Predicting Disease Incidence Based on Demographic and Lifestyle Factors

Bianca Schutz
Rice University
mbs5@rice.edu

Sachin Shurpalekar
Rice University
sss20@rice.edu

Sienna Tu
Rice University
sat6@rice.edu

Abstract

Health outcomes are influenced by multiple lifestyle and demographic factors. However, some factors are more important than others in predicting various types of diseases. Logistic regression models were created to analyze the importance of these features based on the 2015 Behavioral Risk Factor Surveillance System (BRFSS) Survey for four diseases: heart disease, asthma, cancer, and diabetes. The results indicated that the logistic regression models to predict heart disease and diabetes performed more accurately than the machines for asthma and cancer. This study also showed that only some of the previously identified risk factors are heavily correlated with predicting the disease a person has, while other features were unexpectedly less correlated. However, difficulties predicting the incidence of disease suggest that there are factors not in the data analyzed in this study, also impacting whether someone develops one of these four diseases.

1 Introduction

There are clear patterns and influences that can lead to a person's likelihood of being diagnosed by a disease. Heart disease is the leading cause of death in the United States of America (U.S.) as 1 in 5 people died of heart disease in 2022 [5]. Some risk factors include high blood pressure, diabetes, obesity, physical inactivity, and excessive alcohol use. 1 in 12 people in the U.S. have asthma and some risk factors include family history, allergies, obesity, and air pollution [6]. 1 in 5 people develop cancer worldwide and some risk factors include age, diet, sun exposure, smoking, alcohol, and obesity [10]. Finally, for diabetes, 11.6% of the U.S. population had diabetes in 2021. Risk factors for diabetes include genetics, weight, physical activity, cholesterol, etc.

The goal of this project is to predict the disease a patient may be diagnosed with depending on lifestyle factors such as diet, activity level, drinking habits, etc. We hypothesize that the features most closely aligned with the risk factors for each of the four diseases analyzed will most likely be the parameters most important in making accurate predictions for the disease. For instance, since physical inactivity is a risk factor for heart disease, we hypothesize that physical activity levels will have a negative β.

2 Related Work

Logistic regression has been widely used in clinical contexts to predict lung cancer, heart disease, and other diseases [3][7]. Deppen et al. developed the TREAT model to predict lung cancer, based on demographic and behavioral features such as BMI, age, sex, and tobacco use. However, the model also included medical results data, which largely contributed to its success [3]. Latifah, Slamet, and Sugiyanto found that while both random forests and logistic regression performed similarly with heart disease predictions, logistic regression was slightly more accurate than random forests. In contrast to the findings of Latifah, Slamet, and Sugiyanto, in the context of predicting diabetes, Daghistani and Alshammari found that random forests yielded better results [2].

Other supervised learning machines like Random Forests and Support Vector Machines can also use historical data to make predictive models (Pattyam, 2019). Another study showed that Random Forests had higher accuracy in disease prediction than Support Vector Machines (SVM), but SVM algorithms were applied more frequently.

This literature suggests that both logistic regression and random forests have been useful machines in predicting incidence of disease. However, the matter of which performs better is dependent on the data. As a result, we decided to begin with logistic regression as it is less computationally expensive than random forests and is easily interpretable, which is important for the application of this work to a medical context where doctors would need to quickly understand the results.

3 Methods

3.1 Experimental Design

Past research has indicated that certain factors are more predictive of certain diseases than others. This study aims to identify the features important in predicting four diseases: heart disease, asthma, cancer, and diabetes. We are particularly interested in examining the impact of exercise frequency and intensity, diet, alcohol consumption, tobacco use, and demographic characteristics such as BMI, age, and sex.

3.2 Data

To examine the effects of lifestyle and demographic variables on these four diseases, we used the 2015 Behavioral Risk Factor Surveillance System (BRFSS) Survey available on Kaggle. The BRFSS is the largest continuous health survey, and it is conducted by the Center for Disease Control (CDC). The data was collected through landline calls and cell phone calls. The participants of this survey span across all 50 states in the U.S. and includes the District of Columbia, Guam, and Puerto Rico. This dataset includes over 441,000 responses and over 300 factors covering health behaviors, lifestyle, and demographics of participants.

3.3 Algorithms

This study was completed by first cleaning the data. We chose to narrow our research on certain features and so columns that were not evaluated were removed and narrowed to 25 variables. Rows with missing data (either NA or "don't know" or "refused")

were removed. Furthermore, some questions in the survey elicited a multiclass response, which were then narrowed down to binary variables. We filtered our data so that our response variables (the four diseases) were binary as well, with 1 = disease and 0 = no disease.

