Cancer Outcome Analytics Based on Initial Diagnosis and Cancer Type

Maria Lifshits and Sofia Lis
Department of Computer Science
Mount Holyoke College
South Hadley, MA 01075

Abstract

Cancer is a group of diseases with high mortality and without a proven cure. Multiple factors affect the cancer outcome, and the role in predicting the outcome has to be established. We used logistic regression and Bayesian models to determine the significance for each of the features in the open-source The Cancer Genome Atlas Program dataset. Using the developed models, we achieved 75% accuracy in predicting the outcome, and 78% accuracy in predicting the vital status, surpassing the doctors by 16%. Future research, along with further data collection and development of the existing models, can significantly improve the accuracy of cancer outcome prediction.

1 Introduction

In the past century, our society underwent several significant changes. The developments in the healthcare system and higher quality of life increased life expectancy (Ho et al., 2018). Due to the sporadic nature of cancer, higher life expectancy is correlated with a higher prevalence of cancer in the population (Pompei, 2001). The improvements in the medical equipment and testing techniques improved the diagnostic accuracy for cancer. The symptoms that would go undetected before are more likely to lead to a successful diagnosis today. People became more exposed to toxic elements in the form of food, radiation, and various environmental changes that affect air and water (Haster et al., 2015; Wang et al., 2012). The higher number of cancer diagnoses made the research of potential treatments and diagnostics more crucial than ever and attracted more attention and funding to cancer research in the recent past.

Usually, the doctors choose an appropriate patient treatment based on the information about the biology of a specific cancer type and its causes, hoping that it will lead to complete recovery. However, it is extremely difficult to predict the outcome based on the choice of treatment alone (Michiels et al., 2005; Shi & Sargent, 2009; Catto et al., 2003). We argue that such outside factors as age or gender, play a crucial role in determining the disease outcome as well as the treatment protocol. Despite the outstanding developments in multiple areas of cancer research, we still cannot quantify the effect that covariates have on the cancer outcome in a reliable and consistent manner. In order to make accurate predictions, it is vital to understand the underlying biology of cancer and engineer the features appropriate to the specific cancer type.

For our project, we used the clinical dataset that was recently created by The Cancer Genome Atlas Program (TCGA) to drive high-quality survival outcome analytics (Liu et al., 2018). Table 1 demonstrates the abbreviation used throughout the paper as well as basic background information about the cancers. The cancers chosen for this study vary in prevalence and survival rates, allowing us to evaluate the efficiency of our models for diseases with various properties and disease courses. Though the specifics differ between cancers, including the role of the genetic component, location, type of cells affected as well as other properties, they share the same pattern. Due to a sporadic or

inherited mutation in the DNA, the affected cells divide uncontrollably, often with the loss of function and spread into surrounding tissues (Warburg, 1956).

In our study, we will generalize the factors that affect the outcome and compare them across the cancers. Understanding the contributing role of the factors other than the treatment protocol could help medical professionals enhance their decision-making process, thus leading to better outcomes and higher survival rates in cancer patients.

Table 1: Descriptions of the studied cancer types

Name	Abbreviation	Description
Urothelial Bladder Carcinoma	BLCA	A type of bladder cancer. 80,000 people get yearly diagnosed with it with a 77% 5 year survival rate (Dinney et al., 2004)
Head-Neck Squamous Cell Carcinoma	HNSC	A type of squamous cell cancer. 65,630 people get yearly diagnosed with it with a 40-50% 5 year survival rate (Sarini et al., 2001)
Low Grade Glioma	LGG	A type of brain tumor (not a cancer, though often studied alongside cancers). 24,600 people get yearly diagnosed with it. It is almost incurable in adults, with 5 year survival rate varying from 68% to 22% based on the age (Dougherty et al., 2010; Ostrom, 2016)
Lung Adenocarcinoma	LUAD	The most common lung cancer. 87,200 people get yearly diagnosed with it with a 70% 5 year survival rate (Beer, 2002)
Lung Squamous Cell Carcinoma	LUSC	The type of lung cancer. 65,400 people get yearly diagnosed with it with 24% 5 year survival rate (Masuda et al., 2012)
Ovarian Cancer	OV	21,750 people get yearly diagnosed with it with 47% 5 year survival rate (Cannistra et al., 2015)
Stomach Adenocarcinoma	STAD	The most common type of stomach cancer. 24,840 people get yearly diagnosed with it with 32% 5 year survival rate (Sehdev et al., 2013)
Uterine Corpus Endometrial Carcinoma	UCEC	The most common type of uterine cancer. 49,200 people get yearly diagnosed with it with 81% 5 year survival rate (Kurman, 1994)

