Machine Learning Approaches to Breast Cancer Classification with *All of Us* Data

A Senior Honors Project Presented to the Faculty of the

Department of Information and Computers Sciences, University of Hawai‘i at Mānoa

In Partial Fulfillment of the Requirements

For the Bachelor of Science in Computer Science with Honors

By Jeremiah Keith Averia Dy

April 15, 2024

Thesis Committee:

Peter Yiğitcan Washington, PhD; mentor

Mahdi Belcaid, PhD

Co-Authors:

Zain Jabbar

Acknowledgments

I gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. I also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.

I would like to extend a special thanks to my primary research mentor, Dr. Peter Y. Washington in his continued support of all aspects of my undergraduate thesis journey. I also wish to thank my Honors committee member, Dr. Mahdi Belcaid, for his helpful insight and critique of my thesis findings and report.

I would like to acknowledge everyone who has taken the time to assist me in my educational endeavors and contributed to my academic growth and the development of this thesis. Without the support of my friends, family, and academic peers the completion of this thesis would not have been possible.

Lastly, any opinions, findings, and conclusions or recommendations expressed in this material are those of the primary author and do not necessarily reflect the views of the University of Hawai'i, the National Institute of Health, or the All of Us Research Program.

Abstract

Cancer diagnosis is a lengthy process and can contribute to the fatigue of medical staff. Additionally, breast cancer is one of the leading causes of cancer-related deaths amongst women. Therefore, it is imperative to research new, safe ways to expedite patient diagnosis to improve patient outcomes. This study aims to apply machine learning techniques to the field of medical science for the purpose of creating a proof-of-concept diagnostic tool which can help expedite breast cancer diagnosis. Machine learning models were imported from the scikit-learn library and trained on two different sets of cardiovascular health and quantitative liquid biopsy data from the *All of Us* database to predict breast cancer malignancy. The first dataset contained a higher volume of data with a small number of predictive features, while the second dataset contained a smaller volume of data with a relatively larger number of predictive features. Models were evaluated using a test dataset containing all features used in either dataset. All models performed poorly in correctly classifying the test data, regardless of what dataset was used for training. However, models trained on the dataset with more features tended to display a recall score of 1.00 on the test data, which indicates that the models are likely to correctly identify all malignant cases in unseen data. As such, instead of using these models as diagnostic tools, they could instead be further developed into screening tools that help identify patients with a higher risk of malignant cancer.

Keywords: machine learning, classification, breast cancer, *All of Us*

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Specific aims

This study aims to create a predictive model by training and evaluating support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), adaptive boosting (AdaBoost), and gradient boosting (GradientBoosting) classifier models such that the best model scores at least an 0.85 AUC-ROC score, which is indicative of a well-built classification model. The trained model data can then be exported as a prototype for additional development. This study will also examine the model performance difference between two different datasets to see if there are any correlations between observation volume, feature count, and model performance.

Significance of study

As breast cancer is one of the leading causes of cancer-related death among women, research into advancing methods of predicting and diagnosing is imperative in the modern day. Additionally, treatment in later stages of the disease often requires bodily trauma to the patient, such as with surgery or chemotherapy (Miller, 2016). If detected earlier, both the patient and medical staff have more time to react and create a plan of patient care. Creating a diagnostic tool which can compute a diagnosis in hours, as opposed to days or weeks can free up medical staff for more urgent tasks, thus reducing subsequent fatigue and exhaustion error. In reducing the possibility of error from the medical staff, patient outcomes have a better average prognosis (Bell et al., 2023). This study aims to provide a proof-of-concept diagnostic tool made by machine learning models to predict the malignancy of cancer in order to expedite the diagnostic process while also examining the model performance difference between datasets of large volume and smaller feature count versus smaller volume and higher feature count.

Significance of data source (All of Us program)

Notably, this study uses volunteered and anonymized data from the *All of Us* (AoU) program to conduct its analysis. The AoU research program and its database was chosen primarily for two reasons: the ease of accessibility of large amounts of anonymized patient data and the relatively new state of the AoU research program. As the primary student researcher was completing this project for an undergraduate Honors thesis, which has the time constraint of being feasible within a one and a half school years, the AoU research program offered a large swath of usable patient data that was cleaned, processed, and anonymized. Secondly, this study aims to establish more research done using the AoU dataset, specifically the *All of Us Registered Tier Dataset v7*, since the AoU research program is relatively new. This is evidenced on the official AoU website as plans to establish the program go back as recent as 2015 (All of Us Research Program, 2023).

Chapter 2: Background Information and Literature Review

Oncological background

Cancer has been described as "any disease where the cells of the human body acquire the ability where cells of the human body acquire the ability to divide and multiply in an uncontrolled way (Miller, 2016, p. xiii)." Cancer cells are created through mutation in healthy cells caused by genetic damage, which can occur from a multitude of sources, such as radiation or viruses. Multiple mutations are often required for the display of cancerous traits, as is seen simplified in Figure 1. Some of these mutations disable certain functions of the cell, such as enabling rapid cell division and preventing the cells from self-termination. When combined, these traits result in significant tumor growth. Additionally, some mutations enable certain functions of the cell to the point that they become harmful to the body. A commonly known example of these types of mutations are mutated proto-oncogenes that promote cancerous traits, such as malignancy in tumor masses (Miller, 2016).

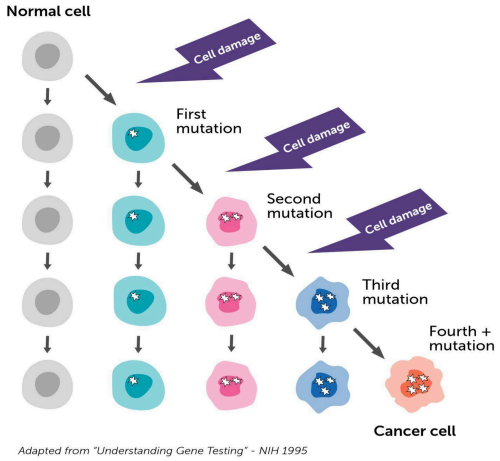

Figure 1. A simplified depiction of a healthy cell mutating into a cancer cell (Understanding gene testing, 1995).

Cardiovascular health and its importance

The word cardiovascular, by its definition, is a part of medical vocabulary that pertains to heart and blood vessels. Components of cardiovascular health include measurable statistics, like blood pressure, heart rate, and heart rhythm. A complete picture of cardiovascular health also includes the patient's previous significant medical events, particularly those that pertain to the heart and blood vessels, such as a heart attack or blood disease. Research shows that

cardiovascular health, which is the health of a person's individual heart and circulatory system, can be used as predictors of breast cancer stage. Research published by Ding et al. (2023) concluded that a patient's heart-rate variability, or HRV, may have significance in being an early diagnostic tool for breast cancer diagnosis. Wu et al. (2021) concluded that patients in more advanced stages of breast cancer had lower HRV when compared to those in earlier stages, indicating that HRV could also be used for cancer stage classification. Additionally, research by Koelwyn et al. (2020) found that a myocardial infarction, commonly known as a heart attack, "induces alterations in systemic homeostasis, triggering cross-disease communication that accelerates breast cancer." They found that early-stage breast cancer patients who experienced a cardiovascular event after cancer diagnosis additionally had an increased risk of recurrence of their cancer and cancer-related deaths.

Liquid biopsies and their importance

Liquid biopsies are defined as "a laboratory test done on a sample of blood, urine, or other body fluid to look for cancer cells from a tumor or small pieces of DNA, RNA, or other molecules released by tumor cells into a person's body fluids." From the same definition, liquid biopsies "may be used to help find cancer at an early stage" (National Institute of Health). However, the proven effectiveness of liquid biopsies is still uncertain, and its implementation is not yet common practice amongst medical providers. Despite this, there have been studies researching the practicality of using liquid biopsies as another detection method for early stage breast cancer. Keup, Kimmig, and Kasimir-Bauer researched a variety of different proteins, DNA, RNA, and cell types which could indicate early stages of breast cancer in relation to their concentration in the blood. Of these, the researchers were especially interested in circulating

tumor cells (CTCs), circulating tumor DNA or RNA (ctDNA and ctRNA), cell-free DNA or RNA (cfDNA and cfRNA), and tumor-derived extracellular vesicles (EVs). Their findings displayed a slight positive effect of using liquid biopsies as a proven method of cancer detection, but still were inconclusive as certain proof (2023). A study by Wu and Chu (2022) also studied the potential of liquid biopsies in breast cancer detection. This study focused on most of the same analytes as predictive measures, while also adding microRNA as another analyte to investigate. Like the previous study, the researchers found some positive correlation between cancer detection and the level of specific analytes, but the study was not conclusive enough to provide enough evidence for the widespread implementation of liquid biopsies. However, like the previous study, the researchers were hopeful for the future of liquid biopsies and their applications, both in breast cancer detection and other diagnostic endeavors (Wu and Chu, 2022).

As the primary student researcher, I used the above cited research on cardiovascular health and liquid biopsies to identify predicting features for my machine learning models. This initial set of features was later expanded upon to compare the performance between two different datasets, a “basic” dataset that has a smaller number of features but a large number of observations, and a more “advanced” dataset that has the additional features but a smaller number of observations. For more information, please see Chapter 3: Research Methodology.

Machine learning background

Machine learning can be roughly defined as the instruction of a computer or program to learn a function based on a set of inputs. Machine learning can be divided into two main categories, supervised learning and unsupervised learning, with exceptions and intersections thereof. Supervised learning is when input-output pairs are known, so the model can be tested on

its general correctness in analyzing general/unknown data. Unsupervised learning is more exploratory, and mainly consists of finding and understanding patterns or groups in the data (Ayodele, 2010). This study will be utilizing supervised learning models for its analysis.

In general terms, a model is the final product of the program or algorithm once trained on the input data. An example of this is a linear equation, in the form y=mx+b. When given input in two dimensions (x-axis and y-axis), a machine learning model can use the data points provided to find the line of best fit with the slope (m) and the y-intercept (b) as learned parameters. (Ayodele, 2010). This study will be utilizing support vector machine, random forest, and multilayer perceptron models for its analysis.