Additionally, we had some features in the questionnaire that were asked following other questions, so if someone responded "No" to the first question, the others were coded as NA. For example, if the person responded that they did not exercise, all other exercise-realted questions became NA, resulting in over 100,000 data points missing. To address this issue without losing that data, we recoded NAs to be zeros in this case, since the question was not asked because the person does not exercise. This ensured that we retained valuable information without losing over 100,000 data points that could cause a disproportionate analysis for diseases with low prevalence.

Four logistic regression machines were created—one for each disease analyzed. The data was split 50/50 into training and test data to ensure the model was generalizable to a large amount of unseen data. For each disease, the data had extremely unbalanced classes, with about 90% of the data not having the disease. Therefore, Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic examples for the minority class, was used to balance the classes. We also standardized the data using StandardScaler(). LogisticRegression() was used with GridSearchCV() for tuning hyperparameters for each disease. A grid of 1,600 combinations of the parameters with a 3-fold cross-validation (CV; 4,800 total iterations) was completed. The hyperparameter grid was intentionally comprehensive to ensure that no significant parameter combinations were overlooked, and the model tuning process was exhaustive.

As a follow-up analysis on logistic regression, we also tested the effect of SMOTE on our results by comparing grid search models with and without it. Because the classes were balanced in the original four models, the number of false positives was incredibly high in many of the models, so we aimed to correct this mistake and improve performance by testing whether SMOTE actually helped. This was crucial for understanding whether synthetic oversampling provided actual benegits or just introduced noise into the dataset, which is a common concern that comes with the use of SMOTE. We also used different class weights based on the ratio of the majority and minority classes to maximize the performance of the minority class predictions. These models tracked F1 scores, precision, and recall and used F1 as its scoring metric.

Our reasoning for tackling SMOTE centers around the findings of Elor and Averbuch-Elor, who find that while SMOTE is useful for weak classifiers (such as logistic regression), hyperparameter tuning can also have a significant impact on prediction quality and on the effectiveness of balancing [4]. Another author also notes: "why would you train a machine learning model on a 50% / 50% balance if the class balance in future data is expected to be 95.5% / 0.05%?" [8].

Given the limitations observed in logistic regression's precision and recall for the minority class, we expanded our approach to include different models. Since a literature review identified Random Forests as being more accurate in predicting diseases in some contexts, we also tested this machine. We first compared a basic random forest model to logistic regression models and found that while it had overall higher accuracy (accuracy = 0.91; Table 4), its ability to predict the minority class (precision = 0.16) was worse than the logistic regression model (precision = 0.75; Table 4, Table 3). We performed a grid search of different hyperparameter combinations in hopes of improving its performance, but did not see a significant breakthrough.

4 Results

Table 1 outlines the results from the logistic regression machines. Asthma had the lowest AUC = 0.58 and heart disease and diabetes had the highest AUC = 0.76.

Table 1: Logistic regression results: Area-under-curve and best scores for each disease.

Condition	AUC	Best Score
Heart Disease	0.76	0.782
Asthma	0.58	0.682
Cancer	0.67	0.692
Diabetes	0.76	0.765

Table 2 highlights the features that had the highest importance when predicting the diseases for the test dataset. Feature 1 had the highest β_1 while Feature 2 had the second highest β_2. A positive β indicates a positive correlation. This means that if β_1 is positive, as Feature 1 increases, the likelihood of the condition increases. The opposite is true for a negative β.

Table 2: Logistic regression results: top two most important features

Condition	Feature 1	β_1	Feature 2	β_2
Heart Disease	Age	0.942	isFemale	-0.837
Asthma	ActivityIntensity	-0.570	ActiveMin	0.559
Cancer	Age	0.777	ActiveMin	0.585
Diabetes	BMI	0.717	Age	0.682

It is also clear from the Classification Reports that the logistic regression machines were not very accurate in predicting diseases. Table 3 shows that while the heart disease machine had a precision of 0.97 in identifying people who did not have the disease, it had a precision of 0.14 in identifying people who did have the disease. The model for heart disease performed the best overall, while the diabetes model performed the best on the minority class. Additionally, upon further analysis, a 5-fold CV was performed with additional parameter combinations which did not show an increase in accuracy.

