UNIVERSITY OF CALIFORNIA,
IRVINE

Application of Machine Learning for Risk Determination in Cancer Occurrence and Recurrence

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSPHY

in Biomedical Engineering

by

Yasaman Fatapour

Dissertation Committee:
Associate Professor James P. Brody, Chair
Professor William C. Tang
Associate Professor Edward Kuan

2023

DEDICATION

To my husband Adam Abiri, my best friend and unwavering support,

To my parents Maryam Tabatabai and Mohsen Fatapour
who have sacrificed daily for me and are thousands of miles away,

To my grandmother Parvin and my aunt Susan
whose love have guided me,

To my beloved siblings, Amirhossein and Sarvenaz

And to all the people who are fighting against cancer,

This dissertation is dedicated to you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude and appreciation to my committee chair, Dr. James P. Brody. His unwavering support, invaluable guidance, and insightful feedback have been the cornerstones of my entire PhD research journey. He consistently provided encouragement, listened attentively to my ideas and concerns, and challenged me to critically examine my work. His expertise, patience, and dedicated mentorship have not only shaped and refined the direction of my research but have also had a profound impact on my personal growth.

I would like to extend my thanks to the members of my committee, Dr. Abraham Lee, Dr. William Tang, Dr. Edward Kuan, and Dr. Christine King. Their constructive critiques, thoughtful insights, and rigorous examination have enriched the quality of my research.

I extend my sincere appreciation to my committee member, Dr. William Tang for giving me the opportunity to work as a teaching assistant in BME 180 with him. Through this experience, he not only taught me the significance of education but also demonstrated patience and innovation in his teaching approach, as well as the application of engineering techniques to various biomedical projects..

I would like to express my gratitude to my committee member, Dr. Edward Kuan for his deep interest in our work and his valuable advice on approaching new ideas with a combined and multi-disciplinary mindset. The collaboration with his team has provided strong support for the completion of this work. His constructive critiques and thoughtful insights have significantly enriched the quality of my research.

I want to express my appreciation to Dr. Christine King and Dr. Khine for affording me the chance to collaborate with them during the teaching of the BME 180 class. This experience has illuminated the profound influence of engineering design on enhancing patient outcomes, while also exposing me to the rewarding challenges of working with diverse teams and effectively imparting knowledge. This opportunity has provided me with invaluable experience that I am truly grateful for.

I would also like to convey my appreciation to Maggie Mulcare, whose consistent assistance with administrative matters has been invaluable.

This research was financially supported by The Henry Samueli School of Engineering at The University of California, Irvine. We are also grateful for the additional funding provided by UC Irvine Graduate Division through the Department of Teaching Excellence and Innovation Fellowship and the Samueli Foundation.

VITA

YASAMAN FATAPOUR

2014	B.S. in Material Science and Engineering Amirkabir University of Technology, Tehran, Iran
2016	M.S. in Material Science and Engineering Amirkabir University of Technology, Tehran, Iran
2017-2018	Graduate Researcher, Department of Biomedical Engineering, University of Houston, Tx
2020-2023	Graduate Researcher & Teaching Assistant, Department of Biomedical Engineering, University of California Irvine
2021	M.S. in Biomedical Engineering, University of California, Irvine
2023	Ph.D. in Biomedical Engineering, University of California, Irvine

FIELD OF STUDY

Machine Learning in Cancer Risk Determination

PUBLICATIONS

Fatapour Y, Abiri A, Kuan EC, Brody JP. Development of a Machine Learning Model to Predict Recurrence of Oral Tongue Squamous Cell Carcinoma. *Cancers*. 2023; 15(10):2769. https://doi.org/10.3390/cancers15102769

Fatapour Y, Brody JP. Genetic Risk Scores and Missing Heritability in Ovarian Cancer. Genes (Basel). 2023 Mar 21;14(3):762. doi: 10.3390/genes14030762. PMID: 36981032; PMCID: PMC10048518.

Fatapour Y, Brody JP. Using chromosomal-scale length variation to predict breast cancer occurrence and recurrence with machine learning, Cancer Research 83 (2023): 772.

ABSTRACT OF THE DISSERTATION

Application of Machine Learning for Risk Determination in Cancer Occurrence and Recurrence

by

Yasaman Fatapour

Doctor of Philosophy in Biomedical Engineering

University of California, Irvine, 2023

Associate Professor James P. Brody, Chair

Despite diagnostic advancements, the development of reliable methods for assessing the risk of cancer occurrence still remains a challenge. Effective risk assessment models can improve monitoring and increase change of early detection and intervention. Existing risk estimate models rely primarily on data collected from single institute and often lack racial and ethnic diversity. In addition, many existing statistical models do not sufficiently incorporate inheritance factors.

With the recent advancements in genetics, big data and artificial intelligence, precision medicine can become a reality. In this study we leveraged the available data from the largest cancer databases to develop machine learning models for predicting cancer occurrence.

In this work, we developed a novel framework for extracting recurrence cases from the SEER dataset and identified cases within a 5-year and 10-year period. Machine-learning prediction models for oral tongue squamous cell carcinoma (OTSCC) cancer recurrence was then developed based on sociodemographic and clinical variables. Among the top trained classification models, the Gradient Boosting Machine model performed the best, achieving

81.8% accuracy and 97.7% precision for 5-year prediction. Moreover, 10-year predictions demonstrated 80.0% accuracy and 94.0% precision.

In addition to the aforementioned model, we also explored a novel strategy that incorporates structural variations in germline DNA, specifically chromosomal scale-length variation (CSLV), to assess individuals' genetic risk scores. This approach enabled comprehensive analysis of copy number variations (CNVs) across large segments of the human genome, capturing variations that may contribute to the inheritance of cancer risk. The strategy was tested on two unique datasets, UK Biobank and NIH All of Us. The viability of the approach first evaluated by developing a machine learning model for predicting breast cancer recurrence based on data from UK-Biobank. The model developed based on CSLV values of 489 patients, all of whom were of white race and had experienced breast cancer recurrence, as well as a negative class consisting of age-matched and under-sampled patients from 13,478 cases who had not experienced breast cancer recurrence. The model showed an average AUC of 0.54 on unseen split of data, however, since the model was developed solely based on CSLV values, it could not comprehensively evaluate an individual's risk for breast cancer recurrence.

In order to determine whether CSLV could be used for developing risk assessment models for occurrence of cancer, we relied on the NIH All of Us dataset. The developed risk estimate model accurately evaluated individuals' risk of developing breast, colorectal, and oral cavity cancer solely based on calculated CSLV values. The AUC of the trained model on unseen split of data was 0.70, 0.68, and 0.69, respectively. By calculating the odds ratio relative to the whole population, we found that patients who were scored by the model in the top 10% were 14, 12, and 13 times more likely to develop that specific type of cancer. The diversity of the

datapoints in the All of Us dataset allowed us to examine our developed model's performance for predicting an individual's risk of breast cancer across different races. This analysis provided valuable insights into the generalizability of our model among different racial groups.

In conclusion, the advancements in machine learning, next-generation sequencing, and big data have allowed the development of effective risk assessment models for various types of cancer. More importantly the techniques introduced in this work are easily translatable to the study of other complex diseases. We hope that this investigation encourages future studies that incorporate clinical, sociodemographic and genetic variables for detection and treatment of cancer. As healthcare datasets continue to grow in size and computational power continues to increase, there is, without a doubt, great promise for significant strides in precision medicine and personalized healthcare.

CHAPTER 1: INTRODUCTION

Healthcare research has made significant progress with recent advancements in computing technologies. These advancements, including improvements in computational power, storage, and data transfer speeds, have opened up new opportunities for conducting innovative research in biomedical and healthcare applications.[1]

Assessing a patient's risk for complex diseases, such as cancer, can have profound implications for disease management, prevention strategies, and optimizing patient outcomes. Stratifying patients based on their risk of specific types of cancer enables early detection, leading to improved survival rates among high-risk individuals. Additionally, a patient's risk for cancer extends beyond the initial diagnosis; assessing the risk of cancer recurrence is crucial for optimizing treatment strategies and making informed decisions regarding disease monitoring.[2,3]

Several risk estimate models have been developed to predict cancer risk across diverse populations.[4,5] However, these developed models possess certain limitations that need to be addressed. The majority of prior risk assessment tools heavily rely on data from Caucasian populations, giving rise to concerns about their accuracy when applied to underrepresented communities. Moreover, studies focusing on cancer recurrence statistics are often confined to small sample sizes at the institutional level, leading to challenges in predicting recurrence on a larger scale. Additionally, some of these models lack the inclusion of inheritance factors in their risk determination calculation, posing further limitations.

For models that have integrated inheritance factors, certain limitations are present. The advent of Next Generation Sequencing (NGS) technologies has resulted in an abundance of data

at reduced costs. [6] However, the current polygenic risk models primarily rely on Genome-Wide Association Studies (GWAS), which predominantly focus on single nucleotide polymorphisms (SNPs) in germline samples. [7–9] Unfortunately, these models often overlook the nonlinear effects of genetic variants, leading to the common outcome of a panel of associated gene mutations or genetic variations in such studies. Consequently, risk scores derived from these models tend to inaccurately predict individualized risks for most patients. These scores usually report odds ratios (OR) or the area under the curve (AUC) of the receiving operating characteristics curve (ROC). For example, in one study the AUC of polygenic risk score for breast cancer is reported as 0.68, while in another study for ovarian cancer, the odds ratio for the top quantile relative to the entire study population is reported as 3.4 1.77. [4,10,11]In light of these limitations, our aim is to enhance the predictions of cancer risk score models on an individualized basis for most patients.

With the rise of big data and advancements in NGS technologies, the amount of available clinical and genetic data has exponentially increased. [1,12,13]In these large, geographically, and racially diverse datasets lies the potential to gain a deeper understanding of various diseases and the efficacy of treatments, including cancer. We believe that by harnessing the power of these extensive databases, we can create an accurate personalized risk estimate model. However, the challenge lies in analyzing such vast amounts of unfiltered and often disorganized data, hindering the development of effective solutions. The main objective of this study is to identify relevant clinical and genetic features and transform them into actionable and reliable information that can aid physicians in enhancing screening and prevention strategies.

Machine learning (ML) has emerged as a game-changer in addressing various medical challenges, including the battle against cancer.[1] ML algorithms have proven their exceptional ability to analyze pertinent information and construct models based on nonlinear effects. In this research, our goal is to leverage the available clinical and genetic data from large-scale, geographically diverse datasets and develop a risk estimation model for cancer occurrence using advanced machine learning algorithms. By employing machine learning classification algorithms, we have adopted a novel approach to assess the risk of cancer occurrence, overcoming the limitations observed in previous models. Through this endeavor, we aim to pave the way for more accurate and personalized cancer risk assessments, ultimately leading to improved patient care and better outcomes in the fight against cancer.

In our initial investigation, we centered our focus on utilizing sociodemographic and clinical variables from the SEER dataset to develop a risk estimation model for cancer recurrence. We applied this framework to two specific types of cancer: oral tongue squamous cell carcinoma (OTSCC) and breast cancer, aiming to predict the 5- and 10-year risk of recurrence for these cancers.

Furthermore, In the second objective of this investigation, we aimed to expand the application of risk estimation models and explore the impact of inheritance factors by adopting a novel strategy that incorporates the effect of structural variations in germline DNA. We focused on chromosomal scale-length variation (CSLV) as a promising approach for assessing genetic risk scores and incorporated it into the development of our machine learning models.[14–16] The CSLV approach allowed us to comprehensively analyze copy number variations (CNVs) across large segments of the human genome, enabling us to capture variations that may

contribute to the inheritance of cancer risk. [16–19] To test our hypothesis, we leveraged the extensive genetic data available in the UK Biobank and All of Us datasets. Utilizing these comprehensive resources, we evaluated the viability of our developed models for predicting cancer recurrence and assessing the risk of developing specific types of cancer, including breast cancer, colon cancer, and oral cavity cancer. Through this investigation, we aimed to provide valuable insights into the role of inheritance factors in cancer risk and contribute to the development of personalized risk estimation models for improved cancer outcomes.

By leveraging the resources provided by the SEER, UK Biobank and All of Us datasets, we have taken significant steps toward developing more accurate and personalized risk estimation models for improved cancer risk prediction and patient care.

CHAPTER 2: OBJECTIVES &SPECIFIC AIMS

Cancer is a complex and devastating disease that continues to pose significant challenges in the field of biomedical research.[20] Early detection of cancer is significantly important in improving patient outcomes and reducing mortality rates. [21] We believe that systematic data analysis, filtering, and feature engineering on large cancer databases, coupled with advanced machine learning algorithms, can allow development of effective and accurate risk assessment models for both cancer occurrence and cancer recurrence. The model can provide risk estimates at an individual level and classify patients into high-risk and low-risk groups. We aim to leverage this knowledge to improve early detection and prognosis.

Objective 1: Assessing the Risk of Cancer Recurrence using Clinical and Sociodemographic Variables

Our first aim of this investigation focuses on developing a technique for analyzing the largest nationally representative medical registry for cancer, SEER (Surveillance, Epidemiology, and End Results), with the goal of developing a reliable, predictive risk determination model, for cancer recurrence. For this purpose, we developed a novel algorithm to identify cases of cancer recurrence from the SEER database. The developed framework can be applied to identify recurrent cases of any type of cancer. For the initial objective of this study, we investigated the possibility of developing machine learning models for predicting time-specific recurrence in cancer patients at 5 and 10-year intervals. We used commonly available sociodemographic and clinical variables as the basis for our models' features. Specifically, we focused on two types of cancer: oral tongue squamous cell carcinoma (OTSCC) and malignant breast tumors. To test the performance of the developed model, we calculated sensitivity, precision, accuracy and area

under the curve (AUC) of trained ML model on unseen test data set. The model with the best performance has been used for further analysis.

Objective 2: Investigating the Impact of Inheritance Factors on Cancer Recurrence Risk

The results of the investigation on the SEER dataset were very promising, however by analyzing the SHAP value and feature importance graph, we found out that the number of prior tumors is consistently one of the most influential features on predicting cancer recurrence. This recognition raised some new questions, specifically regarding genetic predispositions. We hypothesized that inheritance factors may play an important role in the first occurrence and recurrence of cancer. To gain a better insight of how genetic factors can be integrated into a predictive model for recurrence, we investigated the possibility of developing a model that can include inheritance factors as sole features for predicting recurrence of complex diseases such as a cancer. Since genetic data was not present in SEER, this specific aim focuses developing a risk estimate model for cancer recurrence on available genetic data in UK-Biobank dataset.

To test our hypothesis, we incorporated genome structural variations at the chromosome level as predictive features in our model. For this objective, we utilized germline copy number variation (CNV) information from individual patients and transformed it into Chromosomal Scale Length Variation (CSLV) values. This approach allowed us to reduce the dimensionality of the problem and condense the genomic information into a smaller number of parameters. CSLV, which evaluates copy number variations across autosomes chromosomes, enabled us to represent the data with only 22 variables, significantly reducing the complexity compared to using millions of single nucleotide polymorphisms (SNPs). By employing this

approach, we were able to leverage available tools in machine learning algorithms to develop a robust predictive model. To achieve this objective, we conducted an analysis of genetic data obtained from breast cancer patients enrolled in the UK Biobank study. Our primary focus was to identify patients within the UK Biobank who experienced breast cancer recurrence during the study period. Using the calculated CSLV numbers as the sole input, we constructed a machine learning model specifically tailored to this task. To evaluate the model's performance, we compared its prediction results on a completely randomly shuffled dataset.

Objective 3: Developing Chromosome Scale Length Variability-Based Genetic Risk Scores for Predicting Cancer Occurrence

While our initial objectives focused on predicting recurrence, in this effort we would like to take a step further and delve deeper to explore the possibility of predicting the actual occurrence of cancer. Our goal is to leverage germline genetic information to develop a predictive model for assessing the risk of cancer occurrence. By identifying individuals who are at higher risk, we can potentially enhance cancer prevention efforts by targeting screening and other preventive strategies towards those who are most likely to benefit.