Support vector machines

Support vector machines are essentially data plots in high dimensions. For a dataset with n-features, an n-dimensional feature space is used to plot the data. The hyperplane, which is a subspace residing in the feature space existing in n-1 dimensions, is then found using the data points. It is optimized using a variety of loss functions to find the hyperplane of best fit. In other words, the hyperplane is optimized to find that which best separates the data. For example, a 2-dimensional input vector (x-axis and y-axis) will have a 1-dimensional line as a hyperplane. Additionally, a 3-dimensional input vector (x-axis, y-axis, and z-axis) will have a 2-dimensional plane as its hyperplane. These examples can be seen in Figure 2, with the 2-dimensional and 3-dimensional cases corresponding to the left and right sides, respectively.

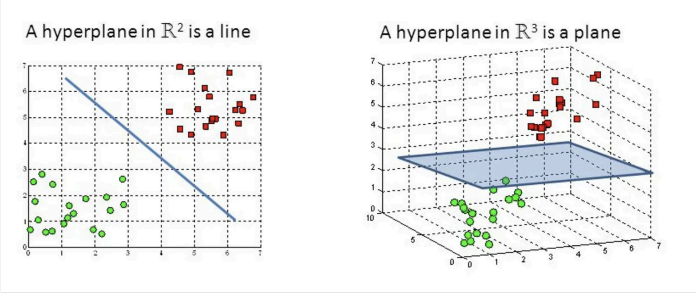

Figure 2. A depiction of a SVM in 2 dimensions and 3 dimensions (Gandhi, 2018).

Support vector machines are usually used in classification models but can be used in prediction and regression models as well, given some modifications (Ayodele, 2010).

Random forest

Random forest models consist of multiple decision trees. Decision trees, like that shown in Figure 3, are mainly used for classifying unknown input based on a variety of factors. For instance, Figure 3's decision tree works to classify an animal based on its physical characteristics, such as height, habitat, or body part size. Decision trees typically use a "splitting" rule or metric that determines which feature, or combination of features, to split on at each individual branch. Typically, the split that best separates the training data into the given output classes is chosen at each individual branch. Branches higher up along the tree (i.e. closer to the root) more significantly affect the classification of the input, while branches lower on the tree (i.e. farther away from the root) typically have less significant information about the classification of an input. As physical forests are made of trees, so too are random forests made of decision trees. Each decision tree in a random forest is typically trained on a subset of the overall features, thus narrowing the scope of that tree to a limited view of the overall input. It is through aggregating multiple decision trees with narrowed scopes that a random forest classifies information. Random forests are typically

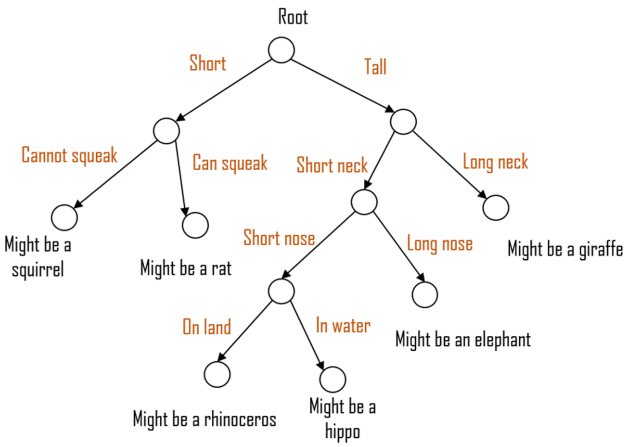

Figure 3. A depiction of a single decision tree predicting the class of a mammal based on individual characteristics (Decision tree, 2021).

implemented for classification but can also be used in regression analysis and feature selection (Ayodele, 2010).

Multilayer perceptron

Multilayer perceptrons are a type of feed-forward neural network. Neural networks are, like the name suggests, modeled after human neuroscience. Each "neuron" is a simple, nonlinear function. Based on the input, the function it represents produces a biased, weighted sum that is given to the next layer as output. For easier visualization, neural networks are often described in "layers". Each neural network has three main components: the input layer, the hidden layers, and the output layer. A neural network can have any number of hidden layers but may only have one input and output layer. This is demonstrated in Figure 4, where the input layer is green, and the output layer is yellow. Additionally, the network in Figure 4 contains only one hidden layer, which is blue. Following the direction of the arrows in Figure 4 traces the process of a feed-forward neural network. Input is received in the input layer (green), then transferred to the hidden layer (blue), where computations are performed, and the values thereof are passed onto the output layer (yellow). The output layer then aggregates the values and determines the overall output from this set of inputs. Figure 4 is representative of a MLP with a single hidden layer. MLP's are feed-forward, which means that information cannot be passed "backward" through the network. This is an important distinction because some data have important temporal or sequential relations, such as video or audio files. In other words, the state

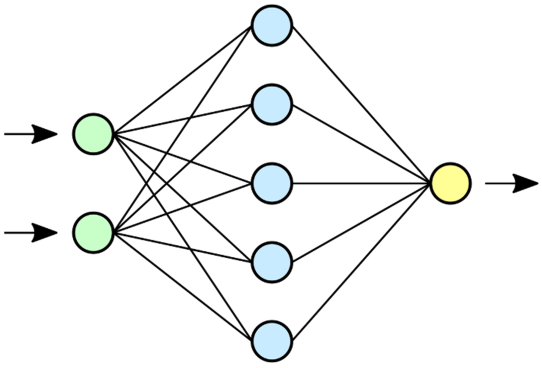

Figure 4. A depiction of a feed-forward neural network with a single input layer, hidden layer, and output layer (Dake and Mysid, 2006).

of the data beforehand significantly affects the state of the data afterward. Feed-forward networks are unable to capture these temporal or sequential relations (Ayodele, 2010).

Boosting classifiers

A popular term in the field of modern data science and machine learning is "boosting." Boosting describes the concept in machine learning in which a weak machine learning model is iteratively and adaptively strengthened with other versions specifically built to account for the previous model's weaknesses. As a simple example, let's say we want to build a boosted model to classify either a positive or negative value on some object. If a classifier guesses randomly, we can expect it to be wrong 50% of the time. Therefore, the first, weak model would learn from the mistakes of the random classifier and achieve a slightly lower error rate, say 49%. This process would then continue, such that each new model learns from the mistakes of its predecessor and theoretically stops at a very low error rate. The final aggregate model would then be a strong classifier with regards to the problem at hand due to being built from many previous weak learners (Friedman, 2001).

In my project, I will use two boosted classifier models: the adaptive boosting (AdaBoost) and the gradient boosting (GradientBoosting) classifier found in the sklearn ensemble library. While both models incorporate boosting and use decision trees as their base classifier model, the method of modifying the additional models differentiates between them. In AdaBoost, the successes and mistakes from the previous model are weighed less and more, respectively, on the additional model. In other words, the subsequent model is a duplicate of the previous model with its weights slightly changed. In contrast, GradientBoosting builds the additional models based on the negative gradient of the loss function. Using this method, the subsequent model is built

entirely from scratch, using the negative gradient of the previous model as a blueprint of sorts (Scikit-learn developers, 2024).

Intersection between medical sciences and machine learning

Breast cancers fall into four categories: hyperplasia, carcinoma in situ, invasive, and metastatic (Miller, 2016). Recently, breast cancer can be attributed to 31% of all new cancer diagnoses in women, with an estimated number of 300,000 new cases expected in 2023, containing both male and female cases (Siegel et al., 2023). As breast cancer is one of the most common types of cancer for approximately half of the population, it is imperative to research further ways to detect and combat this disease.

Aslam and Cui (2020) found great success in using a deep convolutional neural network (DCNN) on two breast cancer datasets supplied by the University of California, Irvine. They found that using a DCNN on these datasets performed well, often reaching 96% or higher in all classification metrics. They noted that the larger the dataset, the more information was available to the DCNN and thus the better it performed on average (Aslam and Cui, 2020). Another study using neural networks was done by Nageswaran et al. (2022). In this study, the researchers used linear discriminant analysis (LDA) to prepare the lung CT scans into computational information vectors, then used an artificial neural network (ANN), random forest, and K-nearest neighbors models to predict the malignancy of cancer in the scan. Amongst the three models tested, the ANN, which is a similar model to the MLP, performed the best, scoring high 90% values in accuracy, specificity, and sensitivity/recall (Nageswaran et al., 2022). Vigier et al. (2021) researched into predicting the occurrence of cancer using heart rate variability. The study did not discriminate between benign or malignant cancers, only training on and predicting the existence

of any cancer. The models trained were LDA, Naïve Bayes (NB), and RF models, respectively. They found that the RF model performed the best with 85% accuracy, but the researchers went farther and aggregated all three models to build a meta-classifier. This meta-classifier outperformed any individual model with an accuracy of 93% (Vigier et al., 2021). Another study testing and comparing multiple models was done by Wu and Hicks (2021). In this study, the researchers were trying to predict the occurrence of triple-negative breast cancer using SVM, KNN, NB, and decision tree models. Of these, the SVM performed the best when the training dataset had more than 50 observations. As is demonstrated above, the aid of diagnostic models powered by artificial intelligence and machine learning have been proven to be statistically significant, as stated by Parimbelli et al. (2021). Kazarian et al. research an alternative method of predicting breast cancer diagnosis; instead of using cardiovascular data to train the models, the research team tried predicting the diagnosis with blood-borne biomarkers such as the CA15-3 antigen or the HSP90A and PAI-1 proteins. Using multivariate logistic regression models, the researchers tested the prediction of prognosis and diagnosis of breast cancer using a single or a combination of blood-borne markers. They found that while their models had significant performance in prognostic scenarios, they were inconclusive in diagnostic scenarios (2017). Additionally, Park et al. also researched the use of blood-based diagnostic tools in predicting breast cancer. Their study involved taking liquid biopsies, otherwise known as blood draws/tests, to test the levels of tumor-associated circulating transcripts in the body across different stages of breast cancer in patients. In their study, they tested four different models: an artificial neural network (ANN), decision tree, logistic regression, and support vector machine models. Of the four types, they noted that the ANN model performed reliably and significantly better than the other three models (2022).

Chapter 3: Research Methodology

Role of the researcher

My role as the primary student researcher was to undergo all the major steps of a data science project, which includes but is not limited to: data wrangling, model training, and data analysis and visualization. Data wrangling is the process in which raw data from one or more sources are "massaged" into a format that is suitable for machine learning. This type of work often includes ensuring type-validation across a specific type of data, handling empty or null values, and unpacking or repacking data storage formats. Model training, data analysis, and data visualization are all self-explanatory sections of the data science process. My work was done using the online *All of Us* Researcher Workbench using the Python programming language. A detailed description of the project methodology is given in the next few sections.

