relative importance of input features and/or reliability of the source of the information. Initial values were selected based on a subjective estimate of relative importance and credibility.

Table I
Baseline Propositions and Corresponding Support

Frames	Assigned Proposition	Support
ML_PREDICTION	NOT_PC	0.5
NGS_TECH	ionTorrent	0.5
SEQ_READ	LOW_GC_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 II
Discount Rates and Corresponding Frames

Frames	Discount Rate
ML_PREDICTION	0.1
NGS_TECH	0.2
SEQ_READ	0.1
NGS_X_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

The grouping of the baseline discount rate and initial inputs gives 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)

Certain experiments are conducted by modifying the baseline propositions and the support in them. Opinions are translated via compatibility relations and a consensus is formed using Dempster's rule. Dempster's rule can be repeatedly applied to the previous combination and combined further with an additional mass distribution if there is any, thus forming a new consensus (Wesley & Graham, Evidence-Based Decision Support For The Biopharmaceutical Industry, 2006). Ultimately, an evidential interval (EI) within the belief function calculus depicting the likelihood of developing pancreatic cancer is computed for each scenario. This is done by conveying, translating, and combining the opinions as mentioned before. Lower (*Spt*) and upper (*Pls*) bounds of an EI describe the degree of support attributed to the given proposition because of the current opinions (Wesley & Graham, Evidence-Based Decision Support For The Biopharmaceutical Industry, 2006). Note that [*Spt*, *Pls*] $\subseteq$ [0,1]. The evidential reasoning model is tested using the following combination of random variables to understand how each of the experiment affects the likelihood of developing pancreatic cancer:

- Machine learning prediction, smoking history, family health history
- Machine learning prediction, and drinking history
- Machine learning prediction, biopsy location, and amount of genetic material
- Machine learning prediction, sequencing technology utilized, and quality of sequencing read

In the experiment with combination of Machine learning prediction, smoking history, and family health history we change the propositions and masses assigned to these three frames and check how the increased belief in an individual with high smoking history and a family health history of having cancer increases support in the ML prediction of pancreatic cancer and increases the evidential interval of belief in an individual having pancreatic cancer. In other experiment, we combine the ML

prediction, drinking history, and personal health history frames to determine how the drinking history and personal health history are correlated to the ML prediction and how their proposition values and masses assigned to them affect the evidential interval. Further, we check how the biopsy location, and cell structure along with the quantity of genetic material affects the evidential interval and how they support the ML prediction as their values and masses change. Lastly, the impact on ML prediction with changes in NGS technology and quality of sequencing reads are tested. By changing their propositions and masses, we combine these two frames to evaluate that the ML prediction becomes less reliable, resulting in a lower evidential interval for pancreatic cancer as the sequencing reads become more error prone by having a high guanine cytosine (GC) content and high homopolymer regions.

EVALUATION OF RESULTS

The results are evaluated by training an SVC classifier and assessing its performance. A cross validation of 10-folds repeated 3 times was applied to the SVC classifier to measure the variance. An accuracy of 84.70% was achieved. The model was stable with a low standard deviation of +/- 1.297%. The mean ROC AUC of 0.834, while an average precision-recall score for this classifier was 0.67. Figure 2 and Figure 3 show the area covered by the receiver operating characteristic curve (AUC-ROC) and area covered by the precision-recall (AUC-PR) curve of the model.

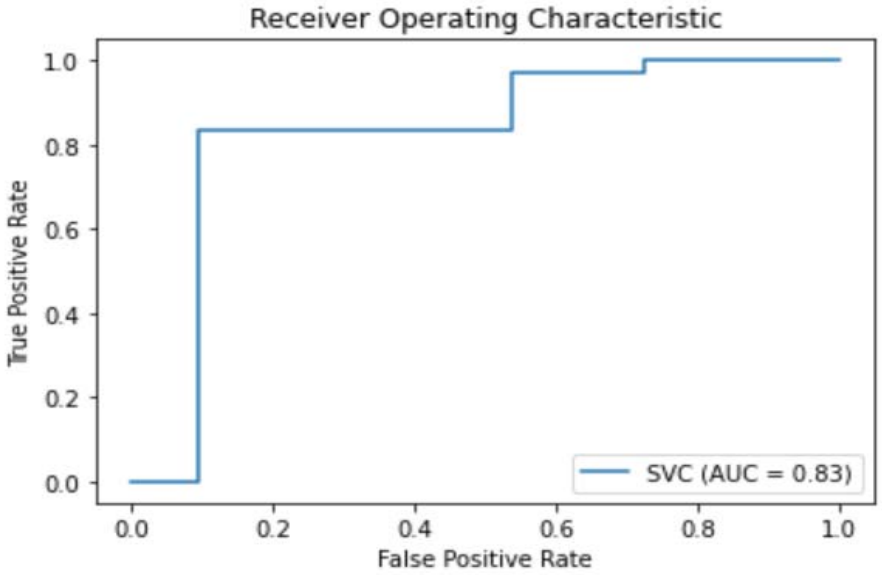

Figure 2. *Support Vector Classifier ROC*

While Figure 2 shows an AUC-ROC of 0.834 suggesting that the model has a good measure of separability as it is nearer to 1. Precision-Recall Curves are an intuitive measure when evaluating imbalanced dataset like the one for this study. In Figure 3, there is a consistent rise in precision as the recall increases. But a sharp drop in precision at around 0.8 recall indicates that there are large numbers of fall positives

at that point. This means that the predicted labels are incorrect when compared to the training labels. In the imbalanced dataset, since the minority class is the positive class, there could be a lot of negative examples that could become false positives. Conversely, a fewer positive examples could become false negatives, hence the steep drop in the precision.