In line with our previous objective, we employ genome structural variations at the chromosomal level (CSLV) as a predictive parameter for model development. For our final objective in this investigation, we have tested our hypothesis using recently released genetic data from the NIH All of Us study. The inclusion of this dataset is particularly valuable as it encompasses diverse groups that have been historically underrepresented in biomedical datasets. This presents a unique opportunity to develop an accurate and personalized risk estimation tool for cancer. In pursuit of this objective, we have explored the development of

various machine learning models for cancer risk assessment at different primary sites. These models have been evaluated using various metrics to analyze their performance. By leveraging the dataset from the NIH All of Us study, our goal is to improve the inclusivity and effectiveness of our predictive models for assessing cancer risk. This research seeks to contribute to the advancement of personalized medicine and improve outcomes in cancer prevention and early detection.

CHAPTER 3: BACKGROUND

Cancer Statistics

Cancer is the second leading cause of death after heart disease since 1970 in the United States. It is responsible for one in eight deaths worldwide. In 2018 there were more than 18 million new cases and 9.5 million cancer-related deaths worldwide. Based on incidence data collected by the Surveillance, Epidemiology, and End Results program; there were 1,898,160 new cancer cases and around 600'000 cancer deaths in the United States in 2020. [22–24]Due to the advancements in early diagnosis and treatment options, the cancer death rate has fallen from its peak in 1991 through 2018 after increasing for most of the 20th century. Figure 1 illustrates the trend in cancer incidence and mortality rates by sex between 1975-2018. [23]

Cancer encompasses more than 100 different diseases with diverse risk factors and epidemiology. It originates from most of the cell types and organs of the human body, which

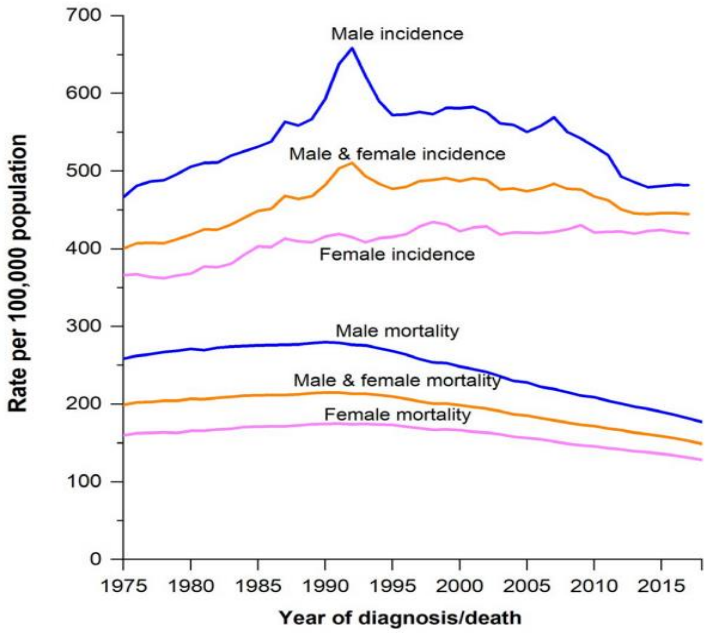

Figure 1: Trends in cancer incidences and mortality rates (1975-2018) in United States.

can invade beyond normal tissue boundaries due to the unrestrained proliferation of cells and metastasize to distinct organs. [25]

Cancer Recurrence

Cancer recurrence is a prevalent and challenging issue that significantly impacts both patients' quality of life and healthcare systems.[26]Typically, cancer recurrence is defined as the return of cancer after a period of undetectability. [27]It can be classified into three main categories based on its location and extent: local recurrence, regional recurrence, and distant recurrence. Cancer recurrence is a complex problem and may be due to numerous factors such as, type of the cancer, stage of the cancer, histology of a tumor genetic factors, age, and types of treatment. The complexity of cancer recurrence necessitates a thorough understanding of its underlying mechanisms and risk factors. For instance, studies have demonstrated that women with estrogen receptor-positive primary breast cancer generally have a higher rate of cancer recurrence, even after 5 years of the initial diagnosis, compared to those with estrogen receptor-negative disease.[28]

The early detection of cancer is crucial for improving patient survival and quality of life.[22,23] Studies have shown that patients diagnosed with cancer at earlier stages (stage I and stage II) have a higher chance of survival compared to those diagnosed at later stages (stages III-IV). Late-stage diagnosis often necessitates more intensive and invasive treatments, leading to longer-lasting side effects and poorer outcomes. [29–31]Unfortunately, the availability of limited early detection tools has contributed to high rates of late-stage diagnosis, resulting in suboptimal patient outcomes. Frequent screening is currently the main approach for early cancer detection. However, it is essential to ensure that screening programs are effective and yield meaningful results before implementation. Some screening tests may produce false positive results, leading to unnecessary invasive procedures and causing anxiety for both patients and healthcare providers. Moreover, many screening tests are expensive, placing a significant financial burden on individuals from lower economic classes. [32,33]

Cancer Risk Estimate Models and Their Limitations

Indeed, there is a crucial need for an individualized and effective risk estimation model that can accurately classify individuals into high-risk and low-risk groups. Such a model would provide precise risk estimates tailored to each individual, enabling early detection of cancer and significantly improving patient outcomes. [4,34] By identifying high-risk individuals, appropriate interventions can be implemented promptly, leading to better treatment strategies and increased chances of survival. Additionally, the utilization of personalized risk estimation models would reduce the reliance on current generalized screening tools, which often produce false positive results and cause unnecessary side effects and anxiety for patients. [35]By embracing advanced technologies and incorporating comprehensive genetic and clinical data, these models can pave the way for a new era of precision medicine, where each patient's unique risk profile informs their personalized care plan.

Artificial Intelligence

Recent advancements in computer science and data technologies have opened up unique opportunities for addressing various medical challenges. The emergence of big data, characterized by its vast variety, high volume, and rapid velocity, provides access to large-scale, geographically diverse datasets that hold the potential to offer deeper insights into the causes and outcomes of numerous medical conditions, including cancer. [26] [36]However, effectively interpreting and extracting meaningful insights from the massive amounts of data can be a daunting task. Analyzing hundreds of thousands of inputs and finding connections between seemingly unrelated information presents a significant challenge. [37] Nonetheless, with the aid of advanced machine learning algorithms and data analytics, we can unlock valuable knowledge and patterns from these datasets, ultimately leading to better understanding, improved diagnostics, and more personalized approaches in medical research and patient care.

Artificial Intelligence (AI), along with its subset, machine learning, is at the forefront of understanding and analyzing the vast amount of data that is being generated across various domains, including healthcare. AI represents the integration of human-like intelligence into machines, enabling them to emulate human decision-making and logic.[38]

Machine learning, as a subset of AI, was first introduced by Arthur Samuel in the late 1950s. Samuel, a pioneer in computer gaming and AI, used this concept to train computers to play checkers without any direct human intervention. The core focus of machine learning is to empower machines with the ability to learn from data, recognize patterns, and make informed decisions on their own. The foundation for artificial neural networks was laid in 1943 by McCulloch and Pitts, who conceptualized a theoretical model based on the connections and communication between human neurons. However, due to the limited performance capabilities of the systems at that time, these ideas remained dormant for a while. [1,39]

Machine learning (ML) has found extensive applications in healthcare, with three main areas of focus. Firstly, ML is applied to medical imaging, including MRI, CT, and PET scans, to enhance diagnostic accuracy and efficiency. Secondly, ML is used for natural language processing of medical documents, enabling the extraction of valuable information from vast amounts of unstructured data. Lastly, ML is employed in genetics to predict diseases and gain deeper insights into their underlying mechanisms.[1]

ML can be categorized into two main learning models: supervised and unsupervised learning. More recently, a third learning method called reinforcement learning has been introduced, which some literature considers as a distinct third category. These diverse ML

approaches have opened up new avenues for addressing complex healthcare challenges and improving patient outcomes. [40,41]

Supervised learning

In supervised learning, a model makes predictions based on labeled data, where the data consists of one or more inputs or features and a corresponding "labeled" output or target variable to be predicted. The data is typically split into training and test datasets. During the training step, the model learns from the labeled data to improve its predictions on new, unseen data in the test dataset.

This learning method is commonly used in classification and regression algorithms, including random forests (RF), decision trees (DT), Naïve Bayes models, linear and logistic regression, support vector machines (SVM), and neural networks. Neural networks can also be trained through supervised learning.[42] The schematic of the supervised learning method is shown in **Error! Reference source not found.**.

Unsupervised learning

In unsupervised learning, we do not have labeled data or known output. Instead, the

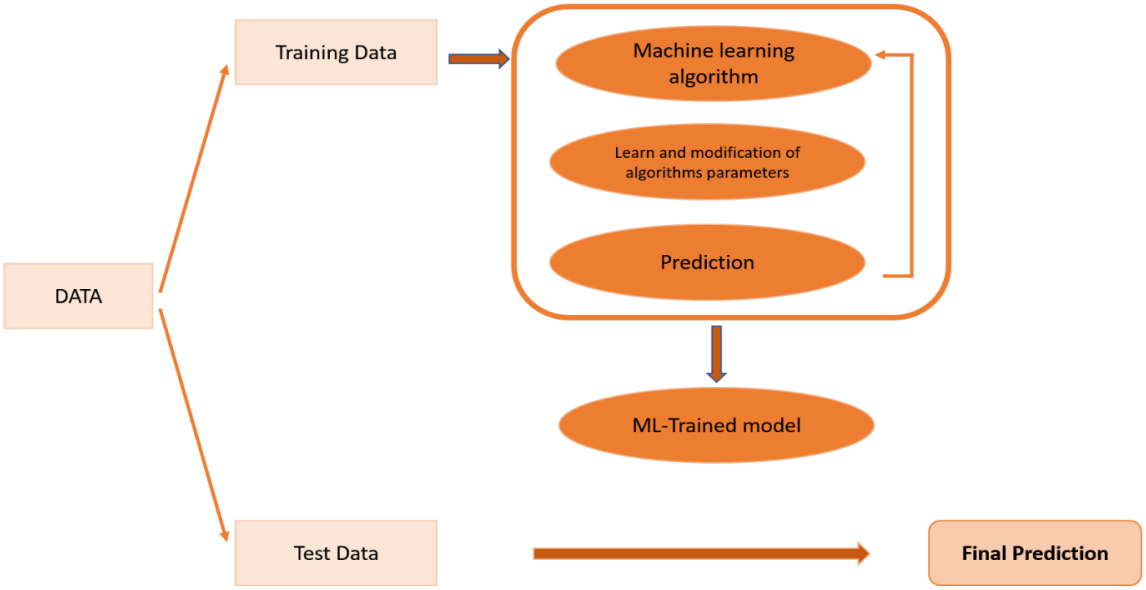

Figure 2: Schematic of supervised learning method

model identifies patterns and relationships in the input data and clusters the data based on their inherent characteristics found within the dataset. The goal is to group similar data points together based on their similarities, without any prior knowledge of the specific categories or labels.

Unsupervised learning is particularly useful when we want to explore the structure and patterns within the data, discover hidden relationships, or gain insights into the underlying distribution of the dataset. Clustering algorithms, such as k-means clustering and hierarchical clustering, are commonly used in unsupervised learning to group data points into clusters based on their similarities.[43]

Reinforcement learning

In reinforcement learning, the model learns by receiving rewards or penalties for its previous actions in an environment. The goal of the model is to adopt the optimal behavior that maximizes the total cumulative reward over time.

In this learning method, the model interacts with the environment and takes actions to achieve a specific goal. After each action, the model receives feedback in the form of rewards or penalties, indicating the quality of its decision. Through trial and error, the model learns to associate actions with the expected rewards and adjusts its strategy accordingly to maximize the total reward.

Reinforcement learning is commonly used in scenarios where the optimal decision-making strategy is not explicitly known, and the model needs to learn from its interactions with the environment to improve its performance over time.[44]

Machine Learning Algorithms

There are several machine learning algorithms that are mainly used for classification problems. In this section we will focus on algorithms that have been utilized in our model development. Since the focus of our study is developing a classification model to identify high risk patients from low-risk patients we only utilized algorithms that work best on binary classification problems.

Decision tree

Most of these algorithms that have been used for this study are developed based on a decision tree. Decision tree algorithms are a popular and intuitive machine learning technique used for both classification and regression tasks. The algorithm works by recursively partitioning the data into subsets based on the values of input features, creating a tree-like structure where each internal node represents a decision based on a feature, and each leaf node corresponds to a class label or a regression output. Decision trees are easy to interpret and visualize, making them valuable for understanding the underlying decision-making process, Figure 3.

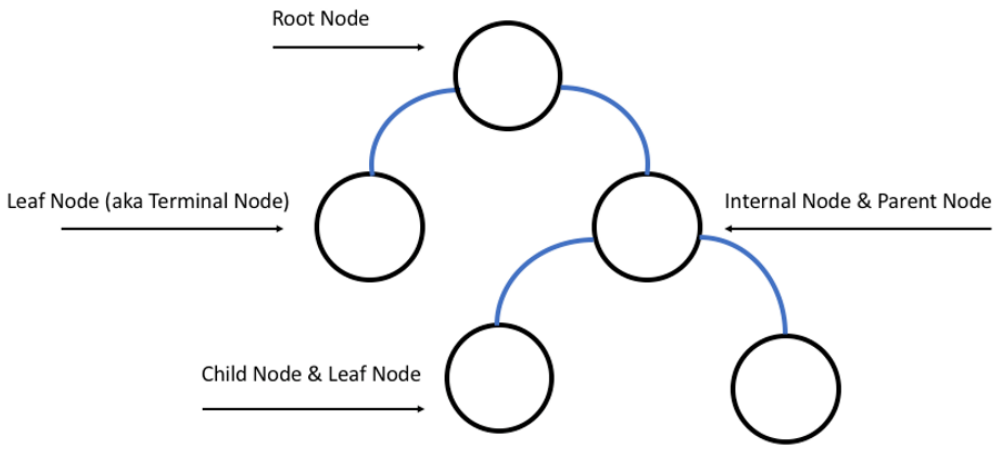

Figure 3: Constructing decision tree schematic

However, they can be prone to overfitting, where the model becomes too complex and fails to generalize well to new data. To address this, ensemble methods like Random Forests and Gradient Boosting are commonly used to combine multiple decision trees and improve overall performance.[45]

Gradient Boosting Machine

Gradient Boosting Machine (GBM) is an ensemble learning technique that combines the predictions from a series of weak classification trees to generate a strong predictive model. Unlike other ensemble methods that use equal-weighted trees, GBM employs a sequential approach where each tree takes into account the errors made by its predecessors and builds upon them. This step-by-step process gradually moves towards the best fit, descending to minimize the error (measured by the minimum squared loss function) and effectively handling imbalanced classes. However, one of the potential drawbacks of GBM is the risk of overfitting, which can be addressed through cross-validation. By verifying results on different randomly

selected datasets and utilizing leave-out testing for final prediction, the risk of overfitting is mitigated, ensuring a more reliable and robust model.[46]

Distributed Random Forest

Distributed Random Forest is a potent algorithm used for solving classification and regression problems. It consists of multiple weak decision trees that are independently ensembled. The predictions made by these individual trees are then averaged to obtain a more accurate prediction for a class or numeric value. One of the strengths of Random Forest is its ability to include hundreds or even thousands of trees, which enhances its performance on noisy data. As the number of trees increases, the variance is reduced, making the model more robust. This feature makes Distributed Random Forest particularly effective in handling noisy data, making it well-suited for tasks involving multi-class objects and bioinformatics, where statistical noise is prevalent. [47]

Generalized Linear Model

The Generalized Linear Model (GLM) is a sophisticated statistical modeling approach that encompasses several other models, such as Linear Regression, Logistic Regression, and Poisson Regression. GLM is versatile and can be applied to both classification and regression problems. To use GLM effectively, the data should meet certain criteria, such as being random, independent, and balanced. While the outcome or labeled variable does not necessarily need to follow a normal distribution, it should belong to an exponential family, which includes distributions like binomial, Poisson, and multinomial distributions. GLM provides a flexible

framework for analyzing various types of data and is widely used in statistical modeling and data analysis.[48]

Deep Learning

An Artificial Neural Network (ANN) is a powerful algorithm composed of multiple layers and nodes that simulate the behavior of biological neurons. ANN has found extensive application in disease genomics research. Deep learning, a subset of ANN, is built on a multilayer feedforward architecture that can include a large number of hidden layers. The training of the feedforward ANN is achieved through stochastic gradient descent using backpropagation. Despite its effectiveness, the downside of ANN lies in the complexity introduced by the number of hidden layers, making it challenging to interpret and understand the model fully. However, ANN tends to perform exceptionally well with large-sized datasets, making it a valuable tool for processing and analyzing vast amounts of genomic data.[49]