2 Related Work

The closely related field of research is the survival analysis. It focuses on predicting the likelihood of a specific event occurring at a specific time. In biology and medicine, the traditional approach to survival analysis is several statistical models. Interestingly, most of the research papers on the topic use statistics and statistical software rather than more advanced technologies. One of the most flexible statistical models is the Kaplan-Meier estimator (Chen et al., 1982). However, though it reaches a very high accuracy, it does not allow to include the population data, only the person and the date of event occurrence. Therefore, it creates a survival curve that applies to the population as a whole, but not to a single person case. Another popular model, the Cox proportional hazard model, allows to incorporate the covariates, but is not flexible and assumes that the hazard rate is constant. Therefore, when it comes to predicting cancer survival, it does not provide very accurate results (Cox, 1972).

Recently, survival analysis became an important field of predictive analytics in machine learning (Alaa & van der Schaar, 2017; Lee et al., 2018). However, most existing models focus on survival analyses for market events (when and how likely a specific event will occur) and not on potential

medical applications. Since these models focus on the new fields with obvious commercial interest, they operate mostly on big amounts of data, which is almost unfeasible to acquire in the medical field. Even now, many hospitals do not have digitalized patient cases. To collect the necessary information, the researchers often have to go and physically scan and process the case files, which is a complicated process due to ethical concerns and privacy issues. Secondly, the medical field changes rapidly, and the standard of medical care, medical notes, and treatment protocols are not consistent from one hospital to another, especially across state lines. At the moment, collecting the amount of information large enough and applying one of the state-of-the-art survival analysis neural networks is impossible. Lastly, the existing models assume that the events are binary (it will occur, or it will not occur) or continuous (when will it occur) rather than several disjoint events, making the existing models not applicable to the existing problem.

In the few studies that addressed the problem of predicting cancer outcomes, the researchers chose genetic information or molecular markers as the input features, disregarding the environmental factors that are affecting the outcome. Additionally, these studies have not developed a consistent methodology that could be applied across the domain to work with different cancer types, which limits the potential applications of their work (Michiels et al., 2005; Shi & Sargent, 2009; Catto et al., 2003).

Our research addresses the issue of small amounts of data and thus improves on the existing models. Additionally, we worked on creating a model that would be able to predict both the overall survival and cancer outcome across several cancer types if provided with correctly formatted data.

3 Dataset

There is a limited number of open-source datasets that have more than one type of cancer. The existing cancer datasets usually have few features, limited to gender, race, age, and year of diagnosis. However, a new dataset was published in 2018 as a result of The Cancer Genome Atlas Program (Liu et al., 2018). This dataset contains data points for 11,160 patients over 33 cancer types. However, only 14 of the cancer types contained more than 300 data points, and out of these 14, only 8 contained the information about the outcome (Table 1). We proceeded with them and made sure to delete the cases that were marked as “inconclusive” by the researchers. Each of the cancers had its own set of input features, presented in Table 2.

Each of the features represents an important characteristic of cancer, and while some are quite straightforward, such as gender, the others require further explanation. The gender differences are known to affect the cancer outcomes due to the hormones and their effect on the body (Micheli et al., 1998). Age is important because the older the person is, the longer they have been exposed to carcinogens. Additionally, our body accumulates mutations the longer we live, and many mutations lead to cell loss of function. Therefore, older people have a higher chance of getting cancer. The grade of cancer addresses how similar the cancer cells are to healthy cells. Therefore, the lower is the grade, the more likely the positive outcome is. The staging corresponds to the severity of cancer or a specific tumor. For most cancers, the staging varies from 1 (localized and small) to 4 (cancer present in multiple locations in forms of metastasis). However, for some cancers, the stages are broken into substages to allow more accurate prediction. Tumor status is a binary parameter that is equal to 1 if the tumor is present and to 0 otherwise. Histological type addresses the type of tissue affected by cancer and to the tumor tissue type. The year of the diagnosis is important since the cancer treatments rapidly evolve. The year when a person gets diagnosed with cancer might affect the outcome significantly. The new tumor and days to new tumor address the recurrence or metastasis of cancer and how long it took to occur.