Data selection from the AoU database

The first step was to select patients of interest and appropriate features to be included in the pared-down version of the *All of Us* (AoU) dataset. As this is a smaller research project, the population of interest was set to be persons assigned female at birth with a current or past diagnosis of either benign or malignant breast cancer. After initializing the population of interest, a variety of features and characteristics were chosen to comprise the training dataset. These features and characteristics were selected based on previous research studies cited in the literature review portion of this report. These features mainly included cardiovascular health indicators, such as systolic and diastolic heart rate, as well as blood/serum/plasma biochemistry values, such as the amount of calcium in serum or plasma per observation. The querying of the data and its features into the notebook was accomplished through SQL queries. These SQL

queries were automatically generated by the AoU website to soften the learning curve required for less SQL-experienced users. Additionally, a secondary focus of this research project, introduced after the initial launch, was to compare the model performance between two types of datasets: a dataset which has an overall larger volume of observations but a smaller number of features and a dataset which has a smaller volume of observations but a larger number of features. To accomplish this, additional features were added to a smaller dataset using Fitbit data submitted to the AoU Research Program. As a Fitbit is a wearable health device, the additional Fitbit data consisted primarily of activity, heart rate, sleep, and calorie monitoring. This smaller dataset included 4152 observations with 49 predicting features, while the larger dataset was approximately five times larger at 20756 observations albeit with 26 predicting features. Appendix A contains a full list of each feature used in any dataset which describes each feature's name and a short description.

Exploratory data analysis

The second step of the research process was to conduct exploratory data analysis, or EDA. The primary method of identifying potentially significant features in either dataset was kernel density estimation. The scipy module was used to model a Gaussian kernel density estimation, or KDE, which serves to "estimate the probability density function of a random variable in a non-parametric way (Scipy developers, 2024)." Once plotted, the KDE plot then shows the general shape of the feature's data distribution as modeled with an approximate Gaussian distribution. The vast majority of the features displayed a high similarity between patient data with malignant breast cancer against patient data with benign breast cancer. One example can be seen in Figure 5, which shows the calcium levels in patients diagnosed with

malignant and benign cancer respectively. In Figure 5, it can be seen that there is very little difference in shape between either distribution. When performing additional KDE on other features, it was found that many predicting features displayed this level of similarity between the data distributions of the malignant and benign patient data. Figure 6 is a non-exhaustive collection of predicting features that exhibit a high similarity between both benign and malignant data.

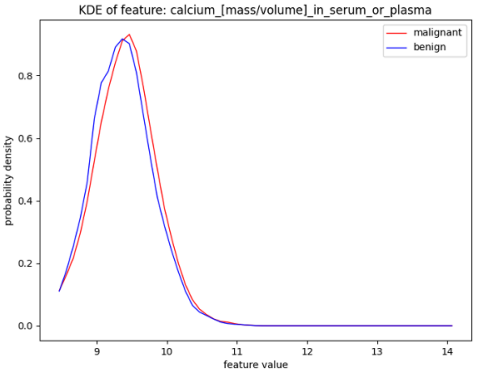

Figure 5. KDE of calcium levels [mass/volume] in patient's liquid biopsy (serum or plasma).

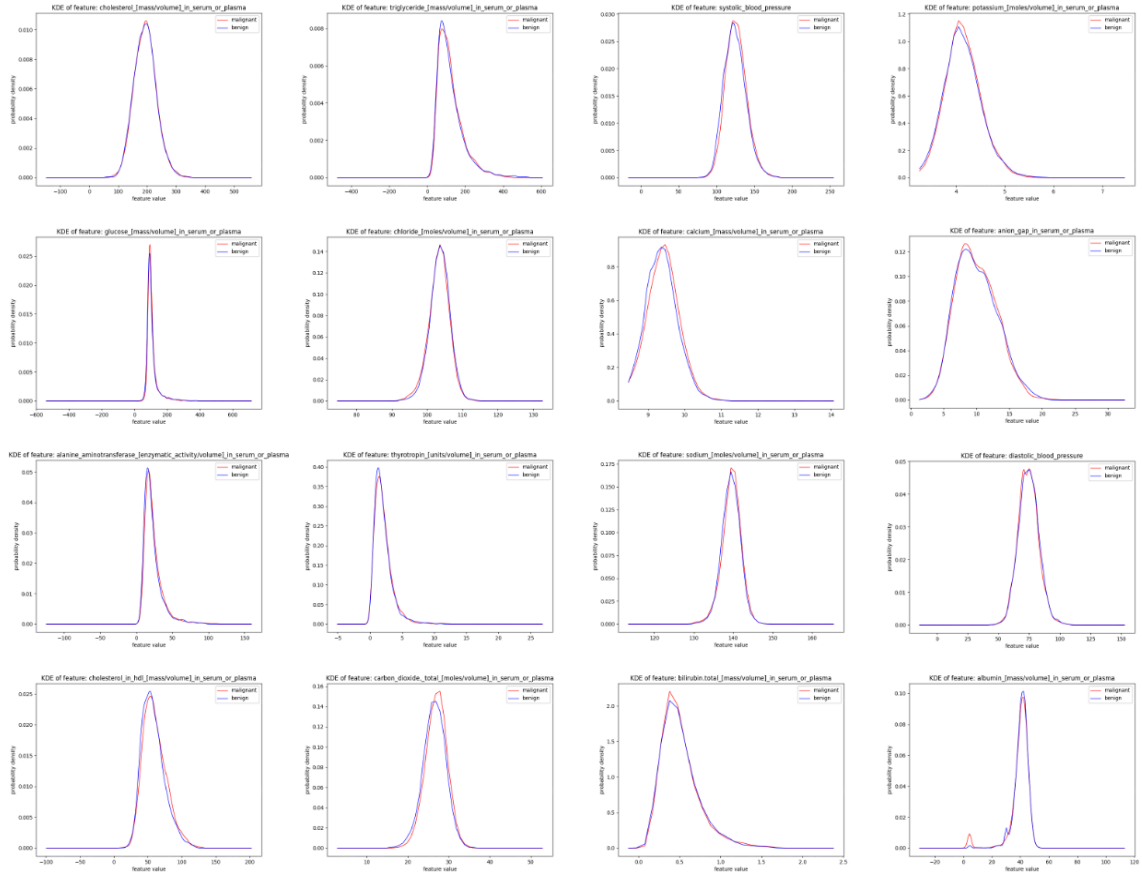

Figure 6. KDE plots of numerous liquid biopsy and cardiovascular health data features that demonstrate high similarity between benign and malignant patient data distributions.

In contrast, some features that displayed a marked difference between the malignant and benign patient data distributions were the levels of basophils, eosinophils, erythrocytes, and

leukocytes, as seen in Figure 7. Interestingly, these are all levels of blood cell count, with basophils and eosinophils being special types of white blood cells and erythrocytes and leukocytes being the overall count of red and white blood cells, respectively. The markedly different distribution shapes indicate that these features may be significant in classifying malignant or benign cancers.

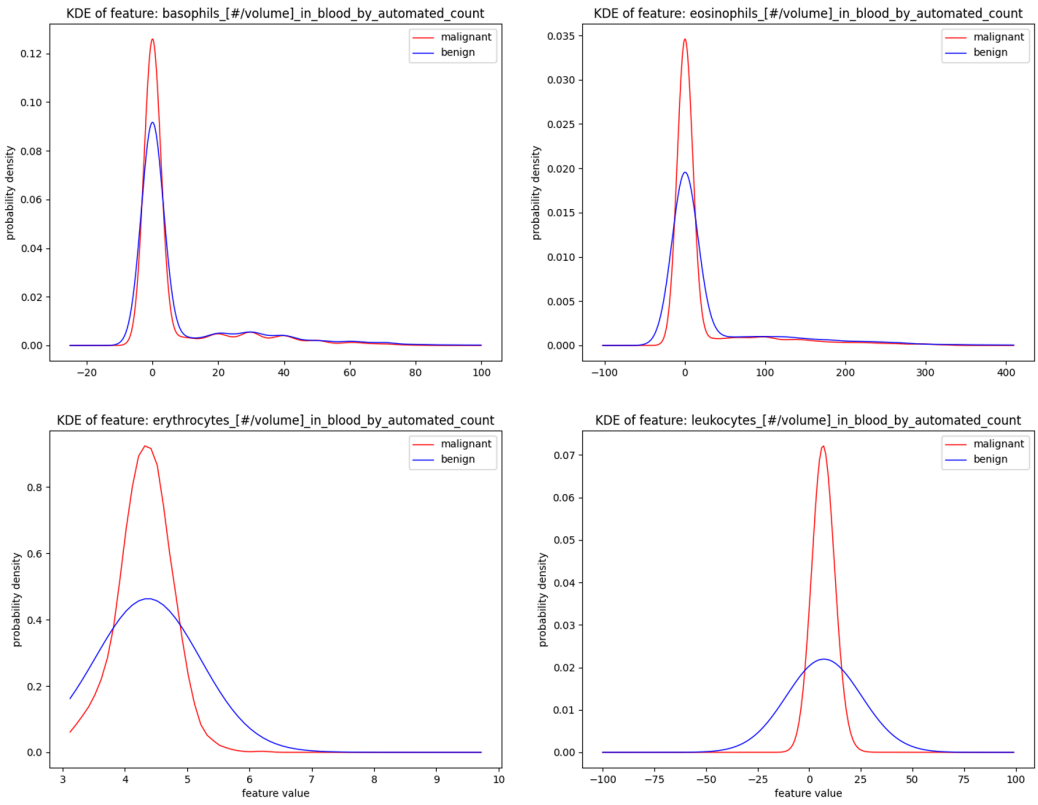

Figure 7. KDE plots of basophil, eosinophil, erythrocyte, and leukocyte levels [count/volume] in patient's liquid biopsy (blood).

Some Fitbit data features that exhibited these traits are the amount of calories burned per day and the amount of minutes spent "very active" in a day. A description of how Fitbit data defines "very active" is located in Appendix A. The data distributions of these features are displayed in Figure 8 .