Stacked Ensemble Learning

The stacked ensemble method is based on the concept of leveraging multiple machine learning algorithms to enhance the overall performance of the final model. It achieves this by using a process called stacking, where predictions from various machine learning algorithms are combined to find the optimal combination. The meta-algorithm, which forms the final model, is constructed from multiple base algorithms, such as Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), and others. By blending the predictions of these diverse algorithms, the stacked ensemble method can significantly improve the model's predictive capabilities. [50]

Application of Machine Learning in Cancer Research

Over the past decade, scientific efforts have demonstrated the utility of the advent rise of available big data in health care and machine learning in guiding cancer diagnosis and management. [8,51] Specifically, many studies have applied advanced machine learning techniques and statistical models for predicting tumor recurrence, patient survival and developing data-driven risk estimate models in the context of various cancers. [14,52–57] Recently, Alabi et al. and Karadaghy et al. showed the capacity for ML to elucidate models for predicting recurrence and survival, respectively, in oral tongue squamous cell carcinoma patients. [58,59] In another work done by Ahmad et al. and Tseng et al. machine learning models were developed for predicting 2-year risk of breast cancer recurrence and identifying risk factors of ovarian cancer recurrence respectively. [60,61]

However, as with many of their predecessors, these studies were limited by the small samples of patients from which their models were trained. Many of these studies were conducted using institutional-level databases, which often lack geographical, racial, and ethnic diversity.[58,60] Even studies developed based on big data may not be medically applicable, as they often fail to include the time interval between the first and second diagnosis in their analysis. Additionally, some of these studies are based on very specific clinical variables, limiting their generalizability to broader populations.[58,62]

On the other side, current risk studies for cancer prediction, have incorporated inherited factors in a very limited way, looking mainly at cancers caused by single mutations. [63–65] Even the most popular risk assessment models for cancer, such as Gail model for breast cancer are

developed based on data from the Caucasians and they have limitation on the way that incorporate family histories and inheritance factors.[66,67]

The current challenge in cancer risk estimate models is to expand them into scalable models that provide accurate risk estimates for individuals. With advancements in technology, there is a vast and diverse biomedical dataset available, allowing for inclusion of data from underrepresented communities to enhance predictions for different backgrounds and ethnic groups. With application of these available data and efficient feature selection we are aiming to improve early cancer diagnosis and increase survival rates. A novel strategy involving inheritance factors, focusing on structural level differences in chromosomes rather than single nucleotide mutations, will address limitations of current genetic score models. By utilizing advanced machine learning algorithms and implementing this strategy in feature selection, we can better understand diseases caused by combinations of mutations. Developing such a risk estimate model can guide disease management, optimize patient outcomes, and significantly impact early cancer detection and preventative treatments. These models can have a significance impact in early diagnosis of cancer and will enable medical community to look further preventative treatments.

CHAPTER 4: ASSESSING THE RISK OF CANCER RECURRENCE USING CLINICAL AND SOCIODEMOGRAPHIC VARIABLES

In recent years, the adoption of machine learning (ML) in the medical field has brought unique perspectives and solutions to various medical challenges. Notably, ML has been applied in tumor diagnosis, tumor recurrence, and patient survival in the context of various cancers. However, studies utilizing ML for predicting tumor diagnosis, recurrence, and patient survival have encountered limitations such as small sample sizes, limited applicability to specific cohorts, low prediction accuracy, and lack of dataset diversity. [1,57,58,68]

The small sample sizes often result in models that may not generalize well to broader patient populations, potentially leading to biased or unreliable predictions. Furthermore, the limited applicability of these models to specific cohorts or populations may hinder their widespread adoption in clinical practice.

Low accuracy in predictions is another challenge faced by many of these machine learning-based models. The complexity and heterogeneity of cancer make it challenging to achieve consistently high accuracy in predicting tumor occurrence and patient outcomes.

Despite significant advancements in cancer diagnostics, reliable prognostic systems for assessing the risk of cancer occurrence remain a challenge due to the limitations mentioned earlier. As a result, the development of robust and generalizable prognostic models has been hindered. Prognostic systems are crucial for providing personalized risk assessments to patients, guiding treatment decisions, and ultimately improving patient outcomes. Moreover, there have been fewer studies focusing on predicting cancer recurrence.

For our first aim in this investigation, we concentrated on leveraging data from one of the largest cancer databases to overcome the limitations associated with recent studies. By utilizing this vast dataset, we aim to pave the way for more precise and personalized cancer risk assessment for predicting recurrence, ultimately leading to better patient care and improved cancer outcomes.

4.1: SEER Dataset: Identifying Cancer Recurrence Cases.

The Surveillance, Epidemiology, and End Results (SEER) program, in particular, provides one of the largest cancer databases in the United States and represents nearly 48% of the national population. The 2000-2018 SEER database is a deidentified registry that reports cancer incidence and survival data of the national population, serving as one of the most comprehensive efforts for tracking oncological cases within the U.S. Due to the massive scale of available data, this work utilized SEER as its target database. [69,70]

To identify patients with specific cancer recurrence throughout the study period a novel algorithm was implemented. A unique framework to leverage the expansive SEER database to generate highly representative prediction models for cancer recurrence. In this strategy, a new approach was implemented for extracting cases from the SEER database with the goal of identifying cases with recurrence of cancer within 5-year and 10-year periods. As further detailed in the sections below, SEER*Stat version 8.3.9 (Surveillance Research Program, National Cancer Institute) was used to extract data from 18 SEER registries from 2000 to 2018.

Each individual case in the SEER dataset was defined by a unique patient identification number. Cases were first grouped according to their patient IDs, and subsequently sorted within their groups using their sequence numbers. Next, a series of validations were performed for all patients and their respective cases. These validations focused on minimizing errors in later classification steps by eliminating conditions where the "state of recurrence" (recurrence = true/false) could not be determined with absolute confidence based on the available SEER data. All cases corresponding to patients with missing or unknown values for any variable critical for analysis, including Total Number of Malignant Tumors, Sequence Number, Survival Months, and Year of Diagnosis, were filtered out. In the final step of the algorithm, we computed the target outcome variable, "Will Recure". This variable, was computed for each

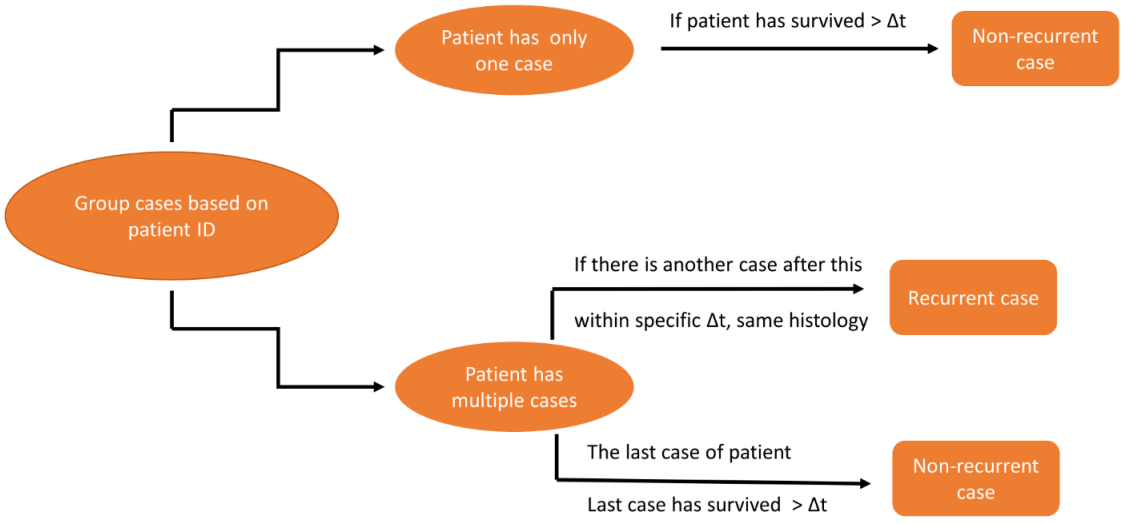

Figure 4: Schematic of developed algorithm to identify patients who have had cancer recurrence in SEER database

individual case, defining whether a case would recure in locoregional sites or the same primary site within the defined period of time (5 and 10 years). A non-recurrence was defined as a patient that had only one primary tumor and survived longer than the target window (e.g., 5 years). Conversely, if there was another recurrence of the cancer within the target window and in the same region as the initial tumor, then the case was marked as “Will Recur” = true. It is worth noting that, based on the algorithm above, the last case for a patient with multiple primary tumors (i.e., multiple case) would be marked as “will not recur” if the patient survived longer than the target window without another recurrence of cancer. This is critical as it tends to indicate a successful treatment. Figure 4 illustrates the schematic of developed algorithm used to identify recurrence cases. We applied this algorithm on the all the primary sites that listed in SEER database and calculated the recurrence rate for each primary sites, Figure 5.

As a proof of concept, we chose to develop a recurrence risk estimation machine learning

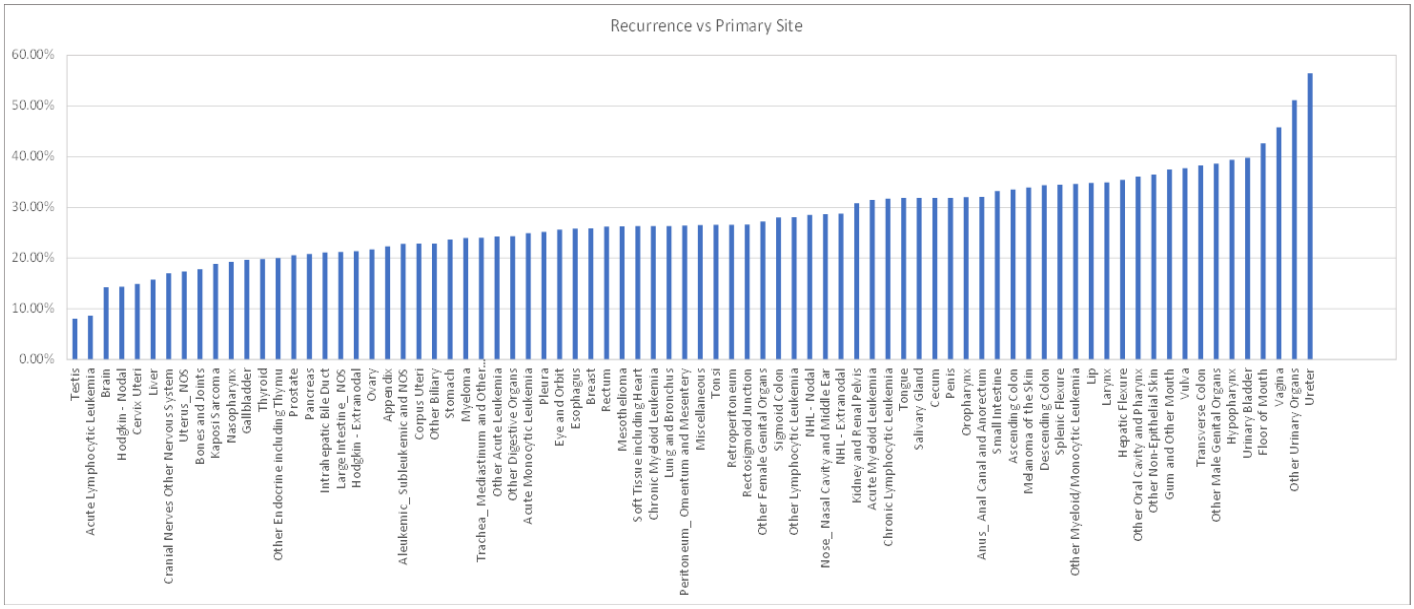

Figure 5:Recurrence rate vs primary sites on SEER database.

model for two prevalent types of cancer: oral tongue squamous cell carcinoma, one of the most common neoplasms of the head and neck, and breast cancer, the most common type of cancer in females.

4.2: Machine Learning Model Development Using H2O.ai

We utilized the H2O AI platform (H2O.ai, Inc, Mountain View, CA) in conjunction with R statistical computing environment (version 3.6.1; The R Foundation for Statistical Computing) to train and test numerous machine learning models. Our primary objective was to identify the most effective model for predicting locoregional recurrence of OTSCC and local recurrence of malignant breast tumors. To ensure proper validation and testing, the dataset was split into training (80%) and test (20%) sets. Employing H2O's Automl function, we explored various machine learning algorithms and assessed different hyperparameters for each algorithm, Figure 6.

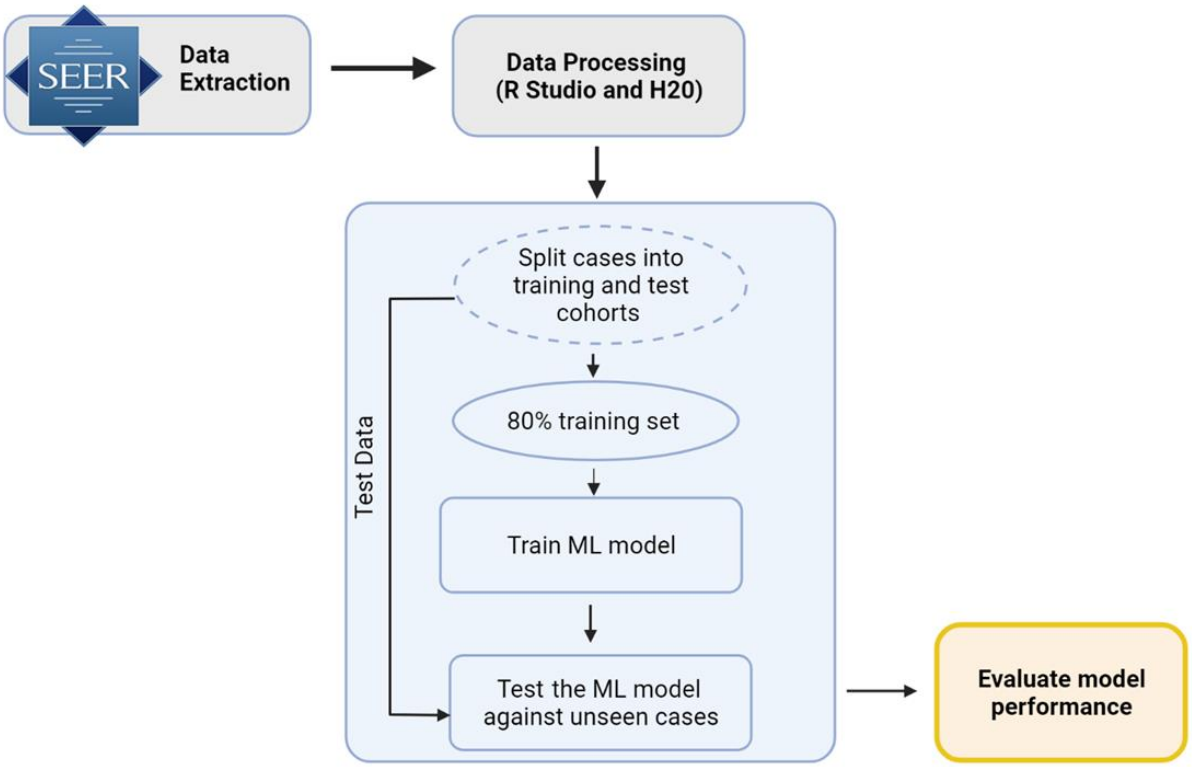

Figure 6: Schematic of data processing and model development. The model development process, data cleaning, and machine learning steps in R studio and H2O.ai tool.

Due to the unbalanced nature of the dataset (i.e., many fewer recurrence cases compared to non-recurrence), two approaches for balancing data were evaluated. The first was oversampling, which involved synthesizing new examples from the existing samples for the minority class. [71] The downside of oversampling is that it introduces a risk of overfitting and/or introducing mathematically valid, yet logically non-sensical sample sets. The second approach was under-sampling, which involved randomly selecting examples from the majority class to remove from the training dataset. In general, under-sampling is the preferred method, particularly for large datasets. [72,73] In this case, the application of the massive SEER data set

helped make utilization of the under-sampling approach a reality, further strengthening the accuracy of the final model.