4 Methods

For each cancer, we predicted the vital status and the outcome. We trained several models for both types of predictions and compared the results. Each model was trained twice: first, with the assumption that the stages and grades of cancer are consecutively connected, and then, with the assumption that they are independent. In the latter version, we used a one-hot encoding to represent the specific stage of cancer or tumor. In this paper, we call the two versions of input 'unvectorized'

Table 2: Descriptions of the studied cancer types

Type	BLCA	HNSC	LGG	LUAD	LUSC	OV	STAD	UCEC
Samples	352	347	430	390	307	450	349	473
Feature 1	Age	Age	Age	Age	Age	Age	Age	Age
Feature 2	Gender	Gender	Gender	Gender	Gender	Clinical stage	Gender	Hist. grade
Feature 3	Tumor stage	Tumor stage	Hist. grade	Tumor stage	Tumor stage	Hist. grade	Tumor stage	Hist. type
Feature 4	Diagn. year	Hist. grade	Hist. type	Diagn. year	Diagn. year	Diagn. year	Hist. grade	Clinical stage
Feature 5	Tumor status	Clinical stage	Diagn. year	Tumor status	Tumor status	Tumor status	Hist. type	Diagn. year
Feature 6	New tumor	Diagn. year	Tumor status	New tumor	New tumor	New tumor	Diagn. year	Tumor status
Feature 7	Days to new tumor	Tumor status	New tumor	Days to new tumor	Days to new tumor	Days to new tumor	Tumor status	New tumor
Feature 8		New tumor					New tumor	Days to new tumor
Feature 9		Days to new tumor					Days to new tumor	Menopause status

and 'vectorized' respectively. We also attempted to train several models to predict the number of days to death but did not derive any significant results due to the large variance in the data.

We only used supervised learning models since we had clearly defined classes and wanted to establish patterns in the data rather than find the patterns to form new classes. We trained the models on randomly chosen data points and used 5-fold cross-validation to decrease the likelihood of unfortunate random assignment. The distribution of classes in every batch was equal to increase the test accuracy.

4.1 Vital Status

Vital status is a binary value $y \in \{0, 1\}$ that represents whether the person is dead or alive in the 10 year span after the primary diagnosis.

4.1.1 Binary Logistic Regression

We started with the implementation of binary logistic regression that we studied in class. This model was reasonable to use since it has binary outputs $y \in \{0, 1\}$, which directly correlates with the dead/alive predictions that we wanted to make.

The hypothesis for binary logistic regression:

$$h_\theta(x) = g(\theta^T x), \qquad g : \mathbb{R} \to [0, 1],$$

where $x \in \mathbb{R}^{m \times n}$ is an input vector, $\theta \in \mathbb{R}^n$ is a vector of weights for the features of x, and g is a logistic (sigmoid) function that maps real values to probabilities between 0 and 1:

$$g(z) = \frac{1}{1 + e^{-z}}$$

With the logistic probability below 0.5, we assumed that the given sample belongs to class 0. Likewise, with the probability above or equal to 0.5, we assumed that the sample belongs to class 1. In this way, we applied the following rule to the outputs of g to get the vector $y \in \mathbb{R}^m$ of binary predictions:

$$y = \begin{cases} 0, & \text{if } h_\theta(x) < 0.5 \\ 1, & \text{if } h_\theta(x) \geq 0.5 \end{cases}$$

In this implementation, we used the negative log-likelihood (NLL) cost function. It scales to small changes in θ better than the squared error cost function and improves the convergence of the model:

$$J(\theta) = \sum_{i=1}^{n} -y^{(i)} \log h_\theta(x^{(i)}) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

To minimize the cost, we used the gradient descent algorithm. It iteratively plugs the weights $\theta_1 \dots \theta_n$ into the cost function $J(\theta)$. For each weight θ_i, it calculates a partial derivative to find the direction of its fastest decrease. Then, it takes a step in this direction and updates θ_i by the following rule:

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta), \qquad j = 0, \dots, m$$

4.1.2 Scikit-learn Logistic Regression

The next model that we trained was the scikit-learn implementation of logistic regression. We tried out every solver provided by the scikit-learn logistic regression package to maximize the accuracy of the model. Eventually, we settled with the 'liblinear' solver since it had the best average performance.