As noted in the previous section, we also employed a random forests model after seeing the poor results for the logistic regression models. To determine if this method would significantly improve our results, we tested it on heart disease, which performed the best out of the logistic regression models. We ran a grid search of 4,800 combinations of varying amounts of $n\_estimators$ or the number of trees, different $max\_features$, $criterions$, and other parameters. The best results from this came from a tree with a greater

Table 3: Logistic regression results: precision, recall, and F-1 scores for Heart Disease

	Precision	Recall	F1-Score	Support
Does not have disease	0.97	0.75	0.85	160604
Has disease	0.14	0.61	0.22	10190

Note. The accuracy was 0.75 based on 170794 instances.

max_depth parameter and more trees, demonstrating that a more complex model was the best for the training data. However, this model did not perform well in classifying the minority class (Table 4). While the logistic regression model over-predicted the incidence of disease, the random forests under-predicted disease, which could be more harmful than false positives.

Table 4: Random forest results: precision, recall, and F-1 scores for Heart Disease

	Precision	Recall	F1-Score	Support
Does not have disease	0.95	0.96	0.95	160604
Has disease	0.16	0.13	0.14	10190

Note. The accuracy was 0.91 based on 170794 instances.

To further investigate the results of the logistic regression models, we performed a comparison between SMOTE and no SMOTE with various class weights. We performed this analysis on diabetes, which performed the best on the minority class. We find that for diabetes, while SMOTE and no SMOTE perform similarly, no SMOTE performs better on the minority class and is a huge improvement from our initial grid search cross-validation in terms of recall (Table 5). The best class weight for no SMOTE was $0 : 0.467, 1 : 0.533$. Still, the precision and F-1 score are similar to the original logistic regression with SMOTE, demonstrating this isn't a perfect solution.

Table 5: SMOTE comparison results: precision, recall, and F-1 scores for Diabetes

With SMOTE	**Precision**	**Recall**	**F1-Score**
Does not have disease	0.93	0.71	0.81
Has disease	0.26	0.68	0.38
No SMOTE	**Precision**	**Recall**	**F1-Score**
Does not have disease	0.95	0.69	0.80
Has disease	0.27	0.74	0.39

5 Discussion

Some of the features found to be most important support the findings of existing medical knowledge. Sex was identified as the second most important feature for heart disease (β = -0.837), meaning that if an adult is female, the log odds that they have heart disease decrease. This aligns with existing research which finds that heart disease is more prevalent in men than in women [9].

Since asthma's AUC = 0.58, this showed that the model was just slightly better at predicting whether a patient had asthma or not, than flipping a coin (50% chance). Such low discriminatory power for asthma may suggest that the features in the model don't adequately capture the unique predictors of this condition. The other diseases had a higher AUC value suggesting the machines were more accurate. The models performed well predicting when people did not have the diseases, but was not precise in its predictions of those who actually have the disease. This can be attributed to one major limitation of this study–the scope of the dataset. This not only affects the accuracy of individual health reports, but may also introduce bias into the model, particularly for diseases that rely heavily on precise diagnostic data rather than subjective lifestyle reporting.

The dataset is a questionnaire, meaning the data is adults self-reporting these characteristics about their health. This can lead to inaccurate data as some adults may not be aware of any diagnoses they have unless they have been to a doctor. Even so, the adult may have a disease without being diagnosed properly (either a wrong diagnosis or missed diagnosis).

The results of this study were also interesting as certain risk factors had a different correlation to the disease than other correlations. For instance, for heart disease, *TotalActiveMinsPerWeek* was positively correlated with heart disease whereas *ActivityIntensity* was negatively correlated. It makes sense that the activity intensity is negatively correlated as a risk factor for heart disease is physical inactivity. However, the total active minutes per week was positively correlated with heart disease indicating the more minutes spent exercising, the higher the likelihood of having heart disease, which does not make sense according to our literature reviews. Features like these should be further analyzed to then better the machine.

Age was a feature that had a heavy influence in predicting heart disease (β = 0.942), cancer (β = 0.777), and diabetes (β = 0.682) prediction (Table 2). Furthermore, since the correlation was positive in all three cases, this means that as a person ages, their likelihood of getting the disease is higher.

It is also interesting that the importance of the feature most correlated to asthma, *ActivityIntensity*, was only β = -0.570 (Table2). This is very low considering the importance of *Age* for heart disease was β = 0.942.