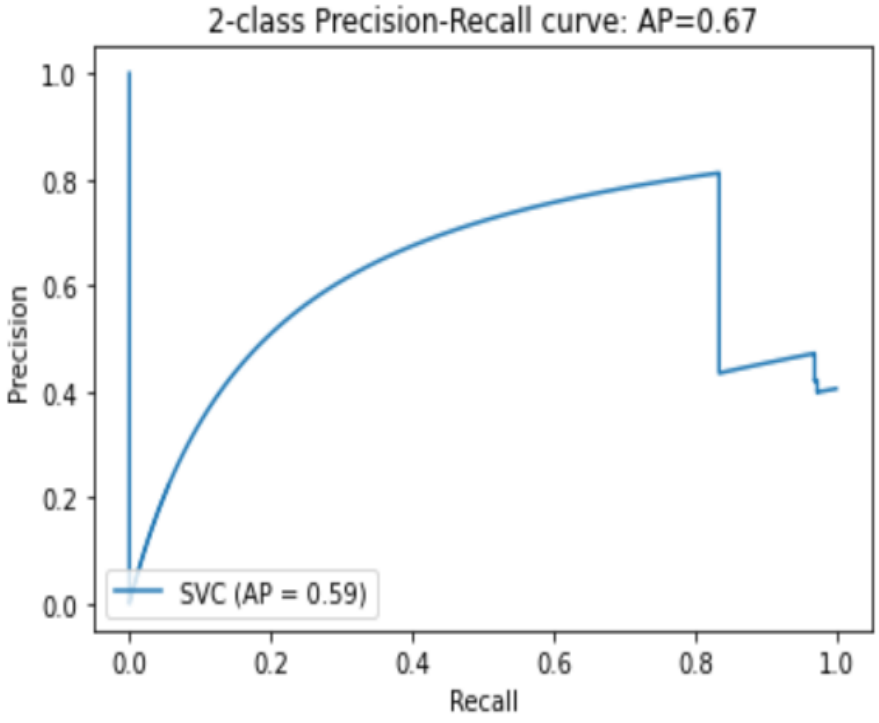

Figure 3. Support Vector Classifier Precision-Recall

Figure 4 and Figure 5 demonstrate the confusion matrix obtained without and with normalization. Higher proportion of true positives and true negatives suggest that the model performs a good classification of the test data. Although the results obtained for the machine learning classifier are promising, they still can be improved further if the GDC data contains more pancreatic cancer related records. For the model to avoid overfitting, there is a need to remove the imbalance present in the data. Efforts to do this were made, but the non-pancreatic cancer records will remain more compared to the pancreatic cancer project records which is quite obvious.