The scikit-learn implementation uses the L2-regularized log-likelihood (LLH) cost function. It is the opposite of NLL, which needs to be maximized rather than minimized, and it takes the form:

$$J(\theta) = \min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1),$$

where C is the inverse of the regularization constant (Scikit-learn, Linear Models).

Besides, the 'liblinear' solver uses a preconditioned conjugate gradient method to train a logistic regression model (Fan, Chang, et al., 2008). It employs the trust-region update rule that utilizes the Newton optimization method to minimize the trust region and converge the model (Hsia, Zhu, et al., 2017):

$$\Delta_{k+1} = \begin{cases} \min((\max(\alpha_k^*, \gamma_1))||\boldsymbol{s}^k||, \gamma_2 \Delta_k), & \text{if } \rho < \eta_0, \\ \max(\gamma_1 \Delta_k, \min(\alpha_k^* ||\boldsymbol{s}^k||, \gamma_2 \Delta_k)), & \text{if } \rho \in [\eta_0, \eta_1], \\ \max(\gamma_1 \Delta_k, \min(\alpha_k^* ||\boldsymbol{s}^k||, \gamma_3 \Delta_k)), & \text{if } \rho \in (\eta_1, \eta_2), \\ \max(\Delta_k, \min(\alpha_k^* ||\boldsymbol{s}^k||, \gamma_3 \Delta_k)), & \text{if } \rho \geq \eta_2, \end{cases}$$

where Δ_k is the trust region size at the kth iteration, $\boldsymbol{s}^k$ is the Newton direction within the trust region, α_k is the step size, ρ is the predicted cost reduction, and $\eta_0, \eta_1, \eta_2, \gamma_1, \gamma_2, \gamma_3$ are positive pre-defined constants.

4.1.3 Naive Bayes Classification

Lastly, we used the Gaussian Naive Bayes classifier as implemented in the scikit-learn package. It operates under the assumption that the predictor values are sampled from a Gaussian distribution. The hypothesis for this classifier uses the Bayes' theorem:

$$\hat{y} = \arg\max_y P(y|x) = \arg\max_y \prod_{i=1}^{n} P(x_i|y)$$

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\Big(- \frac{(x_i - \mu_y)^2}{2\sigma_y^2} \Big),$$

where the parameters σ_y and μ_y are estimated using maximum likelihood function (Scikit-learn, Naive Bayes).

This classifier is considered to be ideal for the small datasets like ours because it can approach the asymptotic error faster than a logistic regression classifier. However, it also tends to work poorly for intercorrelated input features due to the strong (naïve) feature independence assumptions.

4.2 Outcome

Outcome is a discrete multiclass variable $y \in \{0, 1, 2, 3, 4\}$ with the possible interpretations of complete remission, partial remission, stable disease, persistent disease, and progressive disease. For every cancer type, some outcomes were more prevalent than the others, and some classes had less than 10 samples. In those cases, we excluded the minor classes and ran the models on the 2-4 more prevalent ones to ensure the balanced distribution of the training data.

For every cancer, we trained the in-class implementation of the one-vs-all classifier with L2-regularization. For each class $c = 1, \ldots, K$, this classifier fits a logistic regression model to distinguish c from the other classes. The hypothesis is as follows:

$$h_w(x) = \text{logistic}(b + w_1 x_1 + \cdots + w_n x_n) = \text{logistic}(w^T x + b),$$

where $x \in \mathbb{R}^n$ is the feature vector, $w \in \mathbb{R}^n$ is a weight vector, and b is a scalar intercept parameter. Then, the predicted class is the value of c that maximizes the output of $h_w(x)$.