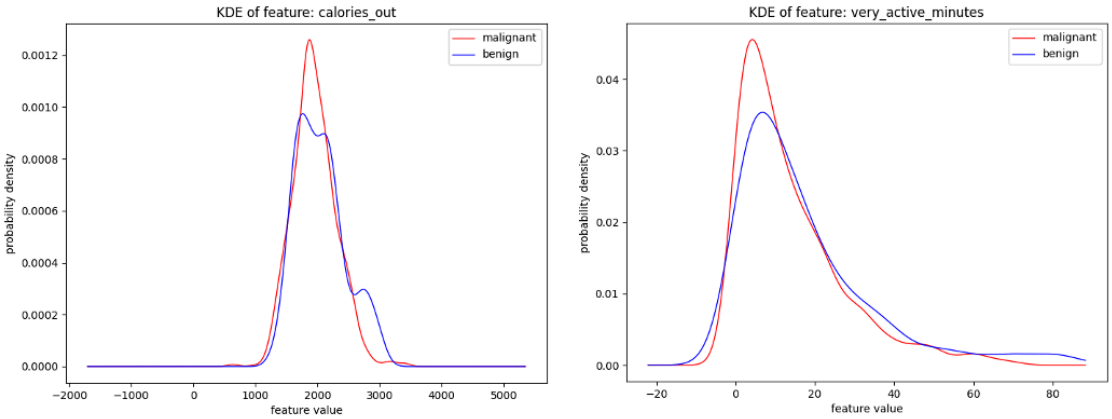

Figure 8. KDE plots of calories_out and very_active_minutes, two Fitbit data features that demonstrate slightly different data distributions for benign and malignant patient data.

In addition to visualizing the general data distribution of each predicting feature, the additional Fitbit data was analyzed against the time of first diagnosis for each individual patient. This was requested by the primary mentor, Dr. Washington, partly to determine if any lifestyle changes were made after a patient received their diagnosis. In Figure 9, the blue curve represents the year difference between the year of that specific row of patient data and the year of the first diagnosis regardless of malignancy. On the other hand, the red curve represents the year difference between the year of that specific row of patient data and the year of the first malignant diagnosis, i.e. the year the patient was first recorded as having a malignant breast cancer. Something of note is that the year difference metric used in Figure 9 was calculated on the aggregated data after preprocessing, the details of which are discussed in the next section: *Data preparation for model training*. As both curves in Figure 9 are centered slightly greater than 0, it implies that most of the Fitbit data was gathered in around the time and somewhat after first diagnosis, implying that patients had more motivation to continually

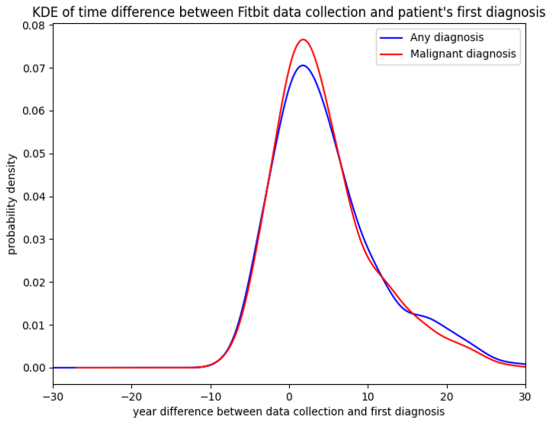

Figure 9. KDE plot of time difference between the aggregate year of Fitbit data collection versus the year of first diagnosis. The red curve indicates only malignant diagnoses, while the blue curve indicates any breast cancer diagnosis.

monitor their health. When analyzing Figures 8 and 9 together, it appears that the malignant data/curves in some Fitbit data features, such as the calories expended per day and the amount of minutes spent in "very active" forms of exercise or activity, seem to concentrate over the expected mean in contrast to the benign data/curve, resulting in higher KDE peaks. This could imply that after first diagnosis, patients with malignant diagnoses put more effort into living healthier lifestyles. However, as these differences are not very pronounced and no hypothesis testing was executed, it could also be that these differences are just due to random chance.

Data preparation for model training

The third step was to preprocess the data under the guidance of my primary mentor, Dr. Peter Y. Washington, in order to prepare it for model training and data analysis. Data stored in the AoU dataset is stored in a row-based format, which is counterintuitive to machine learning. Additionally, many of the original columns were simply metadata, which is information that describes the data, such as its origin and standardized name, rather than being usable patient data itself. After removing the metadata columns and transcribing the data into a column-based format stored in a Pandas Dataframe, patient data was aggregated by individual patient ID and year of observation. Null values were handled using three different types of imputation methods imported from the sklearn library: a SimpleImputer, KNNImputer, and an IterativeImputer. The SimpleImputer was programmed to replace any null value within a column to the median value of that column, KNNImputer would replace the null with the average value of N-number of neighbors' values for that column, and the IterativeImputer would replace the null value based on the known values of surrounding columns in a round-robin like fashion. For more information, please see the respective documentation pages for each type of imputation as linked

below in the *References* section (Scikit-learn developers, 2024). Afterwards, data sampling was manipulated in order to ensure no repeated data crossover and balanced class distributions in both datasets. In simpler terms, no data point appears in more than one dataset; all data points are unique across the smaller, larger, and test datasets. The test dataset is discussed in more detail in the "*Model evaluation*" section below. Additionally, each dataset was balanced to have a roughly equal distribution of positive and negative classifications or, in other words, benign and malignant classifications. This is due to the fact that unbalanced datasets can worsen model performance.

Models and hyperparameters

The machine learning models used in this study include: support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), adaptive boosting classifier (AdaBoost), and the sklearn gradient boosting classifier (GradientBoostingClassifier). Different hyperparameter combinations were tested for performance on each model type using 10-fold cross validation. This was done through the GridSearchCV function in the sklearn model_selection library. The hyperparameters changed per model included: 'C' (a regularization term) and 'kernel' hyperparameters for the SVM model, 'n_estimtators and 'max_features' hyperparameters for the RF model, 'n_estimators' and 'learning_rate' hyperparameters for AdaBoost and GradientBoostingClassifier models, and 'hidden_layer_sizes' and 'tol' hyperparameters for the MLP model. GridSearchCV was instructed to use the F1 score as a metric for determining the best performing model amongst the 10 folds. These models were then evaluated on common classification metrics such as accuracy, precision, and recall. They were also scored on the F1 score and AUC-ROC scoring metrics.

Model evaluation

After training, models were evaluated on their performance on a final test dataset. As industry standard, the test dataset consisted of data excluded from model training to prevent the models from "focusing" on a few specific examples. The test dataset was extracted from the intersection of the two datasets and had 1043 observations. Originally, the test dataset was planned to contain only observations from this intersection that had less than 5% missing values. However, as the additional Fitbit dataset was drastically smaller in observation volume than the main dataset, this criteria had to be loosened to ensure a sizable amount of data was chosen for the test dataset. On average, the test dataset had approximately 20% missing values per observation. To remedy this, the test dataset was treated with the same imputation method as the model whose performance was being evaluated. For example, if a model trained on KNN-imputed data was being evaluated, then the test dataset would also have missing values imputed in the same manner, using a KNN-imputation algorithm, before evaluation. The imputation for the test data is done separate from the rest of the training data for the same reason as its exclusion.

Data visualization

Data visualizations, such as an AUC-ROC and precision-recall curves, were made using the popular matplotlib library. Additionally, SHAP, an open-source explainable AI library was employed to better illustrate and communicate the machinations behind the machine learning models. SHAP stands for Shapley Additive Explanations and uses classic Shapley values from economic game theory as a way to provide both local and model-wide feature explanations. Shapley values are used to determine the contribution of any given "player" in a game theory

scenario. The SHAP library then takes this concept and applies it to explaining machine learning models, substituting the "players" for features in the model with the scenario of classification and/or prediction of unknown data (SHAP developers, 2018).

All of Us Public Data Use Statement

The AoU research program includes a demographically, geographically, and medically diverse group of participants, however, it is not a representative sample of the population of the United States. Enrollment in the AoU research program is open to all who choose to participate, and the program is committed to engaging with and encouraging participation of minority groups that are historically underrepresented in biomedical research.

This study does not directly collect its own data. Instead, research is conducted by accessing and analyzing the AoU dataset in a virtual environment. The AoU dataset is a public dataset with different tiers of accessible data (dependent on credentials and training completed) and different versions of that data as varied by date. The AoU research program is run by the National Institute of Health, an agency of the United State (of America) Department of Health and Human Services. This project uses the *All of Us Registered Tier Dataset v7* from the AoU database.

Lab trainings completed / ethics approval

As the primary student researcher, I have completed the CITI Biomedical and Biological training module, as individual patient data was used for this project. Additionally, I have completed the *All of Us* researcher training for the Registered and Controlled tier, the former

granting access to the Researcher Workbench while the latter granting access to more sensitive data, such as genomic data of individual patients.

This project did not involve direct exposure to humans, hazardous materials (radioactive materials and/or compressed gas (scuba) diving), and vertebrate animals that would require review and approval by the Committee on Human Studies (CHS), Environmental, Health and Safety Office (EHSO), Institutional Animal Care and Use Committee (IACUC), and Laboratory Animal Service (LAS), respectively.

I understand the importance of compliance with the ethical standards held by the University of Hawai'i at Mānoa and the research community as a whole. By completing the aforementioned training, I am informed of and sustained the standard of ethical research placed forth by the university.

Resources and materials

I used my personal computer to access the Researcher Workbench, containing the dataset and analytics software required to carry out the project. The Research Workbench is hosted in a cloud service on the internet and provides virtual machines and resources to run notebook environments. This is a countermeasure to prevent the direct download of patient data onto local storage media, which is strictly prohibited by the *All of Us* program.