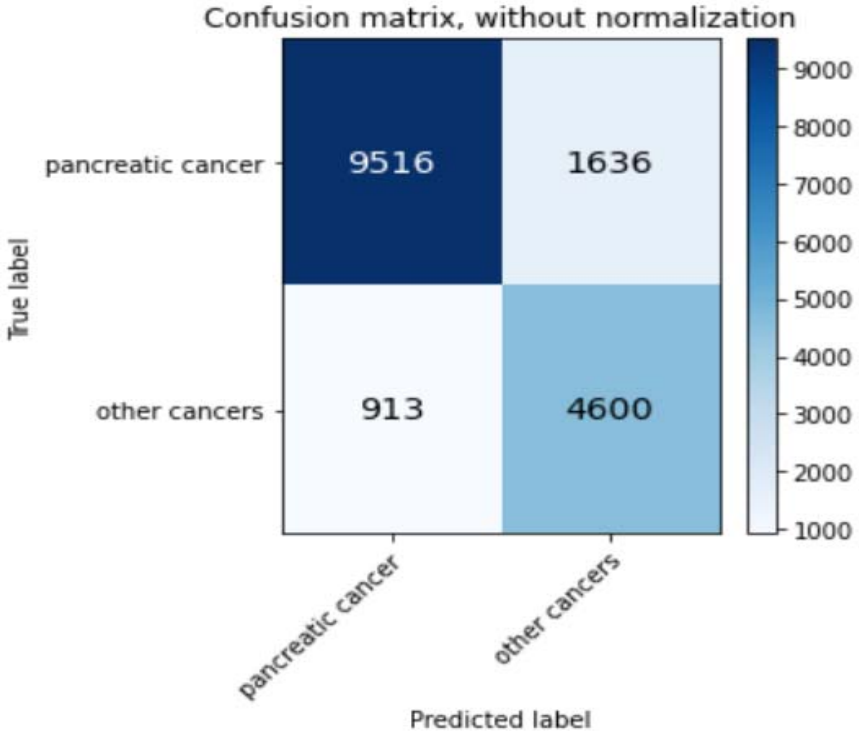

Figure 4. Confusion matrix, without normalization

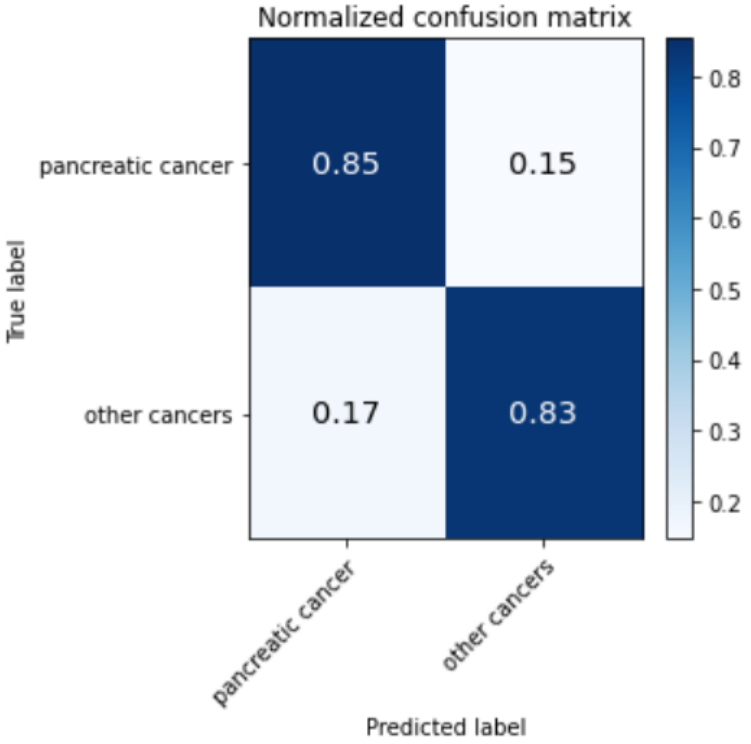

Figure 5. Confusion matrix, with normalization

ER Experiment 1

This experiment focuses on the effect of ML prediction frame on the evidential reasoning model.

- 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)

- The baseline input of ML prediction is changed from NOT_PC to PC with the same mass 0.5 to see how it affects the evidential interval.

  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)

  The results show what after changing from NOT_PC to PC, there is an increase in the evidential interval but since the mass assigned is only 0.5, it is a small increase.

- Suppose we do not have any information about other frames but just the results of the ML prediction. In this scenario, the mass of the ML prediction frame is set to be 0.85 (accuracy of ML classifier) with the proposition PC without changing the baseline beliefs in other propositions to see the effect on evidential intervals.

  Belief Of Having Pancreatic Cancer Lies Between: (0.507, 0.585) (0)|-----*----|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.415, 0.492) (0)|----*-----|(1)

  Now the evidential interval of having pancreatic cancer further increases, while the evidential interval of not having pancreatic cancer decreases as expected.

Thus, it is observed that if the belief in the ML prediction is changed from NOT_PC to PC, while keeping the rest of inputs constant, the evidential interval of pancreatic cancer increases and the interval for not pancreatic cancer decreases to a certain

extent eventually both attaining almost similar values when we are considerably confident about the ML prediction with mass 0.85.

ER Experiment 2

This experiment evaluates the effect of smoking history along with family medical history on the likelihood of developing pancreatic cancer. Assuming that a person does not have an active smoking history and moreover there is no family history of cancer, it is highly unlikely that the person would develop pancreatic cancer. On the other hand, if the smoking history is high with previous family history of cancer, the evidential reasoning model is expected to predict more likelihood of developing pancreatic cancer.

- 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)

- ML Prediction is set to PC with mass 0.85, smoking history set to HIGH with mass 0.9 along with family history set to CANCER with mass 0.8 to check if the high smoking and family history of cancer causes supports the ML prediction of having pancreatic cancer.

  Belief Of Having Pancreatic Cancer Lies Between: (0.956, 0.974) (0)|---------*|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.026, 0.043) (0)|*---------|(1)

  The evidential interval rises significantly showing that the smoking history and family history does support the ML prediction of having pancreatic cancer as was expected.

- Family history is set to NO_CANCER with a mass 0.9 but Smoking History is still HIGH with mass 0.9 to check how the prediction is affected for such scenario.

  Belief Of Having Pancreatic Cancer Lies Between: (0.686, 0.721) (0)|------**--|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.279, 0.313) (0)|--**------|(1)

  Such a combination reduces the evidential interval of having pancreatic cancer as was expected because there is no family history of cancer.

- Smoking History set to MEDIUM with a mass of 0.9 and Family Medical History set to CANCER with a mass 0.5 and ML prediction of NOT_PC with mass 0.5 to observe how setting affects prediction interval.

  Belief Of Having Pancreatic Cancer Lies Between: (0.302, 0.454) (0)|---**-----|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.545, 0.697) (0)|-----**---|(1)

  The evidential interval of pancreatic cancer reduces because the ML prediction is set to NOT_PC and even the smoking history is medium but only the family has a medical history.

- ML Prediction changed to PC with mass 0.85, Smoking History set to LOW with a mass of 0.9 and Family Medical History set to NO_CANCER with a mass 0.9 to check if the interval for PC prediction reduces.

  Belief Of Having Pancreatic Cancer Lies Between: (0.214, 0.248) (0)|--*-------|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.752, 0.785) (0)|-------*--|(1)

  The interval for PC prediction reduces as expected.

From the above experiments, it is observed that as the smoking history changes from LOW to HIGH, Family Medical History from NO_CANCER to CANCER, the evidential interval increases rapidly. It is seen that the ML Prediction results of PC further support increasing the interval. As the variations are made in the propositions in Smoking History and Family Medical History, there are changes in the interval depending also on the ML prediction mass. With HIGH Smoking History with mass 0.9 but Family History of NO CANCER with mass 0.9 the evidential interval

becomes (0.686, 0.721) as the ML Prediction is still set to PC with mass 0.85. If the Smoking History is reduced to MEDIUM with mass 0.9 but with Family History of CANCER and ML prediction of NOT_PC, the interval reduces to (0.302, 0.454). Eventually, with LOW Smoking History with mass 0.9 and NO CANCER Family History of 0.9 along with ML prediction set to PC with mass 0.85, the interval further reduces to (0.214, 0.248) suggesting that the propositions and their masses affect the prediction of evidential reasoning model as expected and the ML prediction is playing a supportive role in this.