The second model that we trained for every cancer was the scikit-learn logistic regression classifier with L2-regularization. Just like the in-class implementation, this library uses classifier chains, combining K binary classifiers into a single multi-label model. For predicting the disease outcomes, the best-performing solver was 'auto' .

Additionally, we trained the in-class implementation of the binary logistic regression classifier for those cancers that had only 2 prevalent outcomes.

As with the vital status predictions, we trained a Gaussian Naive Bayes model for every cancer type, which also works for the multi-label classification.

5 Results and Evaluation

We used accuracy as the main evaluation metrics for all of our models. The output of our models was a class for a specific point. Therefore, we used the accuracy of random assignment as our baseline. The baseline accuracy b can be calculated by the following formula:

$$b = \frac{100\%}{c}, \qquad \text{where } c \text{ is the number of classes.}$$

5.1 Results for vital status prediction

As follows from the formula above, our baseline accuracy for the vital status was 50%. The average training accuracy for the best-performing classifier was 78.4%, which is significantly higher.

Figure 1 makes it evident that the scikit logistic regression model significantly outperformed the in-class model and also showed better results than Naive Bayes. The difference in performance can be explained by the trust-region update rule that the scikit package uses. It is based on fitting a paraboloid to the surface of the function being minimized (A. W. F. Edwards, 1992). This operation involves taking a second-order derivative as opposed to the first-order derivative used for regular gradient descent. It gives us better understanding of the function's behavior, which makes this approach more precise than the one that we implemented in class. The most likely reason for the poor performance of the Naive Bayes model is the invalid feature independence assumption.

There was no significant difference in average performance on vectorized and unvectorized input for the model with the highest degree of accuracy. However, for some types of cancers, there was a slight difference in performance between the two versions, which differed from type to type. It suggested that for some cancers, the assumption that the stages are linearly dependent was valid and useful for making predictions, while for others, there was no direct relationship between different stages and the vital status.

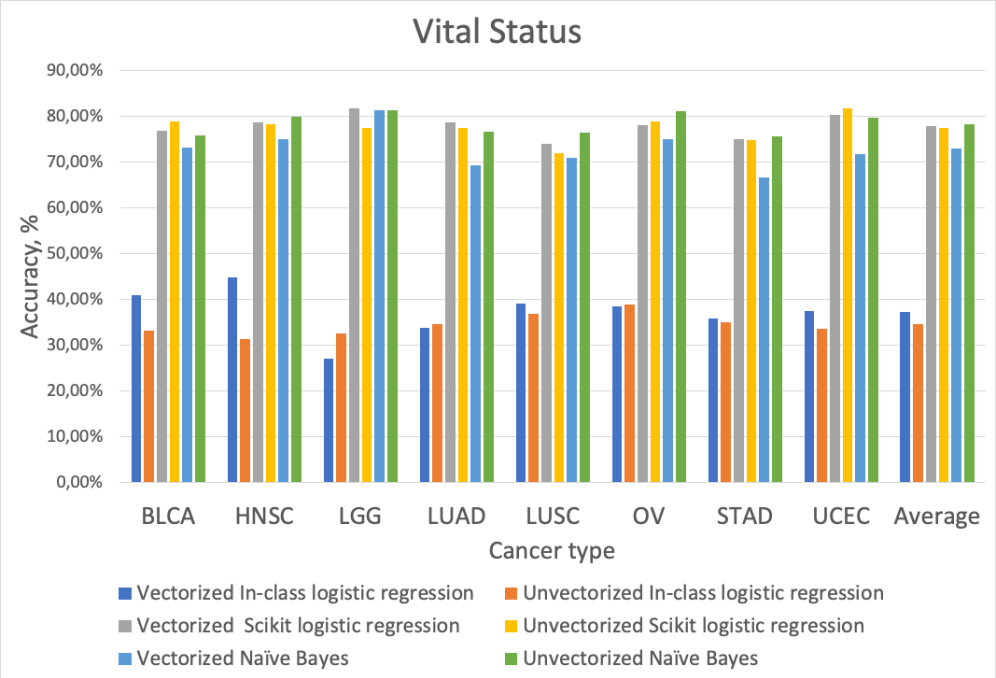

Figure 1: The accuracy of vital status prediction on vectorized and unvectorized data

5.2 Results for outcome prediction

The average baseline accuracy for the outcome predictions was 37%. The individual baseline accuracy varied from cancer to cancer and is displayed in Figure 2 for better clarity. On average, the in-class multiclass logistic regression performed as well as the scikit multiclass logistic regression. These two models significantly outperformed both Naive Bayesian model and the baseline with accuracy varying from 60% to 96%, and 76% accuracy on average.