Chapter 4: Data Analysis

	model	dataset	F1	Accuracy	Precision	Recall	AUC-ROC
1	SVC	Nonzero KNN with Fitbit	0.8450	0.7315	0.7315	1.0000	0.5174
2	RF	KNN with Fitbit	0.8450	0.7315	0.7315	1.0000	0.5148
3	ADA	KNN with Fitbit	0.8450	0.7315	0.7315	1.0000	0.5081
4	SVC	Nonzero Iterative with Fitbit	0.8450	0.7315	0.7315	1.0000	0.5075
5	ADA	Nonzero KNN with Fitbit	0.8450	0.7315	0.7315	1.0000	0.5069
6	GB	Nonzero KNN with Fitbit	0.8450	0.7315	0.7315	1.0000	0.5029
7	ADA	Iterative with Fitbit	0.8450	0.7315	0.7315	1.0000	0.5000
8	GB	KNN with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4665
9	GB	Median with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4662
10	GB	Nonzero Median with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4625
11	SVC	Median with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4535
12	SVC	Nonzero Median with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4495
13	GB	Nonzero Iterative with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4413
14	ADA	Median with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4184
15	GB	Iterative with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4181
16	MLP	Nonzero Iterative with Fitbit	0.8450	0.7315	0.7315	1.0000	0.4160
17	ADA	Nonzero Median with Fitbit	0.8450	0.7315	0.7315	1.0000	0.3982
18	RF	Nonzero KNN with Fitbit	0.8443	0.7306	0.7313	0.9987	0.5002
19	SVC	KNN with Fitbit	0.8443	0.7306	0.7313	0.9987	0.4753
20	MLP	KNN with Fitbit	0.8443	0.7306	0.7313	0.9987	0.3624
21	ADA	Nonzero Iterative with Fitbit	0.8437	0.7296	0.7310	0.9974	0.4987
22	SVC	Iterative with Fitbit	0.8441	0.7306	0.7317	0.9974	0.4594
23	MLP	Median with Fitbit	0.8437	0.7296	0.7310	0.9974	0.3544
24	RF	Iterative with Fitbit	0.8420	0.7277	0.7314	0.9921	0.5042
25	RF	Nonzero Median with Fitbit	0.8403	0.7248	0.7302	0.9895	0.4618
26	RF	Median with Fitbit	0.8382	0.7220	0.7298	0.9843	0.4693
27	RF	Nonzero Iterative with Fitbit	0.8373	0.7220	0.7321	0.9777	0.5130
28	MLP	Nonzero KNN with Fitbit	0.8300	0.7114	0.7292	0.9633	0.4746
29	MLP	Nonzero Median with Fitbit	0.7650	0.6270	0.7096	0.8296	0.3912
30	GB	Nonzero Median	0.7447	0.6280	0.7477	0.7418	0.5164

Table 1. Top 50% machine learning model scores ranked by recall, AUC-ROC, precision, then accuracy.

	model	dataset	F1	Accuracy	Precision	Recall	AUC-ROC
31	ADA	Nonzero KNN	0.7275	0.6079	0.7398	0.7156	0.5149
32	ADA	KNN	0.7272	0.6079	0.7405	0.7143	0.5161
33	GB	Median	0.7069	0.5992	0.7602	0.6606	0.5587
34	ADA	Iterative	0.6986	0.5954	0.7677	0.6409	0.5663
35	ADA	Nonzero Median	0.6982	0.6021	0.7843	0.6291	0.6022
36	ADA	Nonzero Iterative	0.6426	0.5158	0.6985	0.5950	0.4475
37	RF	Nonzero Iterative	0.6618	0.5570	0.7496	0.5924	0.5431
38	SVC	Nonzero KNN	0.6667	0.5714	0.7734	0.5858	0.6021
39	SVC	KNN	0.6647	0.5714	0.7772	0.5806	0.6035
40	MLP	Nonzero Median	0.6672	0.5762	0.7841	0.5806	0.5894
41	MLP	Median	0.6586	0.5676	0.7796	0.5701	0.5868
42	MLP	Nonzero Iterative	0.6482	0.5503	0.7579	0.5662	0.5386
43	RF	Nonzero KNN	0.6505	0.5570	0.7692	0.5636	0.5612
44	GB	Iterative	0.6466	0.5503	0.7606	0.5623	0.5620
45	MLP	Iterative with Fitbit	0.6327	0.5225	0.7234	0.5623	0.4969
46	SVC	Nonzero Iterative	0.6387	0.5379	0.7461	0.5583	0.5425
47	RF	Nonzero Median	0.6468	0.5570	0.7761	0.5544	0.5900
48	MLP	Iterative	0.6337	0.5379	0.7541	0.5465	0.5261
49	RF	Iterative	0.6414	0.5551	0.7815	0.5439	0.5760
50	MLP	KNN	0.6205	0.5216	0.7391	0.5347	0.4816
51	RF	Median	0.6312	0.5484	0.7840	0.5282	0.5950
52	SVC	Iterative	0.6138	0.5187	0.7430	0.5229	0.5257
53	SVC	Nonzero Median	0.6250	0.5513	0.8041	0.5111	0.6309
54	RF	KNN	0.6115	0.5273	0.7668	0.5085	0.5603
55	SVC	Median	0.6179	0.5446	0.8000	0.5033	0.6268
56	GB	Nonzero KNN	0.6154	0.5398	0.7918	0.5033	0.5752
57	MLP	Nonzero KNN	0.5849	0.4938	0.7308	0.4875	0.4886
58	ADA	Median	0.5372	0.4813	0.7734	0.4115	0.5338
59	GB	KNN	0.4508	0.4324	0.7714	0.3185	0.5314
60	GB	Nonzero Iterative	0.2759	0.3356	0.6804	0.1730	0.4674

Table 2. Bottom 50% machine learning model scores ranked by recall, AUC-ROC, precision, then accuracy.

After training was completed, the models were evaluated on the test dataset to test their overall performance. Using the sklearn metrics library, accuracy, precision, recall, an F1 score, and an AUC-ROC score was calculated for each model. The *model* column is shaded with a specific color corresponding to the type of model trained: support vector classifier (SVC) is blue, random forest (RF) is green, Adaboost (ADA) is red, GradientBoosting (GB) is orange, and multi-layer perceptron (MLP) is purple. Additionally, the *dataset* column is shaded as either gold or gray, for datasets with Fitbit data and those without respectively.

Discussion of objective 1: model performance and possible use as a diagnostic tool

Tables 1 & 2 show the overall ranking of the models. The models are ranked with recall being the primary ranking metric. This is due to the nature of this study; a higher recall will result in more "positive" classifications from a model. For context, a "positive" classification

would signal malignant breast cancer in that patient. In diagnosing cancer, there is more potential for life-threatening circumstances when a patient with malignant cancer is misdiagnosed as benign, rather than the opposite case. Therefore, it is more important to focus on obtaining as many true "positives" as possible, so as not to misdiagnose any patients and possibly let the cancer metastasize. Since this study corresponds "positive" to malignancy and "negative" to benign, the maximization of the true positive rate, which is analogous to the recall metric, is the most important indicator of an effective model.

The second metric used in ranking, the AUC-ROC score, is typically used to evaluate the overall performance of a classification model. A theoretical perfect classification model would result in an AUC-ROC score of 1.00. Meanwhile, a random classification model, which would assign observations as either "positive" or "negative" based on a 50-50 coin flip, is expected to have an AUC-ROC score of 0.50. As shown in Tables 1 & 2, none of the sixty machine learning models performed very well at correctly classifying the test dataset. This is evidenced by the highest AUC-ROC score being 0.63, which is only slightly better than random guessing. Figure 10 shows us a closer look at the highest scoring model with regards to AUC-ROC: the SVC model trained on the nonzero-median imputed larger dataset with fewer features. It contains two boxplots which describe the range and variability of the probability scores for the malignant and benign classes, as obtained by the predict_proba() function accessible by many sklearn models. The key to a

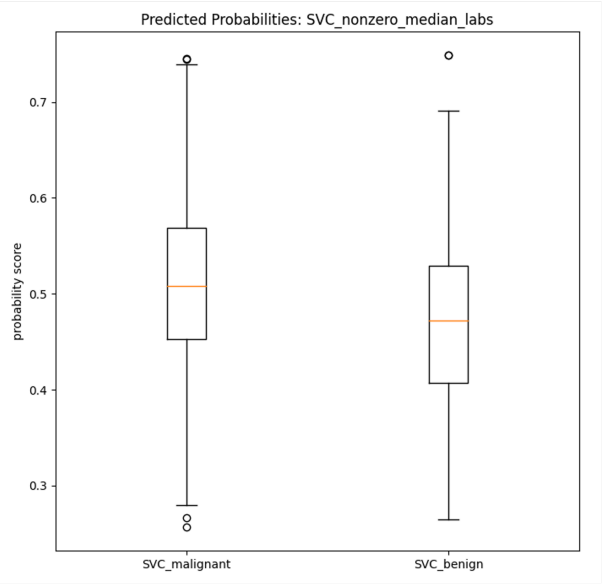

Figure 10. A boxplot of the probability scores obtained from SVC's predict_proba function. Values are separated by their true classification from the test dataset.

good AUC-ROC score, as viewed from the perspective of probability scores, is high differentiation between predictive classes. As Figure 10 demonstrates, there is a slight but noticeable difference between the malignant and benign classes, thus leading to a slight increase in AUC-ROC score when compared to random guessing, which would be 0.500. This is in contrast to Figure 11, which is a similar boxplot on the probability scores of the SVC model trained on the smaller, more complex dataset with Fitbit features imputed using a nonzero KNN imputation algorithm. The boxplots seem nearly identical to each other in shape and value range, which implies low differentiation between predictive classes. As such, this behavior leads to a low AUC-ROC score of 0.5174, which is only very marginally better than random guessing.

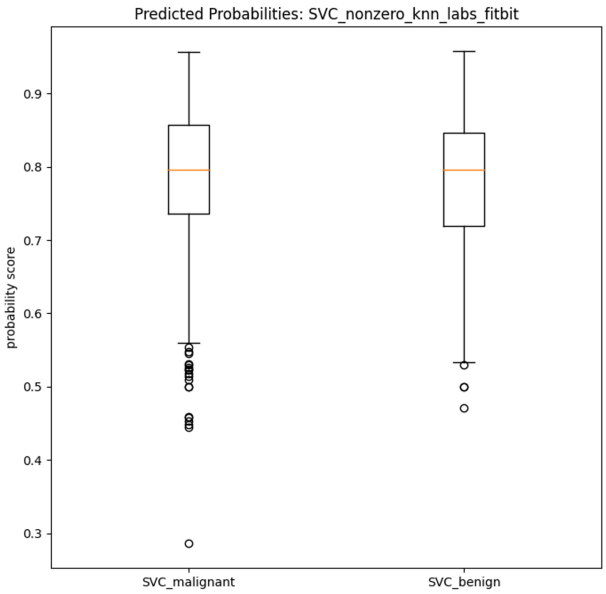

Figure 11. A boxplot of the probability scores obtained from SVC's predict_proba function. Values are separated by their true classification from the test dataset.

Given the sensitive nature of a cancer diagnosis and the poor classification performance exhibited by all sixty models, the use of these models as a diagnostic tool is highly undesirable. However, due to the high recall performance, with many models exhibiting perfect recall scores of 1.00, the purpose of these models could then be shifted from a diagnostic perspective to a screening perspective. Instead of diagnosing patients, the top-performing models could be used to screen patients and assign an increased priority to those classified as having a higher probability of malignant cancer. This repurposing would still serve the spirit of the study, which

was to help expedite the diagnostic process and subsequently help alleviate medical worker fatigue due to labor-intensive diagnostic procedures.