Using the H2O Automl function, we trained and evaluated various machine learning algorithms, including gradient boosting machine (GBM), distributed random forest (DRF), deep learning, logistic regression, and generalized linear model (GLM). To prevent overfitting during the training phase, we initially assessed the performance of the models using a 5-fold cross-validation technique. The trained models were ranked based on their AUC values, and we selected the top four models for further evaluation on an unseen data split, test split.

The evaluation metrics, including accuracy, precision, recall (sensitivity), and area under the curve (AUC) of the receiver operating characteristic (ROC), were computed for the top four predictive models using a separate 20% test set. The method to compute these hyperparameters is outlined in Equations 1-3.

$$Equation\ 1: Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Equation\ 2: Precision = \frac{TP}{(TP + FP)}$$

$$Equation\ 3: Recall = \frac{TP}{TP + FN}$$

H2O was executed with a 5-fold cross validation and configured for a maximum runtime of 600 seconds. For each ML model, 5 different runs were executed, and the average performances of the top four ML models, were compared.

4.3: Prediction of Local Regional Recurrence in Oral Tongue Squamous Cell Carcinoma

Oral tongue squamous cell carcinoma (OTSCC) is a common head and neck neoplasm that accounts for approximately 1% of new cancer cases diagnosed in the United States each year [50]. Despite advancements in cancer therapeutics and surgical techniques, the worldwide incidence of OTSCC is on the rise, Figure 7. Adequate OTSCC management is still a challenge, with patient 5-year survival rates averaging at about 50%. [68,74–76]

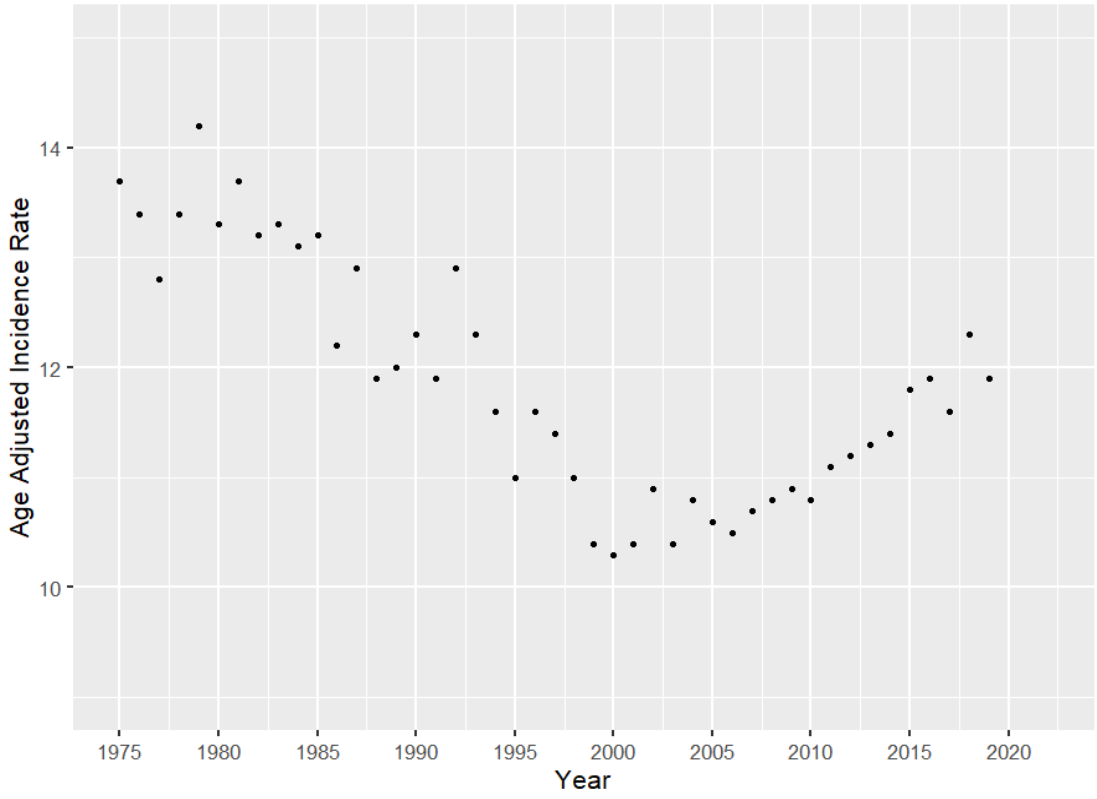

Figure 7: Age adjusted rate of OTSCC between 1975-2019

Although significant progress has been made in cancer diagnostics and treatments, the prognosis of OTSCC is still poor, with many patients experiencing cancer recurrence and surviving less than 10 years after their initial diagnoses.[77,78] With recent studies reporting recurrence rates as high as 37.4%, further investigations aimed at optimizing treatment regimens and post-therapy follow-up will be critical to enhancing patient outcomes.[79,80]

A total of 136,826 cases were extracted from SEER*Stat version 8.3.9 which represented 130,979 unique patients. Two models were trained, one focusing on locoregional recurrence of OTSCC in a 5-year period and the other, a 10-year period. In the 5-year analysis, 14,530 patients met the inclusion criteria, of which 657 suffered from a locoregional recurrence. For the 10-year analysis, 7,100 patients met the inclusion criteria, of which 971 experienced a locoregional recurrence. It is worth noting that only patients alive within the follow-up period (5- or 10-years) were considered in our analyses. Of note, we observed a recurrence rate of ~ 5%, which was lower than the 16-33% recurrence rate that has been previously reported [81–83]. This was due to the stringent exclusion criteria that we applied, which required that patients with certain missing or unknown case information be excluded from analysis.

The database was queried for patients diagnosed with OTSCC using the International Classification of Disease for Oncology, 3rd Edition (ICD-O-3) topography codes for the oral tongue (C01.9-C02.9) and histology/behavior codes for squamous cell carcinoma (SCC; 8010/3, 8020/3, 8021/3, 8070/3, 8071/3, 8072/3, 8073/3, 8074/3, 8082/3). The following demographic and clinical variables of interest were used for training our machine learning models age, sex, race, marital status, year of diagnosis, number of prior tumors, tumor site (e.g., ventral surface of tongue, dorsal surface of tongue, border of tongue), histology, tumor grade, T/N/M stage,

and administered treatments (i.e., surgery, radiation, chemotherapy). To account for variant specific OTSCC behavior and the influence of p16 human papillomavirus (HPV) status, histology was stratified into the following prognostic categories: nonkeratinizing SCC with maturation, undifferentiated nonkeratinizing SCC, differentiated nonkeratinizing SCC, and keratinizing SCC. [84] Furthermore, each case contained a sequence number that provided information on the number of all reportable primary tumors that had occurred over the lifetime of a patient. This variable was used to calculate the "Number of prior tumors", which was defined as the sequence number minus one. All cases with unknown or missing sociodemographic or outcome variables were excluded. Table 1shows a summary of predictors that were used for training the machine learning model. By using simple and commonly acquired prognostic markers as the basis of our models, we enabled our system to be more accessible and easily adoptable by a wide range of practitioners.

Variable	5-Year (N = 14995)	10-Year (N=7342)
	No. (%)	No. (%)
Mean Age, yrs. (SD)	58.4 (11.5)	56.2 (11.5)
Sex		
Male	10,636 (72.0)	4,075 (67.7)
Female	4,129 (28.0)	1,943 (32.3)
Race		
White	13261 (89.8)	5991 (90.1)
Black	706 (4.8)	270 (4.1)
Asian	798 (5.4)	387 (5.8)
Marital Status		
Single	5056 (34.2)	2040 (30.7)
Married	9709 (65.8)	4608 (69.3)
Number of Prior Tumors		
0	14051 (95.2)	6324 (95.1)
1	496 (3.4)	232 (3.5)
2	161 (1.1)	77 (1.2)
3	46 (0.3)	14 (0.2)
4+	11 (0.1)	1 (0.0)
Histology		
Nonkeratinizing SCC with maturation	11468 (77.7)	5276 (79.4)
Undifferentiated nonkeratinizing SCC	86 (0.6)	39 (1.0)
Differentiated nonkeratinizing SCC	824 (5.6)	288 (4.3)
Keratinizing SCC	2286 (15.5)	993 (15.0)
SCC NOS	101 (0.7)	52 (1.0)
Tumor Grade		
Well-differentiated	2262 (18.8)	1067 (19.7)
Moderately differentiated	5752 (47.8)	2585 (47.6)
Poorly differentiated	3896 (32.4)	1710 (31.5)
Undifferentiated	117 (1.0)	64 (1.2)
T-Stage		
T1	4443 (46.7)	1594 (50.0)
T2	3274 (34.4)	1109 (34.8)
T3	1013 (10.6)	262 (8.2)
T4	784 (8.2)	221 (6.9)
N-Stage		
N0	5110 (45.5)	1918 (48.3)
N1	1968 (17.5)	764 (19.2)
N2	3847 (34.3)	1187 (29.9)
N3	296 (2.6)	102 (2.6)
M-Stage		
M0	11200 (99.3)	3913 (99.2)
M1	75 (0.7)	30 (0.8)
Surgery		
Yes	6125 (41.8)	2506 (37.5)
No	8519 (58.2)	4185 (62.5)
Radiation		
Yes	8965 (60.7)	3811 (57.3)
No	5800 (39.3)	2837 (42.7)
Chemotherapy		
Yes	6598 (44.7)	2632 (39.6)
No	8167 (55.3)	4016 (60.4)

SCC: Squamous Cell Carcinoma; NOS: Not Otherwise Specified
Values are based on the number of cases.

Table 1:Summary of the sociodemographic and clinical predictors used in developing the ML models for predicting OTSCC recurrence.

To identify the best predictive model, the AUC of the ROC curve was used as a metric to compare the performance of four trained machine learning algorithms, Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), and deep learning (artificial neural network), **Error! Reference source not found.** on test split. These models were ranked at the top due to their superior performance on 5-fold cross-validation during the training process in the H2O AutoML function compared with other trained algorithms.

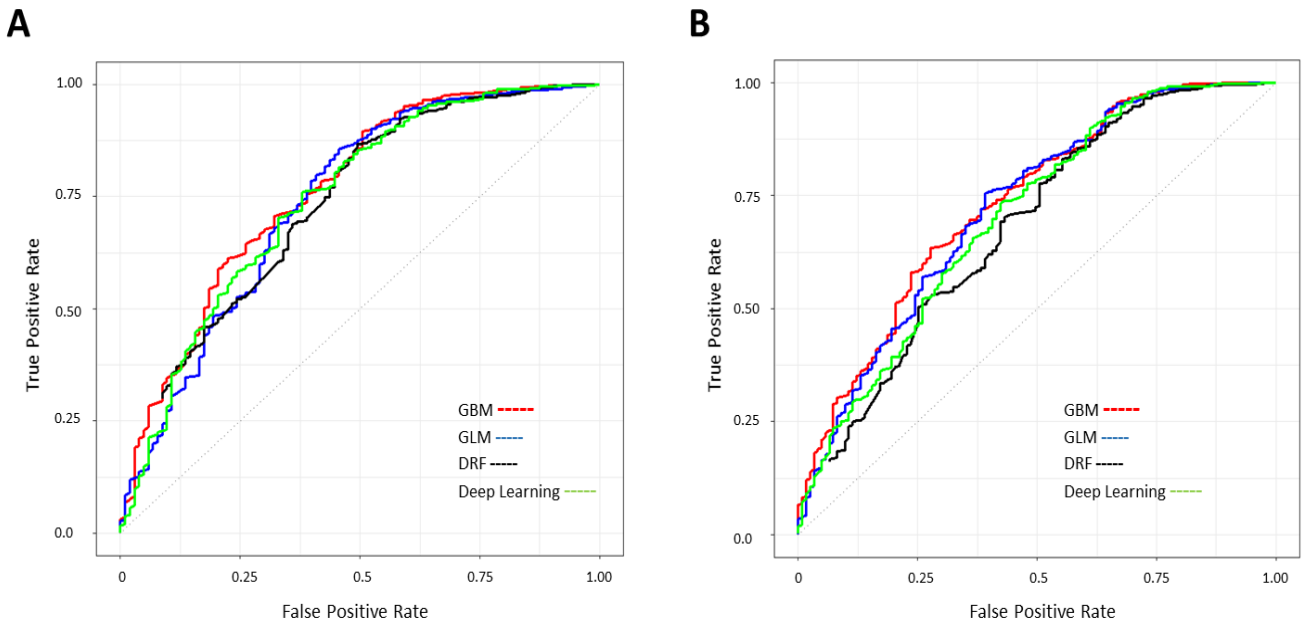

Figure 8: ROC plots of four developed ML models. Performance of Gradient Boosting Machine (GBM),Generalized Linear Model (GLM) , Distributed Random Forest (DRF), Deep learning.

Prediction Window	Classification Model	AUC (SD)	Accuracy % (95% CI)	Recall % (SD)	Precision % (SD)
5 Years	GBM	0.75 (0.01)	81.8 (79.7-83.9)	83.0 (0.02)	97.7 (0.002)
	GLM	0.73 (0.02)	77.4 ((74.5-80.2)	78.1 (0.03)	98.0 (0.002)
	DRF	0.73 (0.03)	72.8 (69.8-75.7))	73.3 (0.02)	97.8 (0.003)
	Deep Learning	0.70 (0.04)	82.1 (74.7-89.6)	83.5 (0.06)	97.6 (0.002)
10 Years	GBM	0.74 (0.02)	80.0 [75.3, 84.1]	82.8 (0.04)	94.0 (0.004)
	GLM	0.73 (0.02)	78.4 [74.2, 82.7]	81.0 (0.04)	94.3 (0.002)
	Deep Learning	0.71 (0.02)	74.4 [70.1, 78.8]	76.6 (0.04)	94.0 (0.002)
	DRF	0.69 (0.01)	70.6 [68.0, 73.3]	72.2 (0.02)	93.8 (0.004)

AUC: Area Under Curve; GBM: Gradient Boosting Machine; GLM: Generalized Linear Model; DRF: Distributed Random Forest
Performance metrics were reported as the average of 5 runs.

Table 2: Performance metrics of the top 4 machine learning models for predicting 5- and 10-year cancer recurrence. The GBM model exhibited the highest AUC and accuracy for both prediction windows.

The performance metrices of top four ML model are shown and compared in Table 2. GBM classification model with AUC of 0.75 (0.01) and 0.74 (0.02) outperformed all other models for both 5-year prediction and 10-year prediction respectively. Of note, the accuracy, recall, and precision of the model can be calculated at different thresholds within the graph of the ROC curve. Thus, the optimum threshold for each model can vary depending on the definition and application of the classification problem. For example, a screening tool may require high recall and precision. For this specific application, we focused on using the model as a screening tool and, therefore, aimed to increase recall without major sacrifice of accuracy however this strategy may vary based on the application of the model. Therefore, the best overall performance for predicting OTSCC recurrence was achieved by the GBM model with 81.8%

accuracy, 83.0% recall, and 97.7 % precision for 5-year prediction, and 80.0% accuracy, 82.8% recall, and 94.0% accuracy for 10-year prediction.