LGG has shown the lowest accuracy, closely followed by OV. The results for LGG confirmed that, while the same features could help accurately predict the outcome for various cancers, it was not true for LGG due to the underlying differences in the course of the disease.

OV cancer has also shown low accuracy, suggesting that the engineered features were less relevant for this type of cancer than for others. We believe that it could happen due to different causes and mechanisms of OV cancer, such as hormones, childbirth, etc.

5.3 Most important features

We established the most important features for each of the cancers by comparing the weights associated with them in the model. The presence of the tumor was the most significant feature across all datasets. It implies that the stage of the tumor and the stage of cancer affects the outcome far less than the mere presence of the tumor and that in combination with other factors, the presence of the tumor is a significant outcome predictor.

The second most important feature for all cancers was the occurrence of a new tumor. However, for LGG, the only disease that is not medically classified as cancer, the new tumor has not been a significant factor. Instead, gender appeared to be a better predictor of LGG outcomes.

In terms of the third most significant feature, it varied based on the type of cancer. Generally, gender was the third most important feature. However, for both OV and UCEC, gender was not applicable since only people with uterus or ovaries get those types of cancer. So for OV, UCEC, and LGG, age appeared to be the third most important feature.

The fact that gender appeared to be one of the most critical features leads to several conclusions. First of all, the differences in biochemistry depended on gender significantly affect the outcome,

and further research is required to establish the mechanism behind this since it can be attributed to hormones and their role in the immune systems, sex-linked genetics, or differences in the lifestyle.

The role of age in determining the outcome is also ambiguous since, in this case, age is a predictor of outcome and not incidence. Potentially, the age affects the immune system, the work of the cells, the biochemistry, or the ability to undergo treatment. A better understanding of the reasons behind this will lead to a better choice of treatment or potential supplements that can affect the immune system, biochemistry, or work of the cells.