ER Experiment 3

This experiment will try to analyze how the personal medical history, drinking history and the ML prediction correlate with each other. If a person does not have any history of cancer and has a low drinking history, the chances that he will develop pancreatic cancer are low. But with the ML prediction of pancreatic cancer and a medium to high drinking history and personal history of cancer, the expected evidential interval of developing cancer is high.

- 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)

- ML Prediction is set to PC with mass 0.85, to see if there is a change in the prediction given that the drinking history set to HIGH with mass 0.8 and Patient_Med_History set to CANCER with mass 0.9

  Belief Of Having Pancreatic Cancer Lies Between: (0.795, 0.845) (0)|-------**-|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.155, 0.204) (0)|-**-------|(1)

- The evidential increases as per expectations and thus the high medical history and drinking history support the PC prediction.

- Drinking history set to MEDIUM with mass 0.3 and Patient_Med_History set to CANCER with mass 0.7 given the ML prediction of PC to see how this affects the evidential interval.

  Belief Of Having Pancreatic Cancer Lies Between: (0.612, 0.705) (0)|------**--|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.295, 0.387) (0)|--**------|(1)

  The evidential reasoning interval reduces a little as only drinking history was changed which was expected.

- Drinking history set to HIGH with mass 0.9 and Patient_Med_History set to NO_CANCER with mass 0.9 and PC prediction to NO_CANCER with mass 0.8 to see if the prediction is impacted.

  Belief Of Having Pancreatic Cancer Lies Between: (0.202, 0.258) (0)|--*-------|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.742, 0.797) (0)|-------*--|(1)

  The evidential interval for pancreatic cancer reduces as per expectations.

- ML Prediction is set to NOT_PC with mass 0.8, to see if the interval for prediction decreases given the drinking history is set to LOW with mass 0.2 and Patient_Med_History set to NO_CANCER with mass 0.9

  Belief Of Having Pancreatic Cancer Lies Between: (0.02, 0.05) (0)|*---------|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.95, 0.979) (0)|---------*|(1)

  This causes the pancreatic cancer evidential interval to reduce sharply as was expected.

Here, different scenarios are created to evaluate the effect of drinking history and personal medical history on the evidential intervals. With ML prediction as PC with a mass 0.85 and Drinking History set to HIGH with mass 0.8 and Personal Medical History set to CANCER with mass of 0.9, there is an upsurge in the evidential interval of belief in developing pancreatic cancer from (0.063, 0.147) to (0.795, 0.845). With MEDIUM Drinking History with a mass of 0.3 the interval decreases by a small amount to (0.612, 0.705). Finally, as the Drinking History is set to LOW with mass 0.2 and the Medical History to NO_CANCER and ML Prediction to

NO_PC both with masses 0.8, there is a decline in the evidential belief interval to (0.02, 0.05), thus suggesting that the model behaves as was expected.

ER Experiment 4

In this experiment the impact of biopsy site and amount of genetic material is observed on the ML prediction and how it affects the evidential reasoning model prediction. It is expected that if the biopsy location is near pancreas with irregularity in the cell result, then there might be a possibility of greater risk of cancer than if the biopsy location is far off pancreas with regular cell result and low amount of genetic material. With lesser amount of genetic material, it might be not enough DNA material available to make a sensible machine learning prediction as opposed to when the genetic material is more. For a greater likelihood of developing pancreatic cancer, it is expected that the amount of available genetic material is high and the cells nearing the pancreatic region are irregular.

- 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 0.6 and change biopsy site and cell result to NEAR PANCREAS and REGULAR with a mass of 0.8 and ML Prediction to PC with a mass of 0.5 to see how it affects the PC interval.

  Belief Of Having Pancreatic Cancer Lies Between: (0.174, 0.251) (0)|-**-------|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.749, 0.825) (0)|-------**-|(1)

  The evidential interval for PC decreased as per the expectations as the amount of genetic material is large along with biopsy site near pancreas with regular the cell structure.

- The proposition amount of genetic material is changed from LARGE to MEDIUM with a mass of 0.7 and change biopsy site with cell result to NOT NEAR PANCREAS and IRREGULAR with mass of 0.9 to see if the PC prediction is changed given ML Prediction is set to PC with a mass of 0.5

  Belief Of Having Pancreatic Cancer Lies Between: (0.37, 0.531) (0)|---***----|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.469, 0.63) (0)|----***---|(1)

  The PC interval increases moderately between 0.37 and 0.53 as was expected given the modified propositions.

- Amount of genetic material is set to SMALL with a mass of 0.8 and biopsy site and cell result set to NEAR PANCREAS and IRREGULAR given a mass of 0.9 and ML Prediction to PC with a mass of 0.85 to see if the evidential interval for having PC supports the input by increasing

  Belief Of Having Pancreatic Cancer Lies Between: (0.838, 0.88) (0)|--------*-|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.119, 0.162) (0)|-*--------|(1)

  The evidential interval for PC prediction surges as expected.