Discussion of objective 2: performance difference between "smaller" and "larger" datasets

Upon examination of Tables 1 & 2, there does not seem to be a significant performance difference between model types, as there are instances of high-performing and low-performing models with each model type. However, there is a clear performance difference when comparing the "larger" dataset without any additional features against the "smaller" dataset which had additional Fitbit data features and less overall observations. This is evidenced by the majority of models using datasets with Fitbit data placed within the top 50% scoring models as seen in Table 1. The opposite is true in Table 2; all but one model in Table 2 have datasets without the additional Fitbit data features. This is further evidenced by the averaged classification metrics over each dataset type in Table 3. Table 3 is also colored such that datasets with the additional Fitbit data features are colored gold and those without the additional features are shaded in gray. It is also interesting to note that despite having lower AUC-ROC and precision scores on average, the Fitbit datasets displayed higher overall accuracy and F1 scores when compared to their counterparts. When examined together, these phenomena reinforce the idea that models trained on the Fitbit datasets are more likely to assign a "positive" malignancy status to patients regardless of their true value.

dataset	F1	Accuracy	Precision	Recall	AUC-ROC
KNN with Fitbit	0.8447	0.7311	0.7314	0.9995	0.4654
Median with Fitbit	0.8434	0.7292	0.7311	0.9963	0.4324
Nonzero Iterative with Fitbit	0.8432	0.7292	0.7315	0.9950	0.4753
Nonzero KNN with Fitbit	0.8419	0.7273	0.7310	0.9924	0.5004
Nonzero Median with Fitbit	0.8281	0.7093	0.7269	0.9638	0.4326
Iterative with Fitbit	0.8018	0.6888	0.7299	0.9104	0.4757
Nonzero Median	0.6764	0.5829	0.7793	0.6034	0.5858
Nonzero KNN	0.6490	0.5540	0.7610	0.5712	0.5484
Iterative	0.6468	0.5515	0.7614	0.5633	0.5512
Median	0.6304	0.5482	0.7794	0.5347	0.5802
KNN	0.6149	0.5321	0.7590	0.5313	0.5386
Nonzero Iterative	0.5734	0.4993	0.7265	0.4970	0.5078

Table 3. Averaged classification metrics across dataset type. Rows are ranked according to recall then AUC-ROC.

SHAP visualizations and discussion

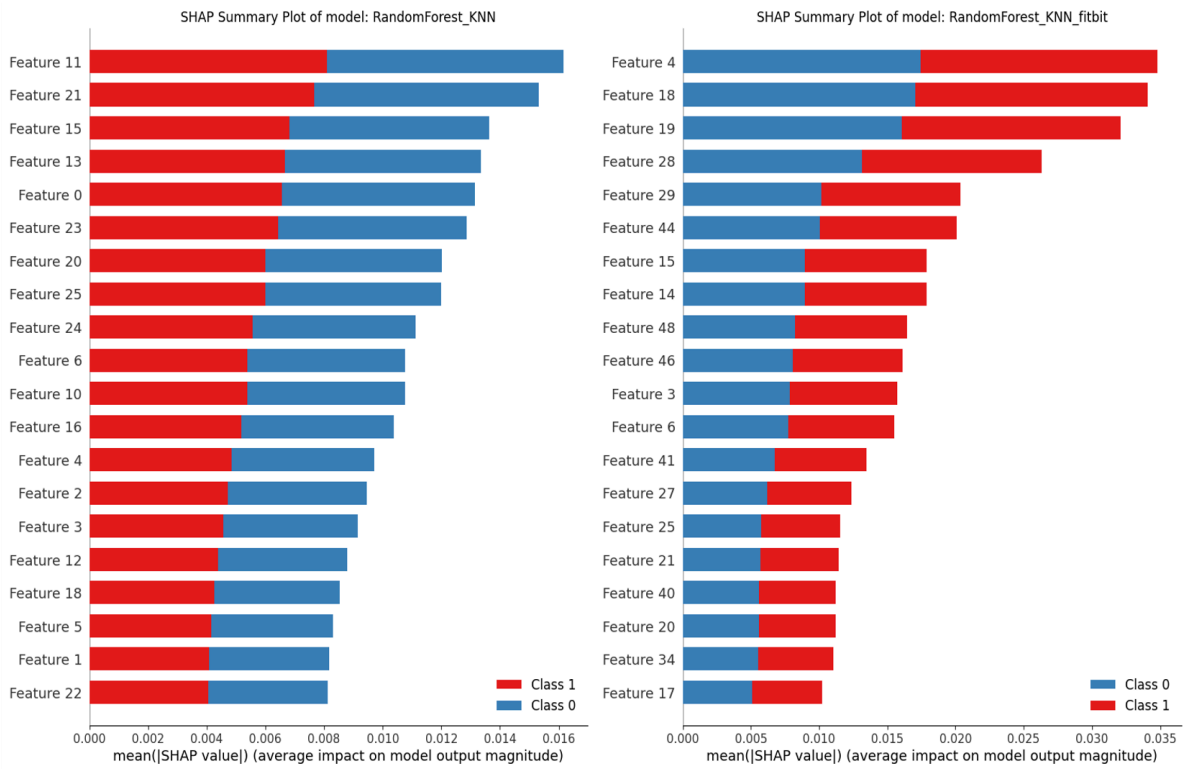

Figure 12. The SHAP summary plots of the top 20 features used by RF models on both dataset types. These models were trained on datasets imputed using the KNN imputation algorithm.

After model training was complete, two top-performing, high-recall models were evaluated using the SHAP open-source library. Additionally, their counterparts, models of the same type trained on the opposing simple/complex dataset with the same imputation method, were evaluated using SHAP as well. As mentioned previously, the SHAP library allows a "grading" of the features in the model according to how much each individual feature contributed to the overall classification. Figure 12 pictured above contains the SHAP plots containing the top twenty features for two RF models trained on KNN imputed datasets. The left plot describes the model trained on the larger but simpler dataset, while the right plot corresponds to the model trained on the smaller, more complex dataset with additional features from Fitbit data. Likewise, Figure 13 pictured below shows the SHAP plots for two GB models,

whose training datasets correspond to the left and right plots in Figure 12, respectively. Something to note is that the feature names have been transformed to show numbers instead of text; this is to prevent messy graphs as many features are long strings of text which would clutter the graph. The predicting features to which each number corresponds are found on Table 4 in Appendix A.

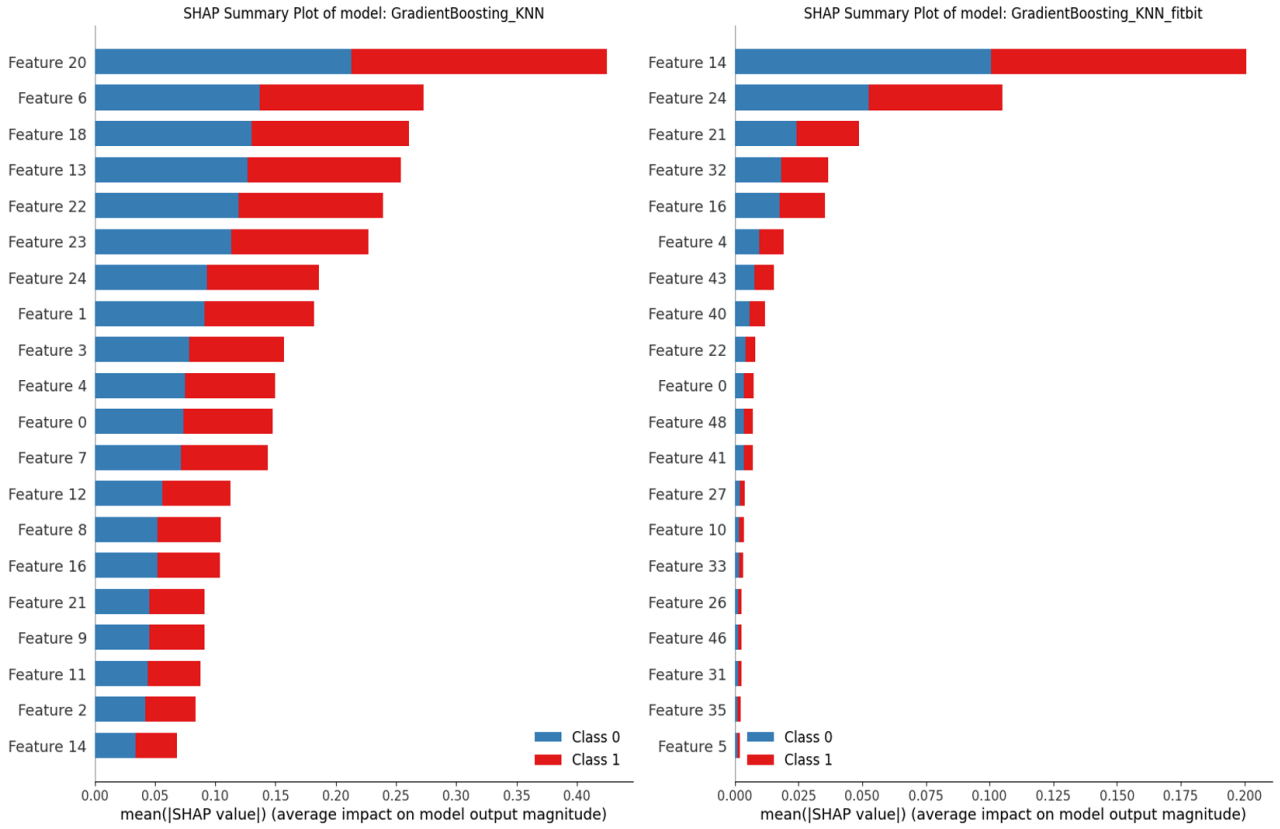

Figure 13. The SHAP summary plots of the top 20 features used by GB models on both dataset types. These models were trained on datasets imputed using the KNN imputation algorithm.

When examining the top twenty features between the GB and RF models trained on the larger, simpler dataset, as denoted by the "_labs", some features are repeated with higher value than others. Amongst these are features number 0, 6, 20, and 23, corresponding to diastolic blood pressure, leukocytes, cholesterol, and alanine aminotransferase. Interestingly, leukocyte count was one of the features identified during the preliminary EDA process as a feature that showed

potential. Additionally, sixteen of the top twenty features from both the GD and RF plots overlapped, resulting in a ratio of 80%, implying that these features have a higher likelihood of contributing significantly to the final classification for models trained on the simpler dataset.