In addition to performance metrics of the model, we were also interested in assessing the impact of each individual feature on the predictive outcome. To achieve this, we applied two different methods to the best predictive model, GBM, which is a tree-based model. The first method involved calculating the GBM model's variable importance, which measures and normalizes the relative influence of each feature. The values are presented in Figure 9. This importance is determined by evaluating whether a variable was selected for splitting during the

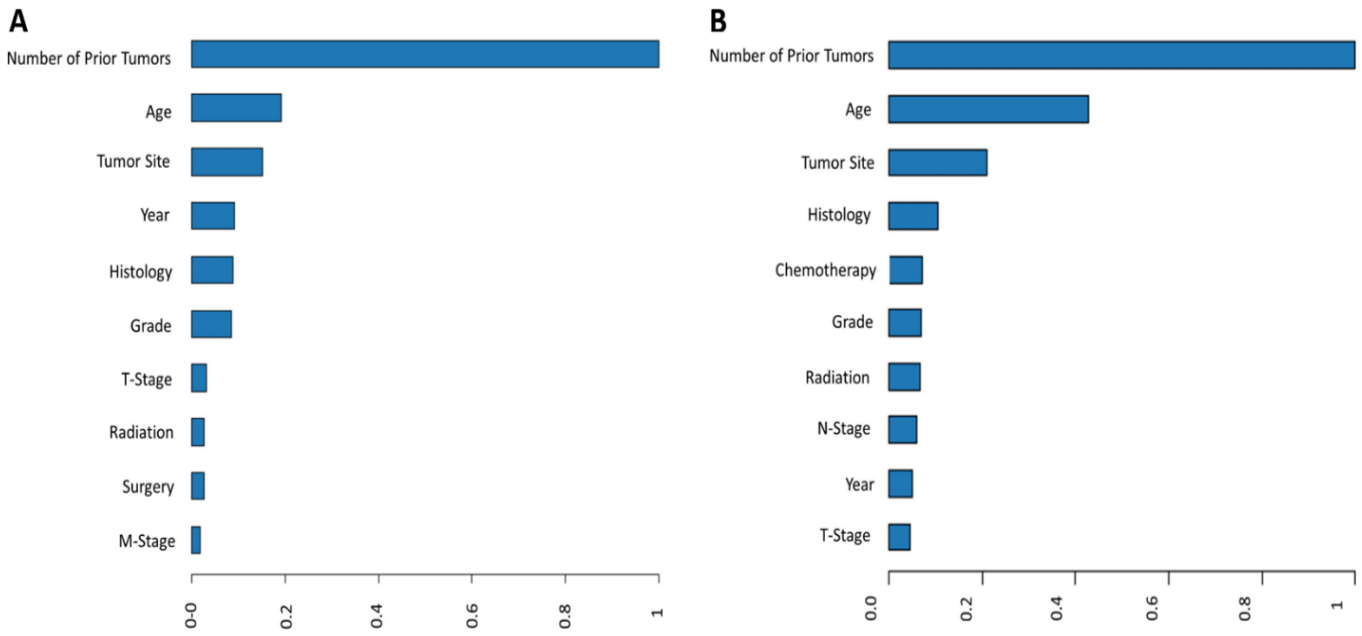

Figure 9: The relative influence of each feature on model's predictability for a) 5-year prediction and b) 10-year prediction of OTSCC recurrence.

tree-building process and how much the squared error improved or decreased as a result over all trees. [45]

In the second method, the Shapley Additive exPlanations contribution plot (SHAP) was utilized to illustrate how the GBM model arrived at its results, Figure 10. This method ranked the importance of each feature in the GBM model based on all the possible pairs of coalitions between predictors of the model, with higher importance scores indicating a higher contribution to the model's predictive ability (from top to bottom).

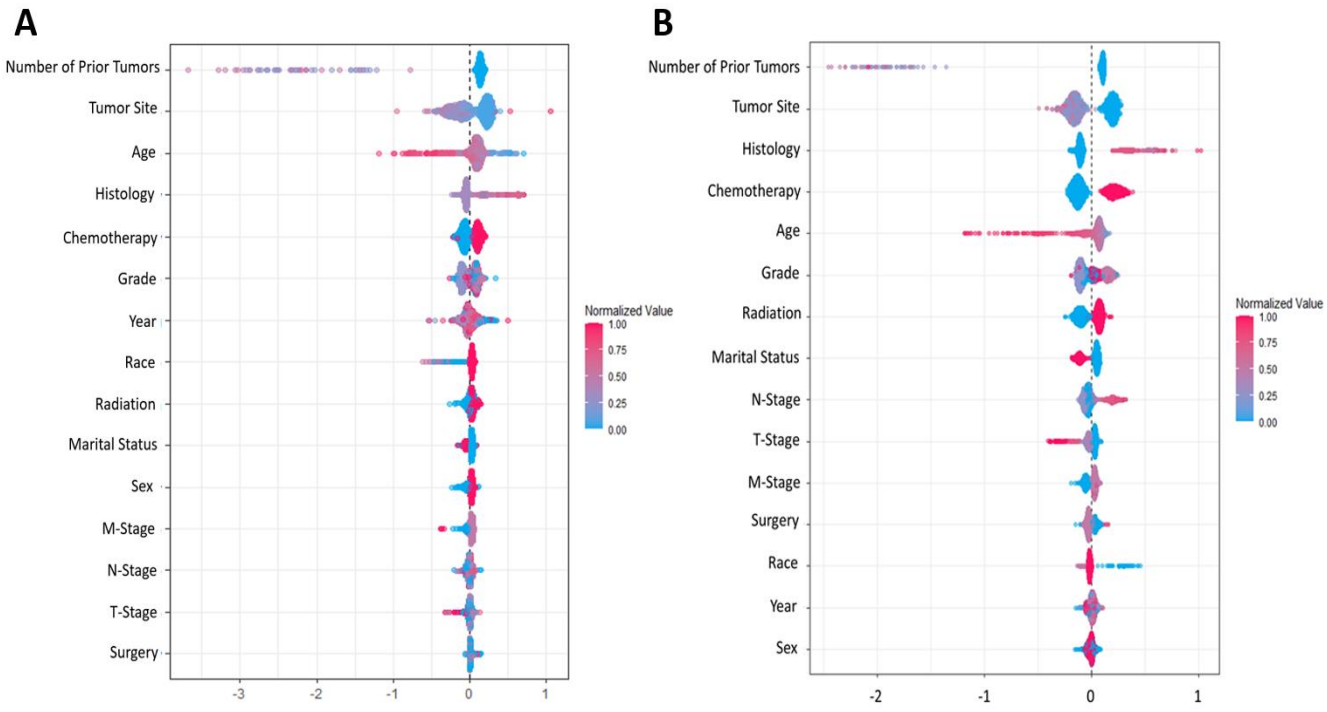

Figure 10: The Shapley Additive exPlanations contribution plots (SHAP) for the GBM model. SHAP plots of (a) 5-year and (b) 10-year prediction models. All pairs of coalitions between features of ML model were calculated and feature's importance were ranked from top to bottom.

As shown by both approaches, the number of prior tumors, age, histology, chemotherapy and tumor site were the most important factors in determining the probability of locoregional recurrence of OTSCC.

In the first aim of this chapter, we used our novel framework for identifying cases of OTSCC recurrence from the SEER dataset. We demonstrated the utility of this framework by developing ML models that predicted 5- and 10-year cancer recurrence with high accuracy and precision using a large population-based cohort of OTSCC patients. Specifically, of the four ML algorithms that we employed, the GBM-based model showed the most promise, demonstrating accuracies of 82% and 80% for 5-year and 10-year recurrence, respectively. Although we observed a low recurrence rate, we do not anticipate this lower prevalence to have influenced our findings. Unlike traditional regression techniques that compute likelihood or risk scores based on a sample's observed event rate, our machine learning model was trained using an under-sampling approach on the majority class (non-recurrence) to be tolerant of deviations from the true population prevalence rates.

By developing a predictive screening tool, treatment teams can be better informed of a patient's risk for cancer recurrence and modify their management strategy accordingly. Additionally, the mortality rate in recurrent cases of OTSCC is highly dependent on the time of diagnosis, with early detection of recurrence being associated with reduced mortality.[85,86] By using our highly representative and sensitive classification models, clinicians can be better informed of which patients are at higher risk for OTSCC recurrence and cater their management and follow-up to ensure timely diagnosis if a recurrence were to occur.

In our analysis, we used SHAP and feature's importance calculation to explain the predictions made by the Gradient Boosting model and interpret the tangled nonlinear relationships between features and local regional recurrence of OTSCC. Consequently, we found that the number of prior tumors, patient age, tumor site, chemotherapy, tumor histology

and tumor grade were consistently the most influential features in predicting cancer recurrence. Thus, by developing an artificial intelligence (AI) model in the context of a highly representative population for cancer recurrence and analyzing the nonlinear effect of features by the SHAP method, we found some of the features to be more prognostic compared to those that have been traditionally considered major prognostic factors in oral tongue cancer recurrence, such as lymphatic invasion or T-stage. [82,83] Importantly, these findings do not discount the prognostic importance of previously reported clinical factors, but rather highlights certain factors that may be generally considered highly prognostic across a more diverse and heterogeneous patient population.

In a recent institutional study, Alabi et al. similarly demonstrated success in predicting locoregional recurrence in OTSCC. [58]However, despite their impressive results, their models were trained using only 217 cases of early-stage OTSCC, which largely limited their system's applicability to more advanced tumors and its external validity against the general population, where the spectrum of disease behavior and progression is much more diverse than what is experienced at a single institution. Interestingly, the authors found that certain specialized histopathological parameters, such as lymphocyte host response, pattern of invasion, depth of invasion, and perineural invasion, were particularly important features in their prediction models. Owing to the limitations of the SEER database, our models were trained without using these clinical features. While the lack of dependence on these specialized histopathological parameters expanded the accessibility of our system to a broader range of clinical facilities where such information may not be readily available, consideration of these features may be

warranted in future generations of ML models where higher prediction accuracy in lieu of increased accessibility is desired.

Earlier research has highlighted the importance of genetic predisposition in head and neck squamous cell carcinoma (HNSCC).[87,88] Moreover, genetic and environmental factors, including a history of prior head and neck cancer, have been shown to be associated with recurrence of HNSCC.[89,90] The influence of patient age on prognosis has also been previously established. In a large retrospective study of OTSCC patients, Mukdad et al. demonstrated that older patients were associated with more advanced disease and worse survival.[68] It was hypothesized that this worse prognosis was partly due to a tendency for clinicians to more aggressively treated younger patients with multimodality therapy. Interestingly, younger patients were also observed to less frequently present with metastatic lymph nodes. Indeed, survival and recurrence rates have been reported to be largely influenced by the presence of nodal disease.[91]As such, cancer recurrence at a regional site can be suggestive of more aggressive disease with tendency to recur following treatment. In a cohort study, Wolfer et al. suggested that aggressive neoplastic behavior is strongly dictated by tumor histology.[92] Specifically, the degree of keratinization in oral squamous cell carcinoma was demonstrated to be an important prognostic factor for recurrence and survival. Other recent studies have reached similar conclusions and have even created recurrence risk models on the sole basis of histological parameters.[93–96]

To our knowledge, this is one of the first studies to develop an algorithm to identify cases of cancer recurrence from the expansive and widely used SEER database, laying a basis for future investigations across a variety of medical fields. Through the use of this novel

framework, we also presented one of the first machine learning-based classification models that accurately predicted 5- and 10-year recurrence in OTSCC patients using only commonly available demographic and clinical features.

There are, however, limitations to this study worth mentioning. Since patients were extracted from a de-identified national database, the data may be susceptible to information bias. Additionally, despite including a number of sociodemographic and clinical variables in our models, we would like to point out that certain potentially valuable histopathological (e.g., lymphocyte host response, perineural invasion, depth of invasion, tumor budding and worst pattern of invasion) and clinical features (e.g., timing of treatments, radiation dose, HPV status, neck dissection) were not accounted for due to the limitations of the SEER database . Despite these constraints, we were able to develop a model with high predictability for locoregional recurrence of OTSCC. We believe that incorporating these site-specific variables along with other clinical and sociodemographic variables can only enhance the predictive power of these models.[58,80,83,89,97] We hope that this investigation will encourage inclusion of such variables in future updates to SEER and other large-scale clinical datasets. Furthermore, in the next chapter of this study, we investigated the impact of genetic variations in individuals in determining the risk of cancer recurrence.

CHAPTER 5: INVESTIGATING THE IMPACT OF INHERITANCE FACTORS ON CANCER RECURRENCE RISK

In the first aim of this investigation, we developed a novel framework for identifying cases of cancer recurrence from the SEER database with which a generalizable and highly representative machine learning model could be generated. By developing a predictive risk estimate tool, treatment teams can be better informed of a patient's risk for cancer recurrence and modify their management strategy accordingly. In the analysis of our models' feature scores and SHAP contribution, we found the number of prior tumors to consistently be the most influential features in predicting cancer recurrence. Previous studies have emphasized on the significance of genetic predisposition in both head and neck squamous cell carcinoma (HNSCC) and breast cancer. [87,88,98]Additionally, genetic, and environmental factors, including a history of prior tumors, have been shown to be associated with different types of cancer recurrence. [89,90,98,99]

For the next specific aim, we are going to include the effect of genetic and inheritance factors in our model development and investigate developing a machine learning risk estimate model for cancer recurrence based on the inheritance factors.

Next-generation sequencing (NGS) has revolutionized the field of genomics by enabling the sequencing of millions of fragments in a massively parallel fashion, leading to improved speed, accuracy, and reduced sequencing costs. This technological advancement has laid the groundwork for a new era of genomic studies. Among the remarkable findings facilitated by NGS technologies is the discovery of extensive genomic structural variations, known as copy

number variations (CNVs). These structural variations involve alterations in the dosage of genomic segments, ranging in size from one kilobase pairs (Kbp) to mega base pairs (Mbp), as compared to a reference human genome. These variations result from deletions, duplications, triplications, insertions, translocations, and inversions of chromosome segments. Such insights into the genomic landscape hold significant promise for advancing our understanding of genetic complexities and their implications in various diseases, including cancer. [100–102]

These structural variations in the genome can have diverse effects depending on their size and location. CNVs can disrupt gene function and lead to complex diseases through different mechanisms. For example, deletions or disruptions within genes or intragenic CNV interactions can result in haploinsufficiency, where the remaining functional copy of a gene is unable to produce enough gene product to maintain normal function. [103] This highlights the importance of considering CNVs in understanding the genetic basis of diseases and their potential impact on gene function and disease development.[104]

CNVs in germ line DNA have been implicated in cancer predisposition through various mechanisms. These structural variations can lead to the formation of fusion genes, disrupt gene regulation, impair DNA repair mechanisms, and affect tumor suppressor genes. However, the precise mechanisms by which CNVs influence cancer development are still being investigated and require further study.[17,18,105] It is worth noting that CNVs are present in all individuals, but research on CNVs in cancer has predominantly focused on rare single region CNVs. Similarly, studies exploring germline CNVs have often concentrated on rare single region CNVs and have identified individual genes associated with specific cancers. Few studies have specifically examined the interactions of intragenic CNVs and the inclusion of structural variations across

chromosome scales in their model development. [104,106] Therefore, there is a need for more comprehensive investigations to better understand the role of CNVs in cancer and their potential implications for disease risk assessment and management.

Indeed, there is a significant distinction between germline CNVs and somatic tumor CNVs. Germline CNVs refer to the structural variations present in an individual's inherited DNA, which can be passed down to offspring. [105]These CNVs are less influenced by environmental factors and provide a foundation for studying inherited risk factors for diseases. On the other hand, somatic CNVs arise in the DNA of somatic cells during an individual's lifetime and are associated with the development of tumors. These somatic CNVs are more influenced by environmental factors and are not inherited or passed on to future generations. [107]To ensure the focus on germline inheritance factors and minimize the impact of environmental factors, this study specifically examines and analyzes CNVs derived from germline DNA.

The majority of existing genetic scores for cancer rely on single nucleotide polymorphisms (SNPs), which are specific variations in a single nucleotide at a particular location in the genome. These polygenic risk scores calculate a score by combining the values of multiple SNPs in a linear manner. [108] While this approach can be effective for diseases caused by single mutations, it may not adequately capture diseases that arise from combinations of mutations or involve complex interactions between multiple genetic factors. In particular, diseases influenced by combinations of mutations, such as the presence of three out of five specific mutations, or diseases affected by epistatic interactions, where the effect of one gene depends on the presence of other genes, cannot be adequately studied using the traditional

SNP-based analysis. [63–65,109] Due to the above challenges the predictability of polygenic risk scores developed solely based on SNPs may be limited. [4]

Machine learning algorithms offer the advantage of incorporating nonlinear combinations of single nucleotide polymorphisms (SNPs) in model development. However, the vast number of SNPs across the human genome, which exceeds 5 million, presents challenges in applying non-linear machine learning algorithms due to the high dimensionality of the data. [110,111] The number of SNPs typically far exceeds the number of available patient data points, which can limit the effectiveness of traditional machine learning approaches.

To overcome the challenges associated with the limited predictability of polygenic risk scores based solely on SNPs, we propose an alternative strategy that incorporates structural variations as predictors in our risk models. By including structural variations, such as CNVs we can capture a broader range of genetic variations that may contribute to disease risk.