Observation of this experiment suggest that the ML prediction of PC are not strongly supported by a LARGE amount of genetic material as well as the results of biopsy NEAR pancreas with REGULAR cell results. Further, with a MEDIUM quantity of genetic material with mass 0.5 along a mass of 0.9 assigned to the biopsy site and cell result of NOT NEAR PANCREAS and IRREGULAR and PC prediction of mass 0.5 increases the pancreatic cancer evidential interval from (0.174, 0.251) to (0.37, 0.531). This interval further increases as experiments are performed with different combinations changing the propositions eventually to soar to an interval of (0.838, 0.88) when the amount of genetic material is changed to SMALL with a mass of 0.8, biopsy site and cell result to NEAR PANCREAS and IRREGULAR with a mass of 0.9 and ML Prediction to PC with a mass of 0.85.

ER Experiment 5

Lastly, this experiment tries to observe the impact of sequence reads quality along with the NGS technology used on the ML model. The anticipated scenario is that the ML should become less dependable as the sequencing reads become more erroneous meaning that they produce high guanine cytosine (GC) content and high homopolymer regions.

- 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)

- The proposition for sequencing reads is set to HIGH GC HIGH HMR with a mass of 0.9 to check to see if there is a change in prediction given the ML prediction of PC with mass 0.85.

  Belief Of Having Pancreatic Cancer Lies Between: (0.332, 0.435) (0)|---**-----|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.565, 0.667) (0)|-----**---|(1)

  It is seen that the evidential interval obtained for pancreatic cancer is moderately low suggesting error prone sequencing reads make the ML prediction unreliable and produce lower interval which is expected.

- The mass assigned to the proposition for NGS tech is set to 0.9 to see if there is a change in prediction given a high GC count and high homomeric region.

  Belief Of Having Pancreatic Cancer Lies Between: (0.332, 0.436) (0)|---**-----|(1)
  Belief Of Not Having Pancreatic Cancer Lies Between: (0.564, 0.667) (0)|-----**---|(1)

  There is no change in the evidence interval showing lack of impact of the NGS tech frame.

- The proposition for NGS tech is changed from ionTorrent to ILLUMINA to check if there is a change in the prediction given a high GC and high homomeric region.

Belief Of Having Pancreatic Cancer Lies Between: (0.332, 0.436) (0)|---**-----|(1)

Belief Of Not Having Pancreatic Cancer Lies Between: (0.564, 0.667) (0)|-----**---|(1)

No change is observed in the evidential interval by changing the type of NGS technology.

- The proposition NGS tech is changed back to ionTorrent and sequencing read is set to MOD GC LOW HMR with mass 0.9 given an ML prediction of PC with mass 0.85 to see if the interval changes

Belief Of Having Pancreatic Cancer Lies Between: (0.768, 0.804) (0)|-------**-|(1)
Belief Of Not Having Pancreatic Cancer Lies Between: (0.195, 0.231) (0)|-**-------|(1)

The evidential interval for PC increases considerably from (0.332, 0.436) to (0.768, 0.804) as expected because lower quantity of GC and homomer regions produce less faulty sequencing readings thus making the ML prediction more reliable.

Changing the ML prediction from NO_PC to PC with a mass 0.85 increases the evidential interval from (0.063, 0.147) to (0.507, 0.585). Further, as the sequencing read frame is changed from LOW GC LOW HMR to HIGH GC HIGH HMR, the evidential interval of belief in developing pancreatic cancer decreases to (0.332, 0.435). The model behaves as expected, meaning that that this change in the sequencing model leads to the possibility of the sequencing read may lead to errors, effectively making the ML prediction less reliable. Another observation is that changing the NGS tech mass or its value from ION TORRENT to ILLUMINA hardly affects the evidential interval. Even changing the sequencing read to LOW GC HIGH HMR does not affect the interval. However, an upsurge in the evidential interval of belief of having pancreatic cancer is observed as the NGS tech value is change to MOD GC LOW HMR with a mass of 0.9 to (0.768, 0.804). Finally, after changing the sequencing read 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 (0.768, 0.804).

Thus, after performing these experiments by changing the propositions and masses assigned to them, the evidential reasoning model behaves as per expectations and produces satisfactory results. Basic knowledge between the relationships of the input factors is needed for verification of the behavior of the model based on the experiment conducted.

CONCLUSION AND DISCUSSION

The machine learning classifier used in this project is an improved version of the previous study where the data related to all the projects in the NIH GDC portal is considered ensuring the variability and heterogeneity of data. Although the previous research by Sharghi reported an SVC prediction accuracy of ~91%, it was achieved without considering the genes and mutations across to all the cancer projects which are a part of the GDC portal. Further, the previous study did not consider the lethality of every mutation. It was trained on a limited dataset of gene-mutation combinations which occurred in only 185 cases of the TCGA-PAAD project.

The present study attempts to overcome these limitations by considering the genomic data with a high 'VEP' impact across all the cancer projects on GDC portal. Results achieved in this study have greater fidelity despite a lower prediction accuracy than what Sharghi's model achieved. The model is also very stable as it has a very low standard deviation. Thus, consideration of the most lethal and more extensive genomic data makes this model a better version of previous research.

Further, based on given inputs, the observed results of the evidential model altered as per the expectations. The experiments conducted presented a positive hope that the evidential model can be used as an effective tool in the early detection of pancreatic cancer. Factors such as accessibility to real sequencing data, accurate family and personal history, along with more powerful and accurate NGS technology are crucial in confirming the feasibility of this approach. An amalgamation of an evidential reasoning approach with machine learning can prove to be a potential solution in early diagnosis of pancreatic cancer.