Specific diseases also have attributes that can lead to unexpected features having a greater/less influence. For instance, diabetes is a risk factor for heart disease, but two separate logistic regression models were created. Cancer is a broad disease with different risk factors for different types of cancers. However, the survey was general about whether a person had cancer or not and did not have the opportunity to specify the type of cancer.

Finally, genetics and allergies are large contributors to the development of asthma [1]. Therefore, the features analyzed from this study would not be as correlated since these two features were not incorporated in this study. Furthermore, asthma typically develops in children and triggers can change over time and be more controlled which can alter whether people know if they have asthma or not [6].

The random forest application had better results in terms of accuracy. However, it is important to note that the random forests model under-predicted having the disease. In terms of application, this can have detrimental effects as it is typically safer to think you do have a disease than miss identifying the disease. False alarms in

a healthcare setting are typically safer than misses in the long run as it is safer to be cautious.

6 Conclusion

Logistic regression machines were created to predict whether or not a person has one of the four diseases analyzed based on correlated features in the training dataset.

6.1 Possible Applications

While the machines were not very accurate, this study emphasizes the difficulty and improvement needed in this field to help improve the accuracy of diagnoses. This shows that there is still potential for predictive models to complement traditional diagnostic approaches by acting as early warning systems in resource-limited healthcare settings. Identifying a disease as early as possible is beneficial for the patient as they can start treatment or change daily habits sooner rather than later, decreasing the potential for the disease to have harsher impacts. Other ways these machines can be useful in the medical field is not only identifying the disease a person may have, but analyzing the level of severity and needs the person may need to personalize their care. Such personalized care models could help optimize treatment plans, allocate medical resources more effectively, and improve patient outcomes by tailoring to their individual needs. Furthermore, these tools can be applied to imaging and scans to identify key patterns or anomalies that may lead to the diagnosis of a certain disease. For instance, using machine learning to identify tumors in MRI scans. In addition to imaging, these tools can be extended to analyze genetic data and lifestyle metrics to provide a holistic view of patient health.

However, something else to consider is the importance of patient privacy with the use of Electronic Health Records as they are susceptible to hacking [11]. There are laws such as the Health Insurance Portability and Accountability Act (HIPPA) that protect patient information. If machines like these will be implemented in hospitals and the healthcare sector, new policies and cybersecurity is a sector that should be explored as well. It would require robust encryption methods and secure data storage to balance the potential benefits of predictive tools with the risk of data misuse and breaches.

6.2 Future Directions

While behaviors can impact the incidence of these chronic conditions, this dataset is limited in that it does not include genetic predispositions, medical examinations and historical factors that could otherwise contribute to disease. Integrating data from medical imaging, lab results, and more could provide a more comprehensive understanding of disease risk factors and progression. This project has shown that it is not feasible to diagnose based solely on behavioral/demographic factors. As noted previously, family history holds many risk factors for these diseases (i.e. genetics, family history of asthma, etc.). Incorporating genetic screening data and polygenic risk scores could allow for a more precise analysis of these hereditary influences on disease incidence.

Furthermore, the correlation between diseases can also be analyzed to see if a specific feature impacting one disease increases the likelihood of another disease. Other genetic features should also be analyzed as genetics and family history are large risk factors that people do not have control over. Finally, using patient data would make this study more rigorous as the data will most likely be more accurate compared to a self-answered survey. Future studies could also leverage other machine learning methods to integrate more diverse data sources and improve accuracy with non-linear relationships.

Contributions

Thank you to Dr. Orchard for his support throughout our project and the Data Science Department at Rice University for the tools used to complete this study.

Bianca: Cleaned the dataset and built the machines, ran logistic regression and random forest, analyzed the data for the presentation and this paper, wrote about the algorithms and drafted the manuscript.

Sachin: Ran logistic regression, analyzed the data for the presentation and this paper, identified key practical applications, communicated with Dr. Orchard in setting up meetings.

Sienna: Completed literature reviews and background research on machine learning in the healthcare sector, ran logistic regression, analyzed the data for the presentation and this paper, drafted the manuscript.

Our dataset was obtained via Kaggle, thanks to the Centers for Disease Control and Prevention.