Recent advancements in next-generation sequencing (NGS) have opened up new possibilities for researchers, offering a wealth of information on various types of structural variations and their impact on genomic architecture. [112] One valuable measurement that can be achieved with these advancements is copy number variation analysis, which involves quantifying changes in the number of copies of specific genomic regions compared to a reference sample. CNV analysis involves two common metrics: CNV ratio and log R ratio. The log R ratio is a logarithmic transformation of the CNV ratio, representing the relative change in copy number for a specific genomic region (locus) in the sample being analyzed compared to a reference sample. This value is calculated as the log base 2 of the CNV ratio, where a positive log R ratio indicates amplification of the genomic region, and a negative log R ratio indicates

copy number loss or deletion. [113] The logarithmic scale provides a more sensitive measurement of the intensity difference compared to a linear scale. To detect these genetic variants, genome-wide genotyping arrays are commonly employed, offering an efficient method for large-scale analysis. [113,114]

5.1: Research Design and Methods

In our approach, we will employ machine learning algorithms to analyze genetic data obtained from DNA microarray measurements, allowing us to capture intricate genetic interactions. By reducing the dimensionality of the problem and taking genome structural variations into account, our goal is to overcome the limitations associated with traditional polygenic risk scores. This comprehensive strategy aims to provide a more accurate and robust assessment of cancer risk.

This comprehensive approach, which includes the integration of structural variations and the utilization of advanced machine learning techniques, has the potential to improve the predictability and accuracy of genetic risk models for complex diseases. It allows us to capture a more complete picture of the genetic architecture underlying disease susceptibility and provide more reliable risk assessments for individuals. As part of our second aim in this investigation, we will utilize available genetic data from the UK Biobank to develop a machine learning model for predicting breast cancer recurrence. While current polygenic risk scores for cancer have mainly focused on identifying individuals at higher risk of developing specific cancers, there has been limited research on developing models to detect the risk of cancer recurrence. Our study

aims to fill this gap and contribute to a more personalized and precise approach in predicting cancer recurrence risk.

5.1.a: UK Biobank, Axiom Microarray, Chromosomal Scale Length Variation Calculation

The UK Biobank is a large-scale population-based study that aims to improve the prevention, diagnosis, and treatment of various diseases. It involves the collection of extensive health-related data, including genetic information, from over 500,000 participants in the United Kingdom. Participants in the UK biobank, recruited at ages 40-69 and were registered with the National Health Service (NHS). [115] It comprises approximately 488,000 individuals with genetic data, with a total of 764,257 CNV values across the 22 autosomes, as well as 18,857 CNV values for the X chromosome and 691 CNV values for the Y chromosome.[115,116] The entire cohort's genotype is being determined using DNA microarrays containing selected genetic variants, including single nucleotide polymorphisms, insertions, deletions, and copy number variants. The UK Biobank utilizes the Axiom Array to probe specific locations of interest for copy number variations (CNVs). These files encompass various information such as normal SNP genotyping data, calls, confidences, intensities, and other relevant data related to the study of genetic variations. Our access to this information is obtained through a renewed application process every three years, and it is provided only to approved researchers through the data showcase.[117] Researchers can download the data through the UK Biobank's Data Showcase, which collaborates closely with the European Genome Archive (EGA). The dataset compromises genotype data for a total of 488,377 participants. We need access to l2r files in order to derive the relative genetic information related to chromosome structural variations.

Each patient in the UK biobank study is assigned a unique patient ID. The patient IDs file has been organized as a list, with one ID per row for each patient. The order of these IDs is crucial as they correspond to the column headings of individual microarray data, which contains approximately 850,000 genetic variants. Each chromosome file is a plain text file without headers, and the data is structured with one column per patient where each row represents the log2 intensity ratio measured at a different SNP location in their array. All the l2r genetic data files for each chromosome along with patient ID file were downloaded by the previous graduate student Christopher Toh in our lab and are stored in our server. [118]

To incorporate structural variations in the genetic score model development, we adopted an unique approach. This involved analyzing CNVs within each segment of the chromosome. To facilitate this analysis, a chromosomal scale length variation (CSLV) dataset was computed by averaging log_2^R across each chromosome. The average log_2^R can be calculated based on the desired number of even splits within the values of each chromosome or within each individual segment of chromosome. This approach enables a comprehensive assessment of structural variations and their potential impact on the genetic score model.

Figure 11 shows the histogram of the relative length of chromosome 1 for 10,000 randomly selected individuals from the UK Biobank. The histogram illustrates that the majority of people have a nominal value of 0 for chromosome 1. However, approximately 200 participants had longer chromosome 1 lengths (around 7% longer) compared to the nominal average. This observation indicates the presence of chromosomal scale length structural variations within the population.

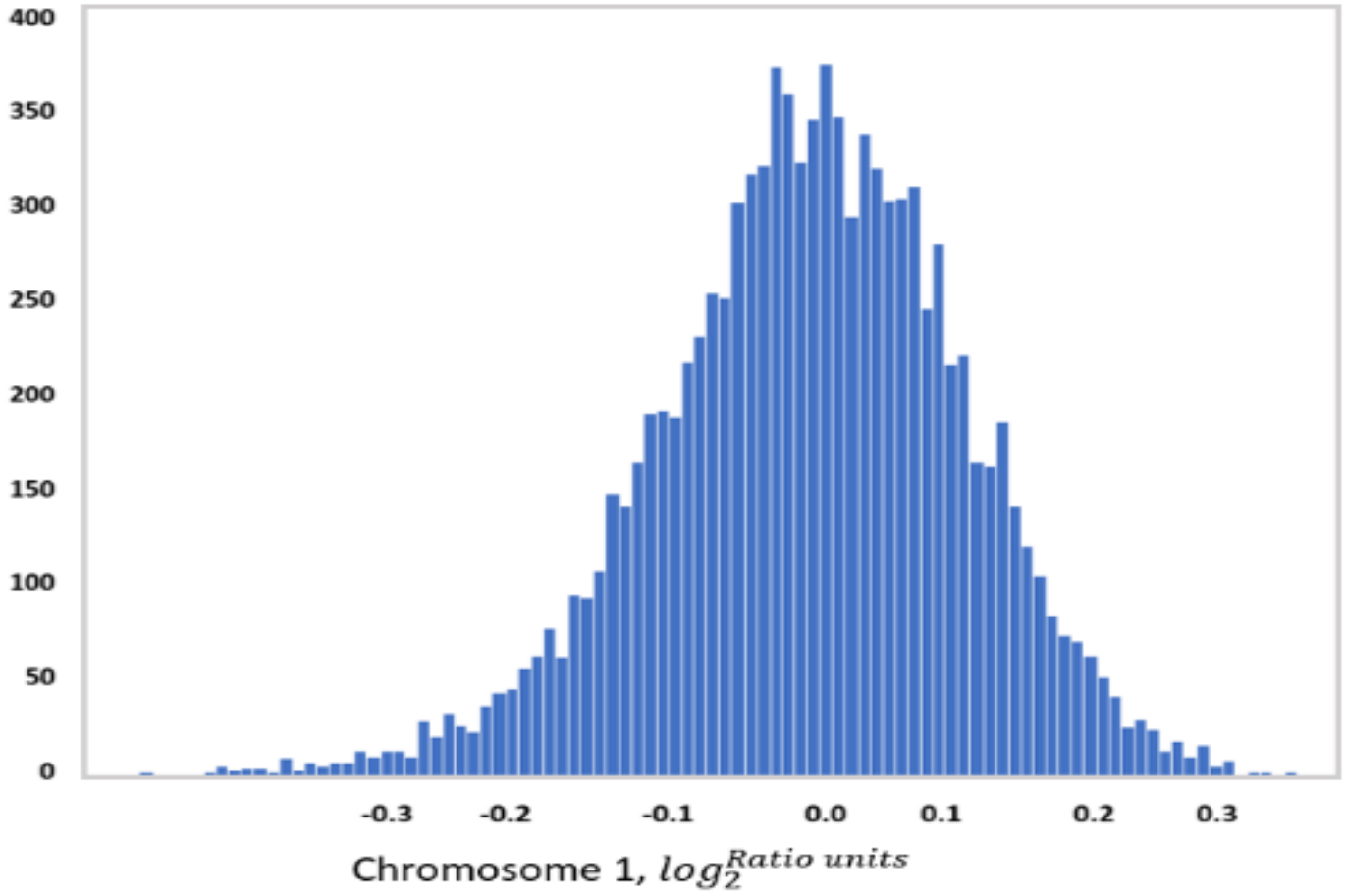

Figure 11: Histogram of the "relative length of Chromosome 1" for 10,000 randomly selected individuals from the UK Biobank dataset.

5.1.b:Data Cleaning and ML Model Development

To test our hypothesis and process the data, we needed to analyze another encrypted file containing information about each patient, such as sex, age, cancer incidences throughout the study period, time of diagnosis of each incidence, cancer type, etc. This information could be linked to the calculated CSLV dataset by using the patient ID. [118] Since our focus was on developing a risk estimate model for cancer recurrence based on the available genetic data in UK biobank, we wrote an algorithm in C# to identify patients who have had cancer recurrence. For that purpose, we used available information on cancer incidences for each patient and identify whether that patient has had a cancer recurrence throughout the study period. The algorithm was executed for all primary tumor sites in the UK Biobank, and the number of patients with recurrence for each tumor site was calculated and plotted, Figure 12.

Based on the data collected from the UK Biobank and analyzing all 488,000 patients with

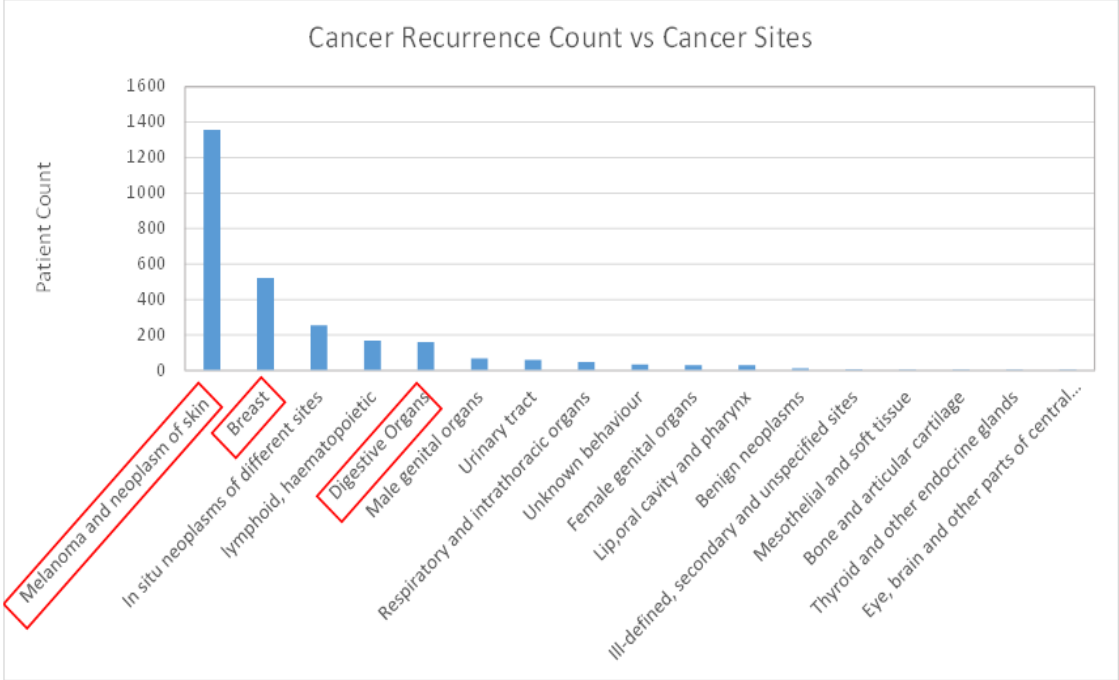

Figure 12: Number of patients who have had cancer recurrence at each cancer site within 10 years

genetic data using our developed platform in C#, we identified the number of target patients who had cancer recurrence within the same primary site group. Among them, melanoma of the skin, breast cancer, and digestive organs showed a higher number of patients with cancer recurrence throughout the study period.

Based on the available recurrence data within each primary site for this specific aim, we focused on developing a genetic risk estimate model for breast cancer recurrence. To achieve this, we divided the study population into two classes: patients who have experienced breast cancer recurrence and an age-matched group of patients who were diagnosed with breast cancer only once throughout the UK Biobank study.

During classification ML model development, for each run, we linked the two classes of patients those with breast cancer recurrence and the age-matched group with a single breast cancer diagnosis, based on their unique patient IDs to the calculated chromosomal scale length variation (CSLV) dataset. The CSLV dataset consisted of average values of copy number variations (CNVs) within large segments of each chromosome, which served as features for the ML model. This approach allowed us to incorporate structural variations from the genetic data into the model, and evaluate their effect solely on determining the risk of individual for developing a breast cancer recurrence.

For the current aim, we utilized the H2O.ai tool and its corresponding R package for developing a machine learning model. The clean dataset, consisting of features and labels, was passed to the H2O automl function in each run, with specific criteria set for model development. In each run, we applied 5-fold cross-validation as an additional method for model evaluation. The model was trained using 80% of the data, while the remaining 20% split of the

data was used for testing the trained model. Additionally, we set the model development time to 900 seconds and included all available machine learning algorithms in the process. We also kept all the cross-fold predictions for further analysis and comparison. This approach allowed us to thoroughly evaluate the performance of the machine learning model and assess its predictive capabilities in predicting cancer recurrence. By using cross-validation and test splits, we obtained multiple metrics to analyze the model's generalization and performance on unseen data.

By analyzing the performance of the model through calculating AUC and plotting ROC curve on test split of the data we can assess the model's ability to distinguish between cancer recurrence cases and non-recurrence cases. Furthermore, we compared the model's performance on the completely shuffled dataset to evaluate its robustness and generalization capabilities. This step helped us ensure that the model's predictions were not biased or overfit to the specific dataset used during training.

Overall, these analyses provided us with a comprehensive understanding of the developed model's strengths and limitations, paving the way for further improvements and applications in cancer recurrence prediction.

5.2: Results

5.2.a: Predicting Malignant Breast Tumor Recurrence using Germline Chromosomal Scale Length Variation

In the UK Biobank, there are a total of 502,536 patients. Through our analysis, we identified 543 different types of neoplasms associated with these patients. Among them, 12,441 patients experienced more than one diagnosis of neoplasm throughout the study. The

latest version of the ICD-10 classification system was employed to categorize neoplasms into various subgroups. [119]This grouping method was utilized to organize each specific neoplasm into a larger subgroup, making it easier to sort and analyze the data for identifying the recurrence cases within each subgroup, Figure 12.

To define recurrence, we identified patients who had more than one cancer incidence within the same group of neoplasm, with a time period of more than 180 days between the two cancer incidences. In particular, patients with malignant neoplasm of the breast, coded as C-50.0 to C50.9 according to the ICD-10 classification, were grouped together. [119]Among them, those who received a second diagnosis within the same group were identified as recurrence cases of breast cancer.

After identifying the target patients, we linked them to the calculated CSLV data table using their unique patient IDs. The finalized dataset for developing a machine learning (ML) model to predict malignant breast cancer recurrence consisted of two main classes, and each case was represented with 22 numbers, each corresponding to the average log2R value within each chromosome.

The negative class consisted of women who had a single diagnosis of malignant breast tumor. In contrast, the positive class included women from the UK Biobank study who had experienced breast tumor recurrence. These women were diagnosed with malignant tumor of the breast more than once within the study period, with a time gap between the diagnoses of more than 180 days. Specifically, the positive class comprised 489 women who had received multiple diagnoses of malignant breast tumor and negative class compromised 13478 cases.

Error! Reference source not found. represents the racial and sex distribution along with average age of women in both groups.

Classes (Recurrence Vs Non-Recurrence)	**Race distribution**			Sex	Age-Average (SD)	**Total**
	White	White	Other	Female		
Positive (Patients with Malignant Breast Cancer Recurrence)	489	489	-	489	60 (6.8)	489
Negative (Patients with Malignant Breast Cancer- No Recurrence)	13478	13260	218	13478	58.7 (7.2)	13478

Table 3: Race, sex, and age distribution of two classes that were used for developing ML model to predict breast cancer recurrence.

chromosome 1, while Figure 14-a displays the distribution in the negative class. The histograms for non-recurrence cases and recurrence cases both indicate that most people in both classes have a nominal value of 0. However, the distribution in the positive class (recurrence cases) appears to be more widespread. One possible explanation for this observation could be related to the relatively smaller number of cases in the positive class compared to the negative class.