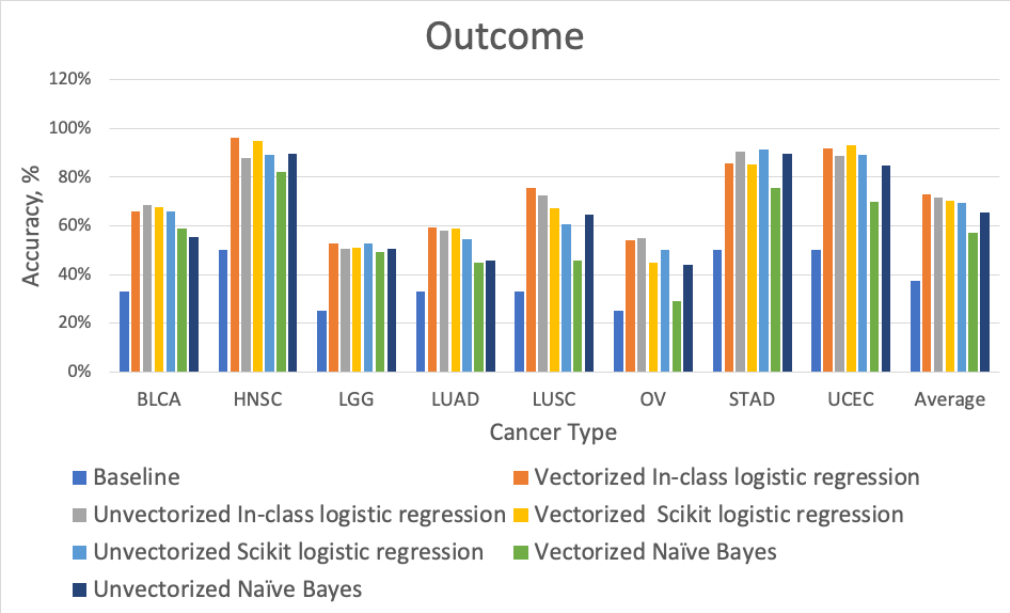

Figure 2: The accuracy of outcome prediction on vectorized and unvectorized data

6 Conclusion

We believe that this project was an important proof of concept. It showed that consideration of the outside factors is essential to predict a patient's outcome accurately. Even on the small amount of data, our models outperformed human physicians by 16% (Clément-Duchêne et al., 2010). It demonstrated the urgent need to collect and study the cancer data to understand cancers' mechanics better and improve the existing treatments and decision-making concerning them. The accurate prediction of the outcome based on the initial factors can also help with choosing and adjusting treatments for a specific group of patients.

6.1 Limitations

In this study, we had to work with only the features provided by The Cancer Genome Atlas Program. The limited amount of data has prevented us from using neural networks or have a higher level of accuracy for the existing models.

However, having access to more relevant medical data poses several questions in regards to ethics. The security system in place right now ensures that the medical data can not be stolen and can only be accessed by the authorized personnel of the hospital or medical network. The absence of a proven secure digital solution explains why the digitization of the records is happening so slowly. Additionally, even with new technology based on blockchain security, the digitization of old records is not a primary goal of the medical community. That said, the thorough discussion of potential ethical implications and possible solutions are required to proceed with this emerging direction of research.

6.2 Future work

In the future, we would like to repeat this study using more relevant features. It includes but not limited to such factors as the number of kids, the blood type, the average resting heart rate, how many nicotine years does the person have, their weight, their height, their weekly number of active minutes, comorbidity, what kind of health care they have, their race and genetics, the histology of the tumor, and the protocol of their treatment. It is not an extensive list of factors that might affect the cancer outcome, but the mere minimum that will allow scientists to understand the outside factors that affect the outcomes.

Additionally, having more data will allow us to create other models such as neural networks that are known to have higher accuracy than models that we worked with.

Another potential direction of future research is to use the approach described in Johnson, (2000) for determining the feature importance. We could use it to verify the established important features for different cancers, as well as evaluate their role in predicting the disease outcomes in a more detailed way and understand they are integrated into the mechanism of cancer itself.

References

[1] Ho, J.Y. & and Hendi, A.S. (2018) Recent trends in life expectancy across high income countries: Retrospective observational study. *BMJ* **362**(8).

[2] Pompei, F. (2001) Age distribution of cancer: The Incidence turnover at old age. *Hum. Ecol. Risk Assess* **7**(6), 1619–1650.

[3] Hastert, T.A., Beresford, S.A.A., Sheppard, L. & White, E. (2015) Disparities in cancer incidence and mortality by area-level socioeconomic status: A multilevel analysis. J. Epidemiol. *Community Health* **69**(2), 168–176.

[4] Wang, Y.C., Wei, L.J., Liu, J.T., Li, S.X., & Wang, Q.S. (2012) Comparison of cancer incidence between China and the USA. *Cancer Biol. Med.* **9**(2), 128–132.

[5] Michiels, S., Koscielny, S. & Hill, C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet* **365**(9458), 488–492

[6] Liu, J., Lichtenberg. T. & Hu, H. (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**(2), 400–416.

[7] Shi, Q. & Sargent, D.J. (2009) Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *International Journal of Clinical Oncology* **14**, 102–111.

[8] Shen, P.S. (2000) Testing for sufficient follow-up in survival data. *Stat. Probab. Lett.* **49**(4), 313–322.

[9] Catto, J.W.F., Linkens, D.A., Abbod, M.F., Chen, M., Burton, J.L., Feeley, K.M. & Hamdy, F.C. (2003) Artificial intelligence in predicting bladder cancer outcome: A comparison of neuro-fuzzy modeling and artificial neural networks. *Clin. Cancer Res.* **9**(11), 4172–4177.

[10] Warburg, O. (1956) On the Origin of Cancer Cells. *Science* **123**(3191), 309-314.