References

[1] American Lung Association. n.d.. What causes asthma? https://www.lung.org/lung-health-diseases/lung-disease-lookup/asthma/learn-about-asthma/what-causes-asthma

[2] Tahani Daghistani and Riyad Alshammari. 2020. Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes. *Journal of Advances in Information Technology* 11, 2 (May 2020), 78–83. https://doi.org/10.12720/jait.11.2.78-83

[3] Stephen A. Deppen et al. 2014. Predicting Lung Cancer Prior to Surgical Resection in Patients with Lung Nodules. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 9, 10 (Oct 2014), 1477–1484. https://doi.org/10.1097/JTO.0000000000000287

[4] Yotam Elor and Hadar Averbuch-Elor. 2022. To SMOTE, or Not to SMOTE? *arXiv* (11 May 2022). https://doi.org/10.48550/arXiv.2201.08528 arXiv:2201.08528 arXiv.org.

[5] Centers for Disease Control and Prevention. n.d.. About heart disease. https://www.cdc.gov/heart-disease/about/index.html

[6] Mayo Foundation for Medical Education and Research. 2024. Asthma. https://www.mayoclinic.org/diseases-conditions/asthma/symptoms-causes/syc-20369653#:~:text=Asthma%20canâĂŹt%20be%20cured,adjust%20your%20treatment%20as%20needed April 6.

[7] Firda Anindita Latifah et al. 2020. Comparison of Heart Disease Classification with Logistic Regression Algorithm and Random Forest Algorithm. *AIP Conference Proceedings* 2296, 1 (Nov 2020), 020021. https://doi.org/10.1063/5.0030579

[8] Christoph Molnar. 2023. Don't 'Fix' Your Imbalanced Data. https://mindfulmodeler.substack.com/p/dont-fix-your-imbalanced-data Accessed: 17 Dec. 2024.

[9] Lori Mosca et al. 2011. Sex/Gender Differences in Cardiovascular Disease Prevention: What a Difference a Decade Makes. *Circulation* 124, 19 (Nov 2011), 2145–2154. https://doi.org/10.1161/CIRCULATIONAHA.110.968792

[10] World Health Organization. n.d.. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer#:~:text=Tobacco%20use%2C%20alcohol%20consumption%2C%20unhealthy,Reducing%20the%20burden

[11] S. P. Pattyam. n.d.. AI in Data Science for Healthcare: Advanced Techniques for Disease Prediction, treatment optimization, and patient management. https://dlabi.org/index.php/journal/article/view/124 Distributed Learning and Broad Applications in Scientific Research.

GitHub Repository

https://github.com/bshoots17/SSB_DSCI303

Appendix

Table A1: Logistic regression results for Asthma

	Precision	Recall	F1-Score	Support
Does not have disease	0.92	0.67	0.78	148308
Has disease	0.13	0.45	0.20	15552

Note. The accuracy was 0.65 based on 163860 instances.

Table A2: Logistic regression results for Cancer

	Precision	Recall	F1-Score	Support
Does not have disease	0.94	0.65	0.77	154617
Has disease	0.16	0.60	0.25	17140

Note. The accuracy was 0.64 based on 171757 instances.

Table A3: Logistic regression results for Diabetes

	Precision	Recall	F1-Score	Support
Does not have disease	0.93	0.75	0.83	145293
Has disease	0.27	0.62	0.38	22192

Note. The accuracy was 0.73 based on 167485 instances.