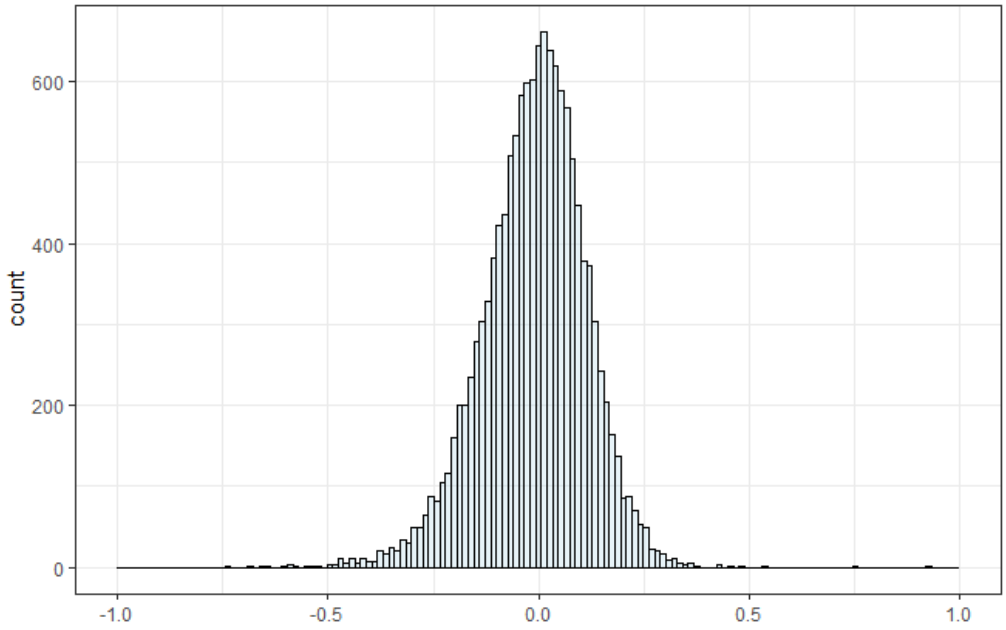

Figure 14-a :Histogram of the "relative length of Chromosome 1" for 13478 patients who have been diagnosed with breast cancer once through the UK Biobank study period (non-recurrence).

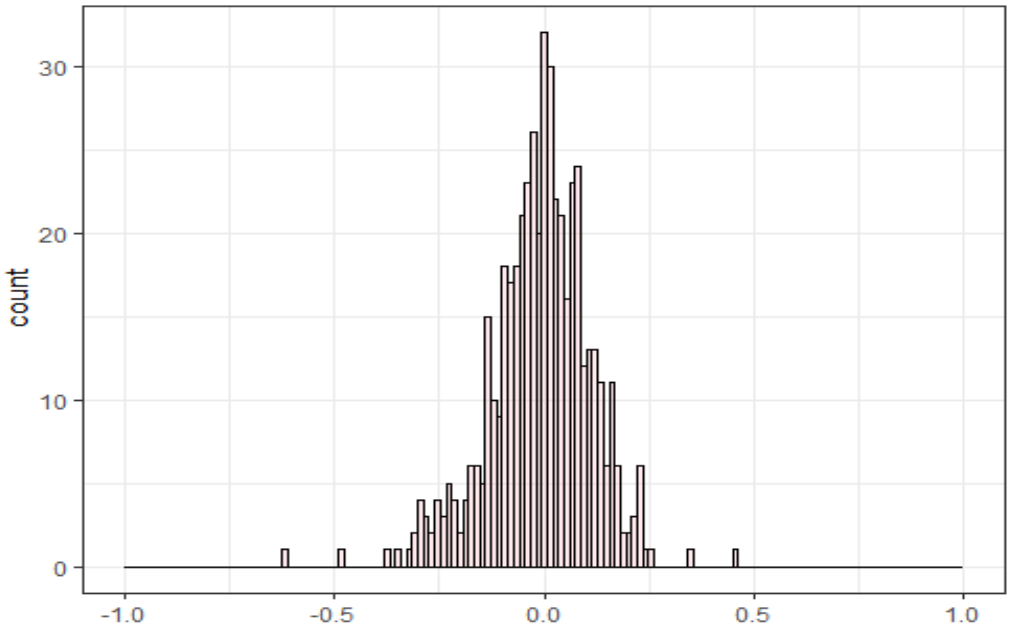

Figure 14-b: Histogram of the "relative length of Chromosome 1" for 489 patients who have had breast cancer recurrence through the UK Biobank study period (positive class).

Figure 15 presents a comprehensive comparison between the two classes using whisker-box plots of calculated chromosomal scale length variations (CSLV) for all the chromosomes. This plot allows us to visually compare the distribution of CSLV values between the positive (recurrence cases) and negative (non-recurrence cases) classes across all chromosomes.

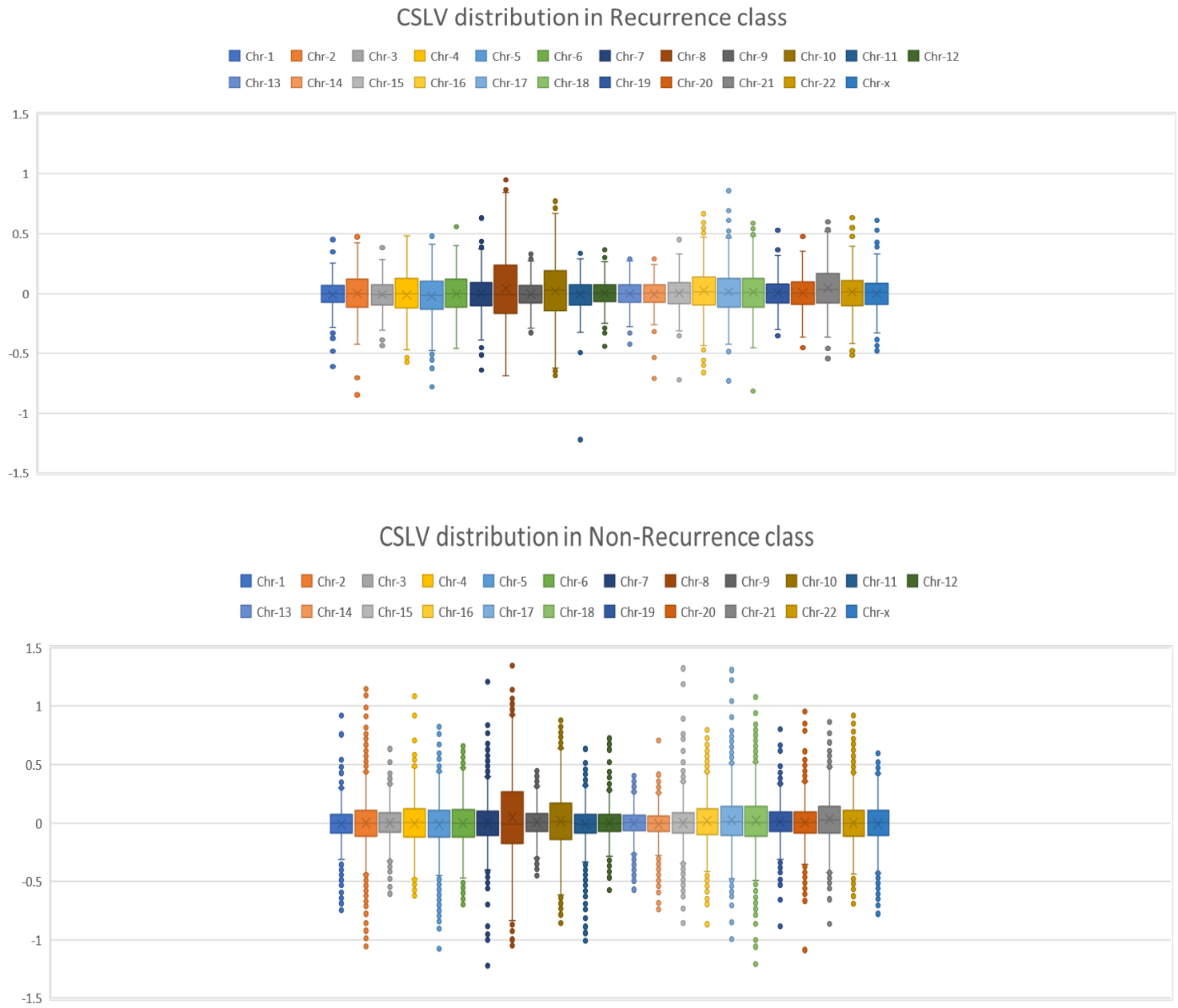

Figure 15: Box plot of CSLV distribution for each chromosome in the positive class (patients who have had breast cancer recurrence in the Biobank study) vs. the negative class (patients who have not had breast cancer recurrence).

Based on Figure 15, the median of the calculated chromosomal scale length variations (CSLV) for both classes was 0. However, it is noticeable that the number of outliers in the non-recurrence class was higher compared to the recurrence class. This could be attributed to the smaller number of cases in the recurrence class or potential structural predispositions in the two groups, leading to differences in CSLV values. Further investigation is needed to understand the underlying factors contributing to these variations by increasing the number of datapoints in the positive class.

These calculated CSLVs were used as feature for model development. During the ML model development process, for each run, an under-sampled age-matched group of participants was selected from the negative class. This under-sampled group was then appended to the positive class, creating a balanced dataset with a rough ratio of 60:40. The balanced dataset, was used as input for the H2O AutoML function, which was implemented using R programming language. The main objective of this step was to ensure that the classes were balanced, thereby preventing any bias towards one class and improving the model's performance. The finalized dataset was divided into 80% training data and 20% test data. Gradient Boosting machine was the algorithm that consistently ranked the highest on the leader board for each run. The performance of the trained model was evaluated by plotting the ROC curve on the test split of the data as shown in Figure 16.

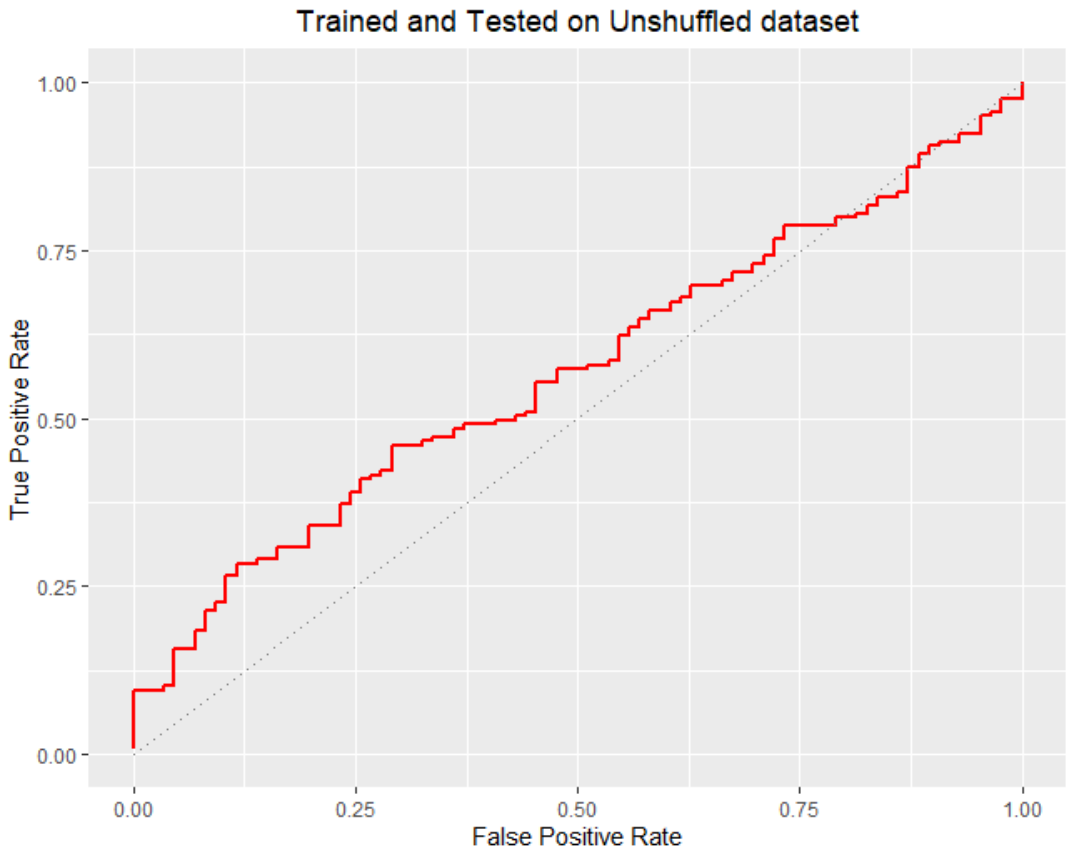

Figure 16: ROC curve of the trained GBM model tested on the unseen dataset with an AUC of 0.56.

The AUC of the top trained model was 0.56 indicating a slightly better performance than a completely random prediction, which would have an AUC of 0.50.

To improve the prediction of the model, we further investigated the number of splits within each chromosome. By increasing the number of splits and capturing more variations with finer subdivisions within each region of the chromosome, we aimed to enhance the model's predictability. In this exploration, we characterized each case in the model with 88 numbers, where each number corresponds to a quadrant of a chromosome.

Figure 17 displays the ROC curve of the top trained model, GBM, on the test split of the data , achieving an AUC of 0.57.

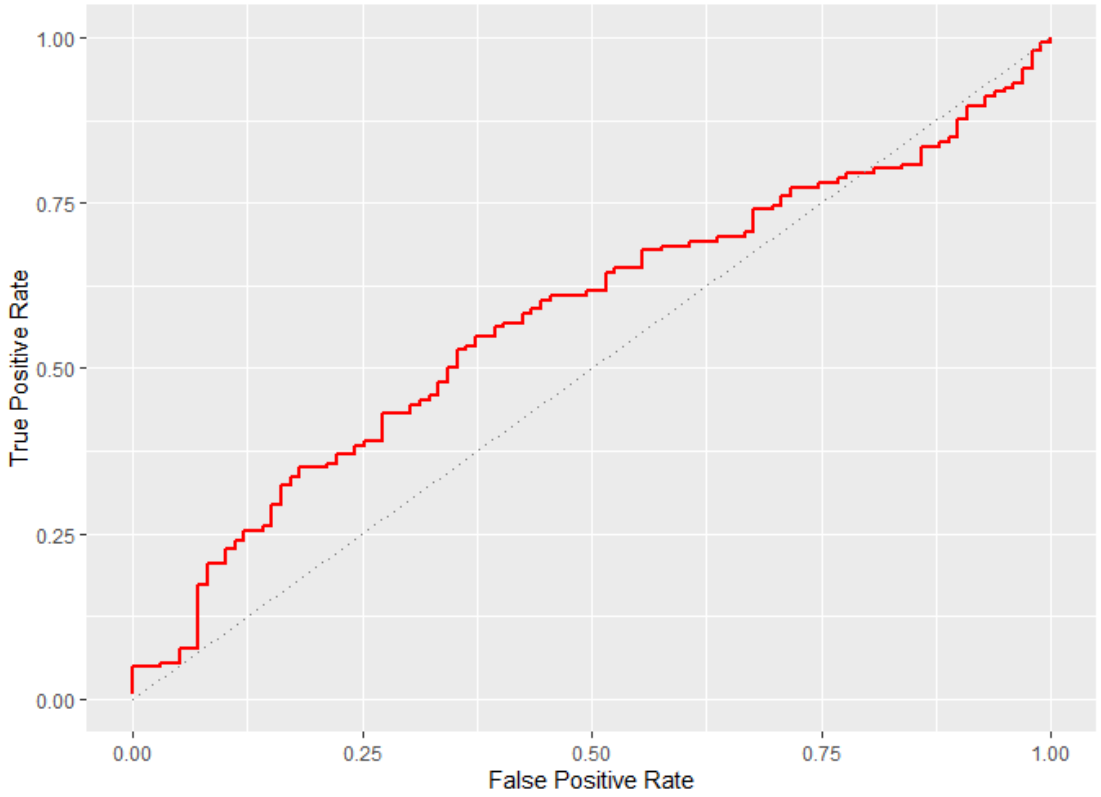

Figure 17: ROC curve of a GBM model trained on a dataset based on 88 numbers associated with 4 splits for each chromosome for each case. The model was then tested on an unseen split of data, achieving an AUC of 0.57.