On the other hand, examining the SHAP plots of the other two models, those trained on the smaller, more complex dataset as denoted by "_fitbit" in the plot name, seem to share less top features. Intuitively, this makes some sense, as there are simply more features to be used as the top twenty features. It stands to reason that different models might prioritize some other data distributions over others. These GB and RF models have an overlap ratio of 40%, and some of these overlapping features are features 27, 40, 41, 46, and 48. These numbers correspond to the minutes spent in bed after waking up from the main sleep in a sleep cycle, calories burned through moderate or higher physical activity, the caloric basal metabolic rate (BMR), total minutes spent at rest, and total minutes spent in intense physical activity over a day. Similar to the previous paragraph's analysis, this list also contains a feature identified as potentially significant for classification: the amount of minutes spent in intense physical activity.

Lastly, some features stood out as being shared across three or four of the evaluated models. Features 4 and 21 were present in the top twenty features of all four models, while features 14 and 20 were present in three models each. The features found in all four models correspond to the measured values for the anion gap and the level of proteins in serum or plasma. Those found in three of the four correspond to levels of albumin and cholesterol in serum or plasma.

Chapter 5: Conclusion

The first goal of this study was to provide a proof-of-concept machine learning model that could predict breast cancer malignancy fairly accurately in order to expedite the diagnostic process for medical workers. The criteria of a successful classification model was set to be an AUC-ROC score of 0.85, which would indicate a well-built classification model with relatively low error. However, none of the sixty models displayed an AUC-ROC score of above 0.63, indicating that each model was performing poorly and could not correctly discern between malignant and benign cancer patients. Despite this, many models displayed a recall, or true positive rate, of 1.00. This occurrence opened a new possibility for real-life application of these models; instead of usage as diagnostic models, these models could be used as a screening tool that can help medical professionals better identify patients that are more likely to have malignant cancers to help expedite the diagnostic process.

The second goal of this study was to compare the performance difference between a larger dataset with less features and more observations against a smaller dataset with relatively more features and less observations. With regards to this project, and the criteria pivot towards recall as the primary indicator of a successful model, it seems that the smaller dataset with additional Fitbit features performed better than their larger counterparts with less overall features. These findings could therefore imply a better performance of diagnostic screening tools if Fitbit data, or other wearable health data, is included in the training dataset.

Limitations

One of the main limitations of this study is the disproportionate ratio of data volume for the Fitbit dataset when compared to the original dataset of labs and biology measurements.

Because the Fitbit dataset was a relatively small size, there was a chance that that data available in the *All of Us Registered Tier Dataset v7* was not indicative of the larger population. Additionally, the relative size difference between the Fitbit dataset and the original dataset forced the criteria for data selection for the final test set to be somewhat lax, resulting in an approximate average of 20% missing value per row. Since the test set has a relatively large average number of missing values per row, the evaluation of model performance and the scores derived using the test set must be viewed through a cynical lens.

A second limitation is derived from the original database itself. As was the time of this writing, the relevant breast cancer cohort data within the *All of Us Registered Tier Dataset v7* contained a disproportionately larger number of Caucasian cisgender female patients when compared to other minorities, such as people of color and LGBTQ+ persons. Since this is the case, there is likely some form of bias in the models towards Caucasian cisgender female patients and, consequently, could likely have a higher chance of misclassifying these underrepresented demographics within the sample.

Future steps

As with any data science research project, future steps could include the expansion of the volume and type of data considered for training. As mentioned in the section above, the main limitations of the study were the scope and volume of the data used for model training. Therefore, collecting more raw data and/or accessing higher volumes of publicly available data could help improve model performance. This is especially true in terms of data volume with regards to the Fitbit dataset and demographic representation in the overall original dataset. Additionally, more features, such as genetic markers and qualitative survey data, could be

included as training features to possibly improve model performance. The range of hyperparameters used for training models can also be expanded to further explore the possible range of model solutions in hopes of finding a better performing model. Lastly, further research into some of the top features across the SHAP evaluated models could provide more domain-specific insight and direction in ways to better inform model training and dataset construction, more specifically the decision on which features to include and which to omit. For example, research into leukocyte levels, cholesterol levels, caloric basal metabolic rate, or anion gap measurements and their potential link to breast cancer can help improve the data processing section, thus having more usable, significant data for models to train on. Research can also compare the difference, if any, between benign and malignant patient data or data from healthy patients versus patients who have breast cancer, regardless of malignancy status.

Appendix A

Feature #	Feature Name	Description
	person_id	anonymized unique patient identification number
	year	year of observation/measurement
	identifier	a unique value obtained by computing the person_id and the year together
	malignant	a binary variable that describes the recorded diagnosis of the given patient_ID and year; 0 is BENIGN and 1 is MALIGNANT
0	diastolic_blood_pressure	measured levels of diastolic blood pressure (blood pressure in arteries during rests BETWEEN beats) in mm/HG
1	systolic_blood_pressure	measured levels of systolic blood pressure (blood pressure in arteries DURING beats) in mm/Hg
2	heart_rate	measured levels of heart rate in beats per minute
3	eosinophils_[#/volume]_in_blood_by_automated_count	measured levels of eosinophils (subset of white blood cells) in liquid biopsy (blood), originally in thousand per microliter
4	anion_gap_in_serum_or_plasma	measured difference of cations versus anions in liquid biopsy (serum or plasma), originally in millimole per liter
5	hemoglobin_[mass/volume]_in_blood	measured levels of hemoglobin protein in liquid biopsy (blood), originally in gram per liter
6	leukocytes_[#/volume]_in_blood_by_automated_count	measured levels of leukocytes (scientifc name for white blood cells) in liquid biopsy (blood), originally in thousand per microliter
7	potassium_[moles/volume]_in_serum_or_plasma	measured levels of potassium in liquid biopsy (serum or plasma), originally in millimole per liter
8	thyrotropin_[units/volume]_in_serum_or_plasma	measured levels of thyroid-stimulating hormone in liquid biopsy (serum or plasma), originally in micro-international unit per liter
9	erythrocytes_[#/volume]_in_blood_by_automated_count	measured level of erythrocytes (scientific name for red blood cells) in liquid biopsy (blood), originally in million per microliter
10	alkaline_phosphatase_[enzymatic_activity/volume]_in_serum_or_plasma	measured levels of the alkaline phosphatase enzyme in liquid biopsy (serum or plasma), originally in unit per liter
11	carbon_dioxide_total_[moles/volume]_in_serum_or_plasma	measured levels of carbon dioxide in liquid biopsy (serum or plasma), originally in millimole per liter
12	chloride_[moles/volume]_in_serum_or_plasma	measured levels of chloride in liquid biopsy (serum or plasma), originally in millimole per liter
13	glucose_[mass/volume]_in_serum_or_plasma	measured level of glucose (type of sugar) in liquid biopsy (serum or plasma), originally in milligram per deciliter
14	albumin_[mass/volume]_in_serum_or_plasma	measured levels of the albumin protein in liquid biopsy (serum or plasma), originally in gram per liter
15	aspartate_aminotransferase_[enzymatic_activity/volume]_in_serum_or_plasma	measured levels of the aspartate aminotransferase enzyme in liquid biopsy (serum or plasma), originally in unit per liter
16	calcium_[mass/volume]_in_serum_or_plasma	measured levels of calcium in liquid biopsy (serum or plasma), originally in milligram per decilter
17	hematocrit_[volume_fraction]_of_blood_by_automated_count	measured levels of hematocrit (percentage of red blood cells in total blood volume)
18	sodium_[moles/volume]_in_serum_or_plasma	measured levels of sodium in liquid biospy (serum or plasma), originally in milliomole per liter
19	bilirubin.total_[mass/volume]_in_serum_or_plasma	measured levels of bilirubin pigment in liquid biopsy (serum or plasma), originally in milligram per decilter
20	cholesterol_[mass/volume]_in_serum_or_plasma	measured levels of total cholesterol in liquid biopsy (serum or plasma), originally in milligram per deciliter
21	protein_[mass/volume]_in_serum_or_plasma	measured levels of protein in liquid biopsy (serum or plasma), originally in millimole per liter
22	basophils_[#/volume]_in_blood_by_automated_count	measured levels of basophils (subset of white blood cell) in liquid biopsy (blood), originally in thousand per microliter
23	alanine_aminotransferase_[enzymatic_activity/volume]_in_serum_or_plasma	measured levels of the alanine aminotransferase enzyme in liquid biopsy (serum or plasma), originally in unit per liter
24	cholesterol_in_hdl_[mass/volume]_in_serum_or_plasma	measured levels of cholesterol in high-density lipoprotein (HDL) in liquid biopsy (serum or plasma), originally in milligram per deciliter
25	triglyceride_[mass/volume]_in_serum_or_plasma	measured levels of triglyceride (type of lipid) in liquid biopsy (serum or plasma), originally in milligram per deciliter
26	minute_asleep_main	number of minutes spent in bed asleep in main sleep period
27	minute_after_wakeup_main	number of minutes spent in bed after waking up from main sleep period
28	minute_awake_main	number of minutes spent in bed awake before falling asleep to main sleep period
29	minute_restless_main	number of minutes spent in restless sleep (continual body movement) during main sleep period
30	minute_deep_main	number of minutes spent in deep sleep phase during main sleep period
31	minute_light_main	number of minutes spent in light sleep phase during main sleep period
32	minute_rem_main	number of minutes spent in REM sleep phase during main sleep period
33	minute_asleep_not_main	number of minutes spent in bed asleep in NOT main sleep period
34	minute_after_wakeup_not_main	number of minutes spent in bed after waking up from NOT main sleep period
35	minute_awake_not_main	number of minutes spent in bed awake before fallling asleep to NOT main sleep period
36	minute_restless_not_main	number of minutes spent in restless sleep (continual body movement) during NOT main sleep period
37	minute_deep_not_main	number of minutes spent in deep sleep phase during NOT main sleep period
38	minute_light_not_main	number of minutes spent in light sleep phase during NOT main sleep period
39	minute_rem_not_main	number of minutes spent in REM sleep phase during NOT main sleep period
40	activity_calories	total number of calories expended during activities (time spent with > 1.5 MET) throughout the day
41	calories_bmr	caloric basal metabolic rate, the rate at which calories are burned during rest to maintain vital bodily functions
42	calories_out	number of calories burned throughout the measurment period (a single day)
43	fairly_active_minutes	number of minutes spent "Fairly Active", registering at around 3.0-6.0 metabolic equivalent task (MET) units
44	lightly_active_minutes	number of minutes spent "Lightly Active", registering at around 1.5-3.0 metabolic equivalent task (MET) units
45	marginal_calories	total number of marginal estimated calories burned throughout the measurement period (a single day)
46	sedentary_minutes	number of minutes spent at rest, registering at 1.0 metabolic equivalent task (MET) units (i.e. the individual's baseline for MET)
47	steps	number of steps taken in a day (pedometer measurement)
48	very_active_minutes	number of minutes spent "Very Active", registering at around > 6.0 metabolic equivalent task (MET) units

Table 4. The complete list of features used for model training. Features are colored according to type of data: blue rows are identification columns, red rows are cardiovascular features, green rows are liquid biopsy features, and orange rows are features obtained from Fitbit Data. Only predicting features used for model training have a feature number.