FUTURE WORK

Scope of this project included sequencing of DNA from real pancreatic tissue in the Evidential Reasoning Model. Unfortunately, the unexpected circumstances prevalent due to COVID-19 prevented acquiring real pancreatic cancer sequence data from the California State University (CSU) East Bay Campus. Even though the NIH Cancer database has protected real pancreatic cancer sequence data, present COVID-19 circumstances has reduced their operational staff and has suspended processing applications to obtain access to such sequence data. Thus, the future scope of this project includes gathering the real sequencing data and incorporating it in the developed evidential reasoning model. The source of the genomic data for this project was The Cancer Genome Atlas Program (TCGA) which, even though being a reliable source, has a limited database of 185 cases in the pancreatic adenocarcinoma project. Hence, in order to improve the efficiency of the machine learning classifier, it is desirable to seek other authentic sources to gather real pancreatic cancer data. There is imbalance in the data currently being considered as cases related to pancreatic cancer constitute a small proportion in the overall dataset evaluated as opposed to non-pancreatic cancer records. Even though this project considers sampling of the records and assigning ‘class weight’ to the model to handle this imbalance, the model still overfits to a certain degree. Thus, the next phase of the project should concentrate on removing the imbalance in the data. The SVC was used as it traditionally has a good accuracy record. Future work may involve exploring other machine learning algorithms. Moreover, to improve the evidential reasoning model, the information related to the verified type of NGS technology should be used to access genetic data. Along with the real sequencing data, efforts should be made to acquire the real personal and family medical history along with

the smoking and drinking history of individuals in order to enhance the evidential reasoning model.

REFERENCES

Amin, S., & DiMaio, C. J. (2016). Pancreatic Adenocarcinoma. In *Pancreatic Masses* (pp. 11-20). Springer, Cham.

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). Cellular and molecular aspects of pancreatic cancer. *Acta Histochemica*.

Gharibi, A., Kim, S. L., Molnar, J., Brambilla, D., Adamian, Y., Hoover, M., . . . Kelber, J. A. (2017). ITGA1 is a pre-malignant biomarker that promotes therapy resistance and metastatic potential in pancreatic cancer. *Scientific Reports*.

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). Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. *BMC Bioinformatics*.

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., N., Krapcho, M., Garshell, J., Neyman, N., Altekruse, S., Kosary, C., Cronin, K (2016). *SEER Cancer Stat. Rev.* Bethesda: National Cancer Institute. Retrieved from SEER Cancer Statistics Review, 1975-2010, National Cancer Institute. Bethesda, MD: https://seer.cancer.gov/archive/csr/1975_2010/

Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., . . . Kim, P. M. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine*.

Kelber, J. A., Reno, T., Kaushal, S., Metildi, C., Wright, T., Stoletov, K., . . . Klemke, R. L. (2012). KRas Induces a Src/PEAK1/ErbB2 Kinase Amplification Loop that drives metastatic growth and therapy resistance in pancreatic cancer. *Cancer Res. col. 72*.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., . . . R, A. H. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature 499*, 214–218.

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

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). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 341.

Risch, H. A., Yu, H., Lu, L., & Kidd, M. S. (2005). Detectable Symptomatology Preceding the Diagnosis of Pancreatic Cancer and Absolute Risk of Pancreatic Cancer Diagnosis. *American Journal of Epidemiology*, 182(1), 26-34.

Sharghi, O. (2019). *Towards Early Detection of Pancreatic Cancer.* San Jose .

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

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. (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.

Wesley, L., Graham Kim (2006). Evidence-Based Decision Support for the Biopharmaceutical Industry, *International Institute of Informatics and Systemics*, 312-218.

Witkiewicz, A. K., McMillan, E. A., Balaji, U., Baek, G., Lin, W.-C., Mansour, J., . . . Knudsen, E. S. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nature Communcations 6*, 6744.

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). Improving Indel Detection Specificity of the Ion Torrent PGM Benchtop Sequencer. *PLos ONE*.

Appendix A

PDAC originates in the ducts of pancreas, meaning that the cells present in the small tubes which line the periphery of pancreas undergo abnormal growth, leading to pancreatic cancer of this type. More precisely, the DNA in these cells undergo mutations causing the anomalous proliferation of the ductal cells which is uncontrollable. S. Amin and C. J. DiMaio describe the pathology of pancreatic adenocarcinoma as the cancer characterized by solid and firm tumors that are highly infiltrative (Amin & DiMaio, 2016). This study also mentions that this tumor invades beyond the main tumor before its diagnosis and thus, the cancer spreads outside of pancreas.

The potential causes or risk factors of pancreatic cancer are broadly classified into 'Environmental' and 'Inherited' (Gharibi, Adamian, & Kelber, 2016). Frequent cigarette smoking and alcohol can be environmental major risk factors. Pancreatic cancer is observed five to six times more frequently with individuals having a high smoking history (Pandol, Apte, Wilson, Gukovskaya, & Edderkaoui, 2012). Individuals with strong drinking history are reported to be prone to develop pancreatic cancer with an elevated risk up to1.5 to 6-fold than the individuals with no drinking history (Gupta, Wang, Holly, & Bracci, 2010). Diabetes mellitus and elevated Body-Mass Index (BMI) may also be a potential cause of pancreatic cancer. In case of inherited factors, Familial Pancreatic Cancer (FPC), a condition in which at least two first-degree family members have pancreatic cancer may lead a person to develop this disease. Other hereditary factors include Lynch syndrome, Peutz–Jeghers syndrome (PJS), hereditary breast-ovarian cancer, Familial atypical multiple mole melanoma (FAMMM), and Familial adenomatous polyposis (FAP) which create a great probability that a person may develop pancreatic cancer having any of these conditions inherited or in his / her family history. In addition, available options

to treat the PDAC are radiation therapy, chemotherapy, removing the tumor by surgery, or more than one of these techniques in combination (Gharibi, Adamian, & Kelber, 2016).