[11] Dinney, C.P., McConkey, D.J., Millikan, R.E., Wu, X., Bar-Eli, M., Adam, L. & Grossman, H.B. (2004) Focus on bladder cancer. *Cancer cell* **6**(2), 111-116.

[12] Sarini, J., Fournier, C., Lefebvre, J.L., Bonafos, G., Van, J.T. & Coche-Dequéant, B. (2001). Head and neck squamous cell carcinoma in elderly patients: a long-term retrospective review of 273 cases. *Archives of Otolaryngology–Head & Neck Surgery*, **127**(9), 1089-1092.

[13] Ostrom Q.T., Gittleman H., Xu J., et al. (2016) CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2009-2013. *Neuro. Oncol.* **18**(Suppl 5), v1–v75.

[14] Dougherty, M.J., Santi, M., Brose, M.S., Ma, C., Resnick, A.C., Sievert, A.J., Storm, P.B. & Biegel, J.A. (2010) Activating mutations in BRAF characterize a spectrum of pediatric low-grade gliomas. *Neuro. Oncol.* **12**(7), 621–30.

[15] Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E. & Lizyness, M.L. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine* **8**(8), 816-824.

[16] Masuda, R., Kijima, H., Imamura, N., Aruga, N., Nakamura, Y., Masuda, D. & Inokuchi, S. (2012) Tumor budding is a significant indicator of a poor prognosis in lung squamous cell carcinoma patients. *Molecular medicine reports* **6**(5), 937-943.

[17] Cannistra S.A., Gershenson D.M., Recht A. (2015) Ovarian cancer, fallopian tube carcinoma, and peritoneal carcinoma. *In Cancer: Principles and Practice of Oncology*, ch.76. 10th ed. Philadelphia, PA: Lippincott Williams & Wilkins.

[18] Sehdev A., Catenacci D.V. (2013) Gastroesophageal cancer: focus on epidemiology, classification, and staging. *Discov Med.* **16**:103-11.

[19] Kurman, R.J., Zaino, R.J. & Norris, H.J. (1994) Endometrial carcinoma. *In Blaustein's pathology of the female genital tract*, (pp. 439-486). Springer, New York, NY.

[20] Chen, Y.Y., Hollander, M. & Langberg, N.A. (1982) Small-sample results for the Kaplan-Meier estimator. *Journal of the American statistical Association*, **77**(377), 141-144.

[21] Cox, D.R. (1972) Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society*, Series **B**(34),187—220.

[22] Alaa, A.M. & Van Der Schaar, M. (2017) Deep multi-task gaussian processes for survival analysis with competing risks. *In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*.

[23] Lee, C., Zame, W.R., Yoon, J. & Van Der Schaar, M. (2018) DeepHit: A deep learning approach to survival analysis with competing risks. *In 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2314–2321.

[24] Micheli, A., Mariotto, A., Rossi, A. G., Gatta, G., Muti, P., & EUROCARE Working Group. (1998) The prognostic role of gender in survival of adult cancer patients. *European Journal of Cancer*, **34**(14), 2271-2278.

[25] Johnson, J.W. (2000) A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate behavioral research*, **35**(1), 1-19.

[26] Hsia, C.Y., Zhu, Y. & Lin, C.J. (2017) A study on trust region update rules in Newton methods for large-scale linear classification. *In Asian conference on machine learning*, (pp. 33-48).

[27] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R. & Lin, C.J. (2008) LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, **9**(8), 1871-1874.

[28] Scikit-Learn. (2019) 1.1.11. Linear Models — scikit-learn 0.22.2 documentation. Retrieved April 26, 2020 from https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

[29] Scikit-Learn. (2019) 1.9. Naive Bayes — scikit-learn 0.22.2 documentation. Retrieved April 26, 2020 from https://scikit-learn.org/stable/modules/naive_bayes.html.

[30] Clément-Duchêne, C., Carnin, C., Guillemin, F. & Martinet, Y. (2010) How accurate are physicians in the prediction of patient survival in advanced lung cancer?. *The oncologist*, **15**(7), 782.

[31] Edwards, A.W.F. (1992) *Likelihood*. Johns Hopkins University Press, p. 129.