Table 4 contains a full list of all features within the dataset(s) used for this project. Rows are colored according to their type: blue rows are identification data or metadata, red rows are cardiovascular health data, green rows are liquid biopsy data, and light orange rows contain Fitbit data. The below paragraphs help explain some units and context for some Fitbit features.

According to the official Fitbit Help Center, activity periods are defined by a user's energy level exceeding the base rate of 1.0 metabolic equivalent task, or MET. MET is a unit used to describe the intensity of physical activity and is defined as the rate of energy expended during activity to the rate of energy expended while at rest. A value of 1.0 MET is defined as the "rate of energy you expend during rest or sitting quietly" (Google, 2024). The values for the MET threshold for "lightly active", "fairly active", and "very active" minutes in the table above were derived from research done by Semanik et al. (2019).

Fitbit data categorizes a user's largest and most common sleep period as their "main sleep". In monophasic sleep cycles, the norm across many societies, this sleep period is typically the nighttime sleep period in which most restful sleep is undergone for the average individual. Any sleep periods outside of this "main sleep" period, such as midday naps, are categorized as "not main sleep" (Google, 2024).

References

All of Us Research Program. (2023, October 18). *All of Us Research Program Overview.* https://allofus.nih.gov/about/program-overview

Aslam, M. A., & Cui, D. (2020). Breast cancer classification using deep convolutional neural network. *Journal of Physics: Conference Series, 1584*, 1-10. https://doi.org/10.1088/1742-6596/1584/1/012005

Ayodele, T. O. (2010). Types of machine learning algorithms. In Y. Zhang (Eds.), *New Advances in Machine Learning* (pp. 19-48). https://doi.org/10.5772/9385

Bell, T., Sprajcer, M., Flenady, T., & Sahay, A. (2023). Fatigue in nurses and medication administration errors: A scoping review. *Journal of Clinical Nursing*. https://doi.org/10.1111/jocn.16620

Dake, Mysid, (2006). Neural network [Online image]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Neural_network.svg

Decision tree [Online image]. 2021. Huawei. https://forum.huawei.com/enterprise/en/machine-learning-algorithms-decision-trees/thread/710283-895

Ding, L., Yang, Y., Chi, M., Chen, Z., Huang, Y., Ouyang, W., Li, W., He, L., & Wei, T. (2023). Diagnostic role of heart rate variability in breast cancer and its relationship with peripheral serum carcinoembryonic antigen. *PloS one*, *18*(4), e0282221. https://doi.org/10.1371/journal.pone.0282221

Friedman, Jerome H. (2001) "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Gandhi, R. (2018). Hyperplanes in 2D and 3D feature space [Online image]. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Google. (2024). *What are Active Zone Minutes or active minutes on my Fitbit device?*. Fitbit Help Center. https://support.google.com/fitbit/answer/14236509?sjid=4601488186208110167-NC#zippy=%2Chow-does-my-fitbit-device-calculate-active-minutes

Google. (2024). *How do I track my sleep with my Fitbit device?*. Fitbit Help Center. https://support.google.com/fitbit/answer/14236407?hl=en&ref_topic=14236503&sjid=10342584190828550496-NC#exp&zippy=%2Ccan-my-fitbit-device-log-a-nap:~:text=Each%20month%2C%20wear,Versa%204%20only).

Kazarian, A., Blyuss, O., Metodieva, G., Gentry-Maharaj, A., Ryan, A., Kiseleva, E. M., Prytomanova, O. M., Jacobs, I. J., Widschwendter, M., Menon, U., & Timms, J. F. (2017). Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples. *Br J Cancer*, 116(4), 501-508. https://doi.org/10.1038%2Fbjc.2016.433

Keup, C., Kimmig, R., & Kasimir-Bauer, S. (2023). The diversity of liquid biopsies and their potential in breast cancer management. *Cancers*, 15(22), 5463. https://doi.org/10.3390/cancers15225463

Koelwyn, G. J., Newman, A. A. C., Afonso, M. S., van Solingen, C., Corr, E. M., Brown, E. J., Albers, K. B., Yamaguchi, N., Narke, D., Schlegel, M., Sharma, M., Shanley, L. C., Barrett, T. J., Rahman, K., Mezzano, V., Fisher, E. A., Park, D. S., Newman, J. D., Quail,

D. F., Nelson, E. R., … Moore, K. J. (2020). Myocardial infarction accelerates breast cancer via innate immune reprogramming. *Nature medicine*, *26*(9), 1452–1458. https://doi.org/10.1038/s41591-020-0964-7

Miller, M. E. (2016) *Cancer*, Momentum Press.

Nageswaran, S., Arunkumar, G., Bisht, A.K., Mewada, S., Kumar J. N. V. R. S., Jawarneh, M., & Asenso, E. (2022). Lung cancer classification and prediction using machine learning and image processing. *Biomed Research International, 2022*, 1-8. https://doi.org/10.1155/2022/1755460

National Institute of Health. (n.d.). Liquid Biopsy. In *NCI Dictionary of Cancer Terms*. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/liquid-biopsy

Ojha, U., & Goel, S. (2017). A study on prediction of breast cancer recurrence using data mining techniques. *7th International Conference on Cloud Computing, Data Science & Engineering – Confluence, 1*, 527-530. https://doi.org/10.1109/CONFLUENCE.2017.7943207

Parimbelli, E., Wilk, S., Cornet, R., Sniatala, P., Sniatala, K., Glaser, S. L. C., Fraterman, I., Boekhout, A. H., Ottaviano, M., & Peleg, M. (2021). A review of AI and Data Science support for cancer management. *Artificial Intelligence in Medicine, 117*, 7-8. https://doi.org/10.1016/j.artmed.2021.102111

Park, S., Ahn, S., Kim, J. Y., Kim, J., Han, H. J., Hwang, D., Park, J., Park, H. S., Park, S., Kim, G. M., Sohn, J., Jeong, J., Song, Y. U., Lee, H., & Kim, S. I. (2022). Blood test for breast cancer screening through the detection of tumor-associated circulating transcripts. *International Journal of Molecular Sciences*, 23(16), 9140. https://doi.org/10.3390/ijms23169140

Scikit-learn developers. (2024). *sklearn.ensemble.AdaBoostClassifier*. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

Scikit-learn developers. (2024). *sklearn.ensemble.GradientBoostClassifier*. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

Scikit-learn developers. (2024). *sklearn.impute.IterativeImputer.* scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

Scikit-learn developers. (2024). *sklearn.impute.KNNImputer*. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

Scikit-learn developers. (2024). *sklearn.impute.SimpleImputer*. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

Scipy developers. (2024). *scipy.stats.gaussian_kde*. scipy. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html

Semanik, P., Lee, J., Pelligrini, C., Song, J., Dunlop, D. D., & Chang, R. W. (2019). *ACR Open Rheumatology, 2(1)*, 48-52. https://doi.org/10.1002/acr2.11099

SHAP developers. (2018). *Welcome to the SHAP documentation*. SHAP. https://shap.readthedocs.io/en/latest/

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians, 73*(1), 17–48. https://doi.org/10.3322/caac.21763

Understanding gene testing [Online image]. 1995. FORCE: Facing Our Risk of Cancer Empowered.

https://www.facingourrisk.org/info/hereditary-cancer-and-genetic-testing/hereditary-cancer/genes-and-cancer

Vigier, M., Vigier, B., Andristch, E., & Schwerdtfeger, A. R. (2021). Cancer classification using machine learning and HRV analysis: preliminary evidence from a pilot study. *Scientific Reports, 11*(1), 1-12. https://doi.org/10.1038/s41598-021-01779-1

Wu, H. J., & Chu, P. Y. (2022). Current and developing liquid biopsy techniques for breast cancer. *Cancers*, 14(9), 2052. https://doi.org/10.3390/cancers14092052

Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. *Journal of Personalized Medicine, 11*(2). https://doi.org/10.3390/jpm11020061

Wu, S., Chen, M., Wang, J., Shi, B., & Zhou, Y. (2021). Association of short-term heart rate variability with breast tumor stage. *Frontiers in physiology*, *12*, 678428. https://doi.org/10.3389/fphys.2021.678428

Glossary

AoU	acronym for *All of Us*, a research program conducted by the National Institute of Health to aggregate anonymized patient data in a single place for public study
machine learning	the instruction of a computer or program to learn a function based on a set of inputs
model	a program/algorithm that can find patterns or make decisions on previously unseen data
training	the act of feeding input data to a machine learning model for it to find patterns or make decisions
hyperparameters	parameters that affect the training process of a model (manually set by the data scientist)
KDE	acronym for kernel density estimation, a statistical approach "to estimate the probability density function of a random variable in a non-parametric way" (Scipy developers, 2024)
SVM	support vector machine model (in depth explanation in Chapter 2: Background Information and Literature review)
RF	random forest model (in depth explanation in Chapter 2: Background Information and Literature review)
MLP	multilayer perceptron (in depth explanation in Chapter 2: Background Information and Literature review)
AdaBoost	adaptive boosting model (in depth explanation in Chapter 2: Background Information and Literature review)
GradientBoosting	gradient boosting classifier, specifically the algorithm implemented in the sklearn library (in depth explanation in Chapter 2: Background Information and Literature review)
AUC-ROC score	a metric to judge the performance of a classification model which emphasizes the amount of correctly classified observations; the best score is 1.00 (stands for *area under the "Receiver Operating Characteristic" curve*)
F1 score	a metric to judge the performance of a classification model emphasizes correctly classifying most to all observations of the positive class; the best score is 1.00 (this score is the harmonic mean of precision and recall)