Appendix B

Visualization techniques such as computer tomography (CT) and magnetic resonance imaging (MRI) are primarily used as the initial steps in pancreatic cancer evaluation after patients start showing symptoms. These techniques are commonly known as ultrasonography imaging of the abdomen. Some of the other visualization methods are magnetic resonance cholangiopancreatography (MRCP), or endoscopic ultrasound (EUS) used with CT and MRI. As per a study, EUS and fine-needle aspiration (FNA) biopsy of the mass is mostly undergone by patients as EUS is highly accurate while detecting small tumors of <=2 cm and focal lesions (De La Cruz, Young, & Ruffin, 2014). EUS is specially known for detecting tumors less than 10 mm with a sensitivity of 84% to detect 25 small tumors of 10mm coupled with a case where EUS-FNA was used to detect masses with a size less than 10mm in 23 patients had an accuracy of 96% (Hijioka, et al., 2017). However, detection of cancer should be before the visualization of pancreatic masses is possible even though visualization is a powerful aid in diagnosing pancreatic masses in order for a longer 5-year survival rate.

Appendix C

Sequencing is the method to determine the sequence of nucleotide bases of genome or exome which can be performed on DNA or RNA (Gharibi, Adamian, & Kelber, 2016). Sequencing has enabled researchers to distinguish between normal and abnormal tissues by analyzing he nucleotide bases of genomic and transcriptomic variations, thus assisting in identification of cancerous tissues (Sharghi, 2019). Next-Generation Sequencing (NGS) is the terminology for contemporary sequencing technologies.

As compared with digital polymerase chain reaction (PCR) method, detection of the CTCs using NGS techniques seems to be less sensitive. Analysis of ctDNA / ctRNA is better done with sequencing with high coverage. Sequencing data is also affected by attributes of the NGS technique used such as read length, depth of coverage, etc. It plays a major role in defining the accuracy and precision of the data which is to be sequenced. A large volume of chromosome loci can be evaluated using NGS. There are some disadvantages when using whole exome sequencing (WES). Identifying copy number alteration (CNA) can be negatively impacted by WGS methods. It is challenging to identify noncoding variants and rearrangements affecting gene regulation when whole exome sequencing (WES) is used over WGS (Wesley L. , 2019). Court et al. found out an ADO (allele drop out) rate of 85% and the reason behind the failure of most sequencing cases to be WGA (Court, et al., 2016).

Illumina is another powerful sequencing technology known for its highly accurate and precise throughput (Sharghi, 2019). Even though Illumina is known to have biases along with 1-2% error rate, it is believed to be used 90% of the times while sequencing. Erroneous results are accountable to reasons such as crosstalk, phasing, fading, and T accumulation, where substitution errors lead over insertion/deletion

errors (Heydari, Miclotte, Van de Peer, & Fostier, 2019). Biases include concentrated errors towards ends of DNA reads, whereas substitution errors happen with incorrect detection of a base near the end of a sequence. Homopolymer errors are commonly occur in the site of true polymorphism regions along with a case resulting in reoccurrence of same nucleotide. It is necessary to reduce the sensitivity of technology if these errors are to be minimized (Yeo, et al., 2012). While comparing sequencing technologies, it was observed that Illumina sequencing produced errors in the analysis of long polymers > 20 bases, whereas IonTorrent sequencing methodology could not accurately predict bases in homopolymers > 8 bases nor could it read homopolymer regions > 14 bases (Quail, et al., 2012).

Appendix D

Each mutation impacts the protein differently. As per the NIH GDC portal, every variant impacts the mutated protein in a certain way. The nature of its impact is observed and is distinguished based on its effect on the protein. The impact associated with the mutation is categorized into VEP, PolyPhen, and SIFT. The scope of this project is limited to the VEP category. The VEP impact is the effect on the structure and the behavior of protein. It is further classified into 4 sub-categories which are high, medium, low, and modifier. High VEP impact indicates that the variant disrupts the protein in a way such that the protein may undergo truncation, decay, or function loss. A moderate impact means that the variant may change the protein effectiveness in a non-disruptive manner. On the other hand, low impact may not change behavior of the protein and is harmless. Lastly, a modifier impact relates to the non-coding variants, that do not leave any evidence of impact by modifying non-coding genes.