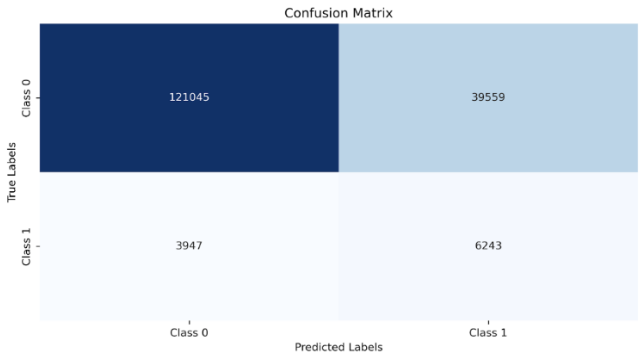

Figure A1: Confusion Matrix of Heart Disease

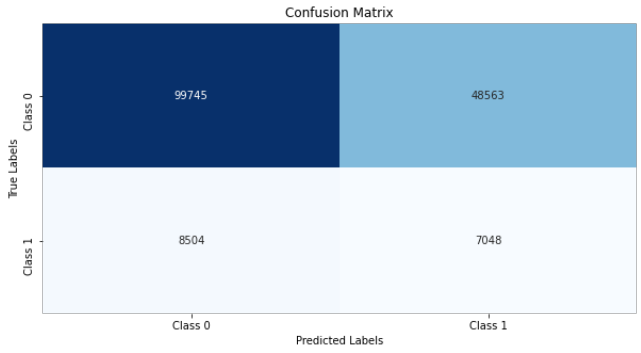

Figure A2: Confusion Matrix of Asthma

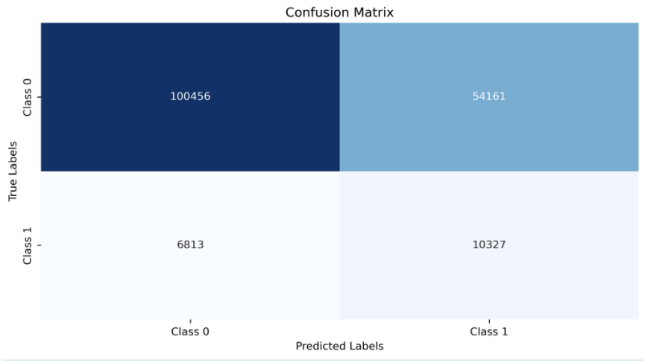

Figure A3: Confusion Matrix of Cancer

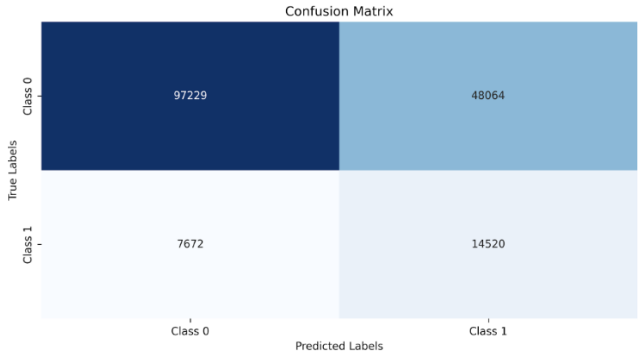

Figure A4: Confusion Matrix of Diabetes

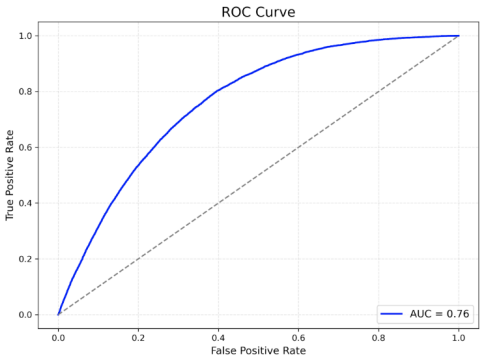

Figure A5: ROC Curve of Heart Disease

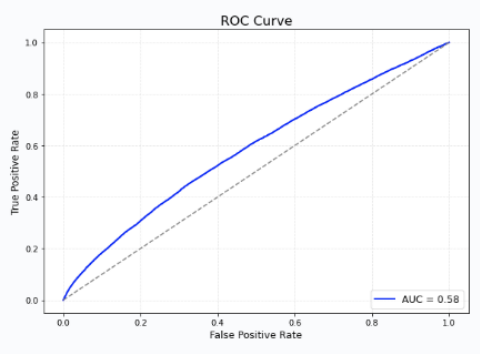

Figure A6: ROC Curve of Asthma

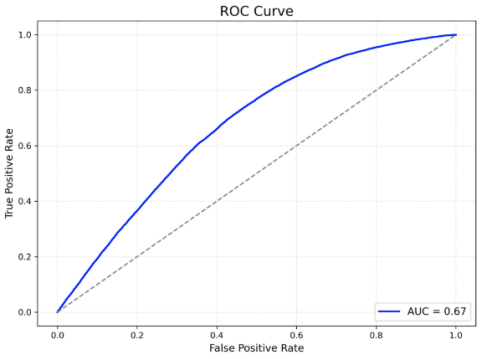

Figure A7: ROC Curve of Cancer

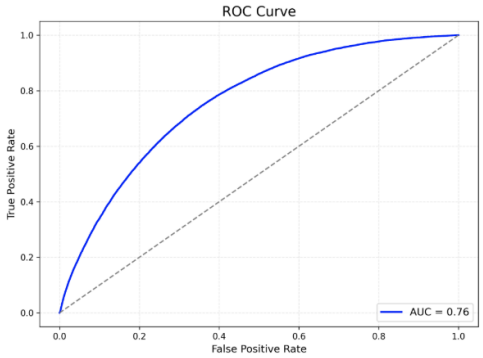

Figure A8: ROC Curve of Diabetes