Appendix E

The machine learning classifier was developed to classify the mutations leading to pancreatic cancer from other cancerous mutations. An SVC was trained on the data from GDC portal. Different factors such as the gene, genetic mutation in that gene, the project (or the disease) in which its occurrence is observed along with the impact of the variant on the affected genome were considered. The data was taken for all the genes that are available on the GDC portal which resulted into about 2.8 million records. The records with the value ‘TCGA-PAAD’ in the project column were the targets with value ‘1’ whereas the records with any other project were labelled as ‘0’. To obtain this data, a program was built which initially queried all the genes in the GDC portal. For every gene, the module queried GDC data portal for all mutations associated with the specified gene. With every mutation, the project in which it occurred was also associated. Every mutation impacts the genome differently, this data is also captured in the portal. Impact is categorized into ‘VEP impact’, ‘SIFT impact’, and ‘PolyPhen impact’ and this dataset only considers ‘VEP’ impact of the genetic mutation. (See Appendix D for more details on the impact field.) A script was developed to compute the VEP impact for each record. Since the data size was extremely huge, there was a need to consider only the data which was more credible to avoid consideration of data which added little value to the set of highly impactful mutations. Hence, the dataset was sorted as per the project the mutation was found in, based on its impact, and finally based on the genes. Out of the sorted dataset, 5% of data from each project was considered in the final dataset which ensured all the highly impactful records from every project were considered. This technique ensured heterogeneity of the data and focus on the mutations with either high, or moderate impact. In this dataset, there was an imbalance as the records belonging to all other projects were way more than the ones belonging to pancreatic

cancer project. The model was built by using the support vector classifier with the ‘rbf’ kernel, and the class_weight parameter equal to ‘balanced’. To measure the variance, a cross validation of 10 folds and 3 splits was applied. Since there was a large difference in the proportion of pancreatic cancer records and non-pancreatic cancer records, the imbalance in the data led to overfitting of the model. To balance the data, 2% records from every project and all the records belonging to TCGA-PAAD project were considered as the final dataset and the same class classifier was trained on it. L2 penalty or Ridge regularization was further used to evaluate the overfitting. The data distribution of the final dataset considered is shown in the Figure 6. The data selected was found to be normally distributed.

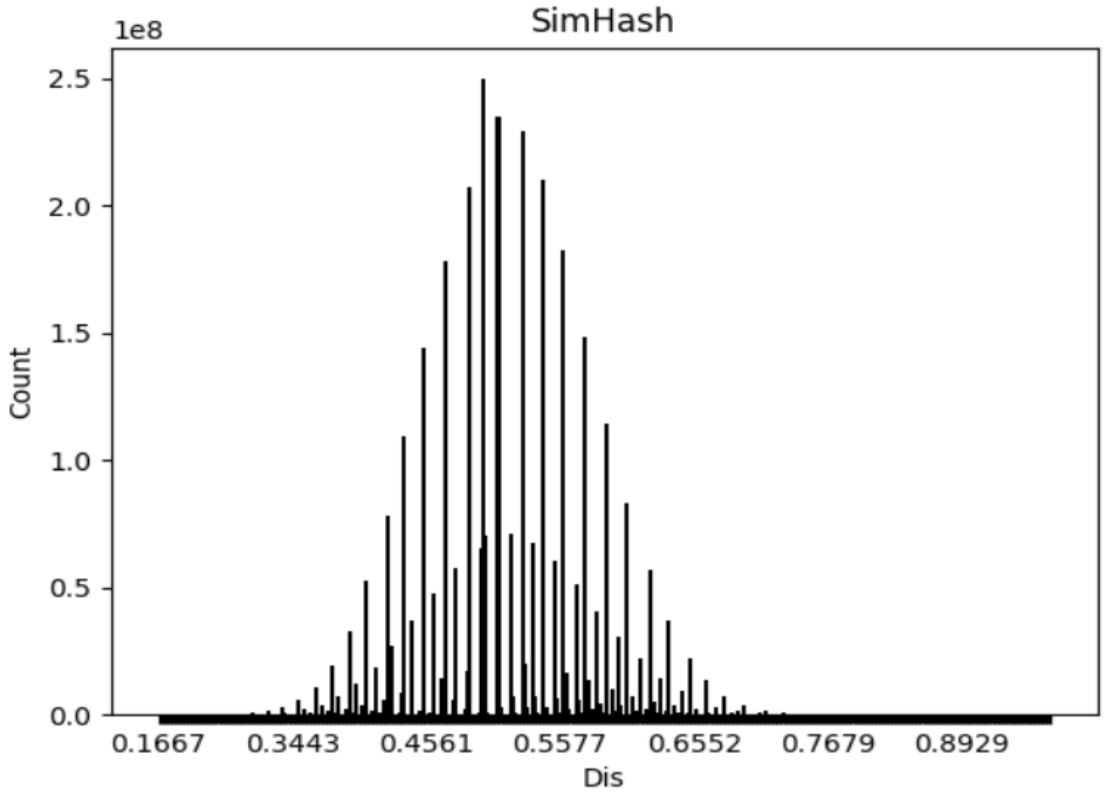

Figure 6. Distribution of the final dataset

Appendix F

The evidential reasoning model consists of various input features such as ML prediction, smoking history, medical history, drinking history, biopsy site, NGS technology used, etc. Every input feature is considered as a frame and for every frame has certain possibilities which are called as propositions. A frame is created for every factor and saved in a text file called gallery_input by defining its name, propositions, its data type such as continuous or discrete, the original frame called as parent frame which is the source of origin of the current frame, the frame in which the current frame will merges into called as the result frame, and lastly the compatibility relations of the frame with other propositions. Each proposition is assigned a certain mass which are all stored in a mass distribution file. To lessen the impact of certainty or credibility in the belief of a frame, a discount is assigned to every frame which is also stored in a text file. Using Dempster's Rule, the frames undergo fusion to form a new body of evidence which can be further used for fusing with other bodies of evidence (Yager, Liu, Dempster, & Shafter, 2008). The final output is computed and displayed as an interval of evidence which designates the level of belief in the propositions which are provided as input scenarios. This is a result of propagation of fused frames which transfuses from start to the end frame.