Through multiple runs and a comparative analysis of the average AUCs from 5 runs between the 1-split and 4-splits analysis, along with statistical testing, we observed that there is no significant difference in the model's performance for predicting breast cancer recurrence when we allow for capturing more detailed information about the Chromosomal Scale Length Variations (CSLVs) by increasing the number of splits.

We explored the impact of increasing the training time on the model's prediction performance. We adjusted the training time to 12 hours and 24 hours, and although the

model's performance improved slightly on the training split of the data, there was no significant change in its prediction on the test split. The AUC value only increased by approximately 1-2%, indicating that further increasing the training time did not substantially enhance the model's predictive capabilities on the unseen data.

To validate our results and ensure that the predictions are not solely due to overfitting of the model on the input data, we conducted another validation method to assess the generalizability of the model. This approach allowed us to assess the model's performance when the target labels were randomly assigned, effectively eliminating any meaningful relationship between the input data and the predictions. By comparing the AUC of the ROC curve from the unshuffled dataset with the shuffled dataset, we could validate whether the model's predictions were genuine and not a result of overfitting.

Dataset	Mean AUC (SD)	P value of t test
1139 data points(Recurrence vs Non-Recurrence, 40:60)	0.54 (0.04)	0.322
1139 data points randomly shuffled on Target column	0.52 (0.03)	

Table 4: Comparison of Mean AUC Values for ML Models Trained with Unshuffled and Shuffled Target Columns.

Figure 18 displays the ROC plot of the top model that was trained and tested on the completely shuffled target column. **Error! Reference source not found.** presents a comparison of the mean AUC values from 5 individual runs of ML models. The models were trained using the unshuffled target column and then tested on the remaining split of data. Additionally, the table includes results from the same dataset, but with the target column randomly shuffled. This allows us to determine if the model's predictions are genuinely generalizable.

Figure 18: ROC curve of the top ML model that was trained and tested on shuffled target column, with an AUC 0.522.

By analyzing the mean AUC values of the two separate approaches and conducting a t-test, we found that there was no significant difference between the predictions generated using the unshuffled target column and those obtained when the target column was randomly shuffled. The calculated p-value further supported this conclusion.

Based on the above results, we have concluded that there is no significant genetic predisposition for breast cancer recurrence. Instead, we found that recurrence of the malignant tumor is more strongly correlated with factors such as age, breast cancer subtype, stage, grade, histology, and treatment strategy of the initial diagnosis. [120,121]

There are several limitations associated with our study. One of the main limitations is the limited availability of recurrence data within each specific primary site in the UK Biobank study. Due to this limitation, our selection criteria for recurrence cases could not be as precise as we would have liked. If we had applied stricter filtering based on the time of recurrence after the first diagnosis and the histology type of the recurred tumor compared to the primary tumor, we would have had less data available for model development and analysis. As a result, the specificity of our model may be affected. Another limitation is the possibility that some non-recurrence data could actually be recurrence cases where the second or third primary tumor was not reported to the UK Biobank study. This could introduce a degree of misclassification in our dataset and affect the accuracy of our model.

Additionally, the lack of complete and detailed information on treatment regimens and follow-up care in the UK Biobank dataset may limit our ability to fully account for the impact of different treatments on recurrence risk.

Furthermore, the genetic data used in this study were derived from germline DNA, and environmental factors were not incorporated in risk determination. As a result, the model may not fully capture the complex interplay between genetic and environmental factors in breast cancer recurrence.

Overall, these results emphasize the importance of considering clinical, treatment and environmental factors when dealing with malignant tumors after the first diagnosis.[82,120,121] While genetic predisposition may still play a role, its impact may be less significant, or it may be more specifically related to certain SNP mutations rather than germline structural variations for breast cancer recurrence. Due to the limitation of available data for oral OTSCC cases in UK biobank we could not tested our hypothesis on this type of cancer. Our study highlights the need to take a comprehensive approach, considering both genetic and non-genetic factors, to better understand and predict breast cancer recurrence. Future research should aim to address these limitations and explore additional factors that may influence breast cancer recurrence and analyze genetic predisposition for OTSCC recurrence.

CHAPTER 6: DEVELOPING CHROMOSOME SCALE LENGTH VARIABILITY-BASED GENETIC RISK SCORES FOR PREDICTING CANCER OCCURRENCE

In this chapter, we expand our investigation beyond predicting cancer recurrence to develop a risk estimate model for predicting cancer occurrence. The primary objective of the final aim in this study is to investigate the development of a model capable of accurately classifying patients into high-risk and low-risk categories for the development of specific types of cancer. By utilizing advanced machine learning techniques and incorporating relevant genetic features, our aim is to enhance our ability to identify individuals at a higher risk of developing cancer. This objective aligns with our second aim, which focuses on investigating a novel approach for incorporating inheritance factors into the development of polygenic risk score models for cancer recurrence. Instead of relying solely on single nucleotide polymorphisms (SNPs), we explore the utilization of structural variations within chromosomes as a mean to enhance the predictive power of the risk score model. [105,122] However, we apply this approach for developing a ML model for predicting cancer occurrence. By incorporating these structural variations, we aim to capture a more comprehensive representation of inheritance factors and improve the accuracy of our risk estimation. This model holds great potential for improving early detection and implementing targeted prevention strategies to reduce the burden of cancer.

Currently, there are several risk estimate models for different types of cancer. One of the most popular risk assessment tools for breast cancer is the Gail model, which is based on a statistical model. The model is named after Dr. Mitchel Gail the Senior Investigator. This tool allows professionals to estimate a 5-year and lifetime risk of developing invasive breast cancer

in women. This statistical model uses an individual's own personal information such as age, race, age of first menstrual period, number of past breast biopsies, etc. to calculate the risk of developing breast cancer. [123,124] However, similar to other developed risk assessment models, this tool does not give a good risk estimate for women with a personal history of invasive breast cancer due to the limited way that the model incorporates family history. [66] The model was originally developed based on data from white women and there are some concerns that may underestimate risk for women with other racial and ethnic groups.

For colorectal cancer, doctors and other health care providers use a risk assessment tool which gives a risk estimate of colorectal cancer over the next 5 years and the lifetime for men and women. However similar to the Gail model there is not much data available for black/African Americans, Asian Americans, and Pacific Islanders which limits the predictive ability of this model for these groups.[125] R package calculates the individual risk of lung cancer based on age, education, race, smoking history, body mass index, etc. Comparable with other models, inheritance factors and family history are not incorporated into the calculation of risk.[126]

It's been shown that genetic factors play an important role in developing lung cancer, which could add great predictive value to the model. [127]

It's been shown that structural variations at germline DNA have valubale prediction ability in developing complex diseases. [15,52,122] By incorporating these structural variations (CSLV) in nonlinear way, we aim to capture a more comprehensive representation of inheritance factors and enhance the accuracy and precision of our risk estimate model. The current developed polygenic risk scores are derived from the available genetic data in large

healthcare datasets. However, one of the main limitations of these datasets is the presence of data inequality. This inequality arises from various factors, including differences in participant demographics, representation of diverse populations, and data collection strategies. As a result, the polygenic risk scores may not accurately capture the genetic diversity and risk profiles of underrepresented or marginalized populations. [128]In large-scale biomedical research programs such as The Cancer Genome Atlas (TCGA) and the UK Biobank study, there is a notable imbalance in data representation, with a disproportionate number of participants from European American and White populations. This lack of diversity poses challenges in terms of generalizability and applicability of findings and risk assessment models to underrepresented populations.

For instance, in TCGA, more than 80% of participants are of European American ancestry, while the UK Biobank study has over 94% of participants from the broad category of White ethnicities. This imbalance limits the accuracy and relevance of polygenic risk scores and other genetic findings for individuals from diverse backgrounds.

To address these limitations, it is crucial to promote inclusivity and diversity in genetic research. Efforts are underway to expand data collection initiatives and encourage participation from underrepresented populations. By including individuals from diverse racial and ethnic backgrounds, researchers can improve the accuracy and applicability of risk assessment models, ensuring that they are more equitable and reliable for all individuals, regardless of their ancestry or background.

6.1: Research design and methods

6.1.a: All of Us Dataset

In this investigation, the All of Us dataset was utilized to compute genetic risk scores for various types of cancer. The All of Us study, sponsored by the National Institutes of Health, has enrolled more than 600,000 participants as of June 18th, 2023, with 80% of them coming from underrepresented communities. Figure 19 (provided by All of Us study) showcases the self-reported races and ethnicities of the participants who have completed the initial steps of the program, providing a diverse representation. The recruitment process spans all regions of the United States. The All of Us workbench encompasses a wealth of information gathered from electronic health records, including data from Fitbit devices, survey responses, and socioeconomic factors. Notably, a recent release of data in April 2023 included approximately 245,400 whole genome sequencing records and 312,940 genotyping microarrays, further enhancing the dataset's depth and potential for analysis.

By application of available data in All of Us, we will include data from communities and backgrounds in our model which has always been underrepresented in biomedical research.[129]

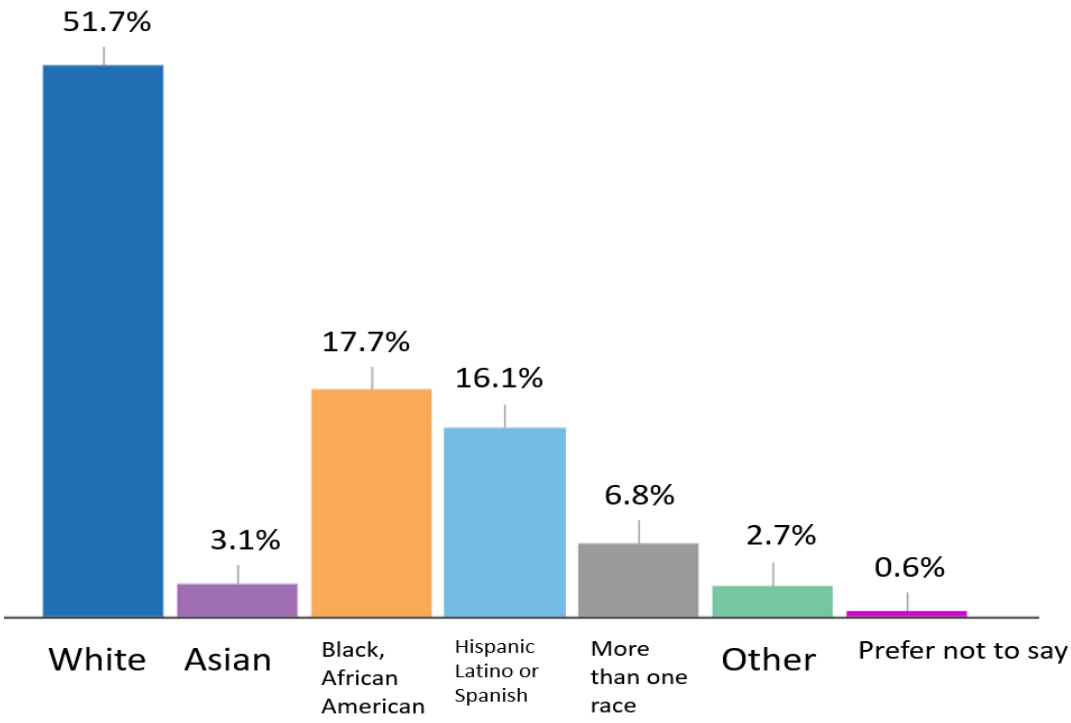

Figure 19: Racial distribution of participants in All of Us dataset

In our innovative approach, we have incorporated CSLV in model development and utilized advanced machine learning algorithms to overcome the limitations of current risk assessment tools. This strategy directly addresses two major limitations of existing risk assessment tools: the lack of diversity in the data used for model development and the failure to incorporate inheritance factors and family history. By leveraging the power of machine learning and considering a broader range of genetic information from the All of US dataset, our approach holds promise for more comprehensive and personalized risk assessment in the field of healthcare.

6.1.b: Genetic Data in All of Us

The All of Us dataset has released its genetic data in two phases. The initial release, known as controlled tier V6, included DNA microarray data for 165,127 participants. More recently, the controlled tier V7 was released in April 2023, and it is anticipated that additional data will be added to the dataset by the end of 2023. The microarray data in controlled tier 6 contains measurements on 1,814,517 genetic variants for each of the 165,127 participants. Illumina Global Diversity Array GDA has been used for microarray data analysis. The GDA is a cost-effective variant coverage solution within the Illumina human array portfolio. It offers comprehensive coverage of disease-associated and pharmacogenetic variants for clinical research. Built on a high-density SNP backbone, the GDA provides optimal cross-population imputation coverage and enables the development of polygenic risk scores. It is a valuable tool for studying genetic architecture and characterizing genetic traits in diverse populations.[130,131]

The header of the array Variant Call Format (VCF) file contains information about the sample processing. VCF is a common file format used to represent genetic data from multiple individuals, including data from the All of Us genetic dataset. The VCF header includes format specifiers for different fields, such as the key parameter for CSLV (log R ratio or LRR). This information can also be processed via Hail matrix, a scalable and flexible framework for genetic analysis. Hail MatrixTable is a tabular data structure that is often used to represents a matrix of genetic data. It organized genetic data into three dimensions. This allows for efficient querying, filtering, and transformation operations on large-scale genetic datasets.

6.1.c: Data Extraction & Processing by using Hail MatrixTable

For data analysis and model development within the All of Us dataset, a dedicated researcher workbench was established. The Researcher Workbench is a cloud-based platform that grants registered researchers access to both Registered and Controlled Tier data. Equipped with robust tools, it enables efficient data analysis and promotes collaborative research efforts. In our study, we extracted microarray genetic information for each participant by analyzing the data stored in the Hail Matrix table. This step was conducted on integrated cloud based Jupyter notebook environment using Python programming language on the designated workbench. [132] Figure 20 depicts a Hail matrix table, which is structured with three dimensions.

Figure 20: Hail matrix representation within Jupyter notebook on all of us workbench.

The column fields represent individual participants in the study, allowing for the identification of specific individuals within the dataset. The row fields, on the other hand, contain constant information that applies to entire rows of entries. In this table, the row field represents the locus, which refers to the specific location on a chromosome where a genetic

marker is situated. This field can be used to efficiently query or manipulate subsets of the rows based on their genomic location or other annotations. Lastly, the entry fields are indexed by both row and column and encompass various attributes such as Genotype (GT), Illumina GenCall Confidence (IGC) Score, Raw X and Y intensities as scanned form the original genotyping array, normalized X and Y intensities, normalized R value, normalized Theta value, Log R ratio, B allele frequency (BAF) etc.

In this investigation, our focus was on incorporating structural variations into the development of the machine learning (ML) model. Incorporating structural variations, such as insertions, deletions, translocations, and copy number variations, provides insights into the individual's chromosome length. These structural variations slightly modify the overall length of a person's chromosomes. To achieve this, we specifically extracted the log R ratio (LRR) values from the entry field for each patient at different loci and excluded other values. The LRR values represent the logarithm of the observed signal intensity ratio, providing information about the copy number status or dosage of genetic material at a specific locus. By computing the average LRR values over a chromosome or a portion of it, we obtained the nominal length known as chromosome scale length variation (CSLV). A value of 0 indicates the presence of two copies at the locus, while higher values indicate duplications and lower values indicate deletions.

To begin our analysis, we filtered the genetic data for each chromosome and stored it in separate Hail Matrix tables within our workbench. We then calculated the average LRR values within all segments of the chromosome along each column of entries, where each column corresponds to a specific participant. The resulting average values were stored as new column annotations in a new Hail Matrix table. We subsequently analyzed the column fields (column

fields are the average values of all the LRR values of that specific chromosome and patient IDs) of this new table individually, focusing on the average LRR values for each chromosome and patient ID.

To facilitate further analysis and reduce computational load, we converted the column fields table into a Pandas DataFrame format. This format offers greater flexibility for data manipulation. The resulting DataFrame consists of 165,127 rows, representing each participant, and two columns: patient ID and average LRR value for the analyzed chromosome. These steps were repeated 22 times to calculate the average LRR values for each chromosome.

Figure 21 presents a histogram that illustrates the distribution of relative chromosome lengths obtained from DNA samples in the All of Us dataset, specifically for chromosomes 1, 7, 13, and 19. A value of “0” represents the nominal average chromosome length. This visualization provides insights into the variations in chromosome lengths within the dataset.

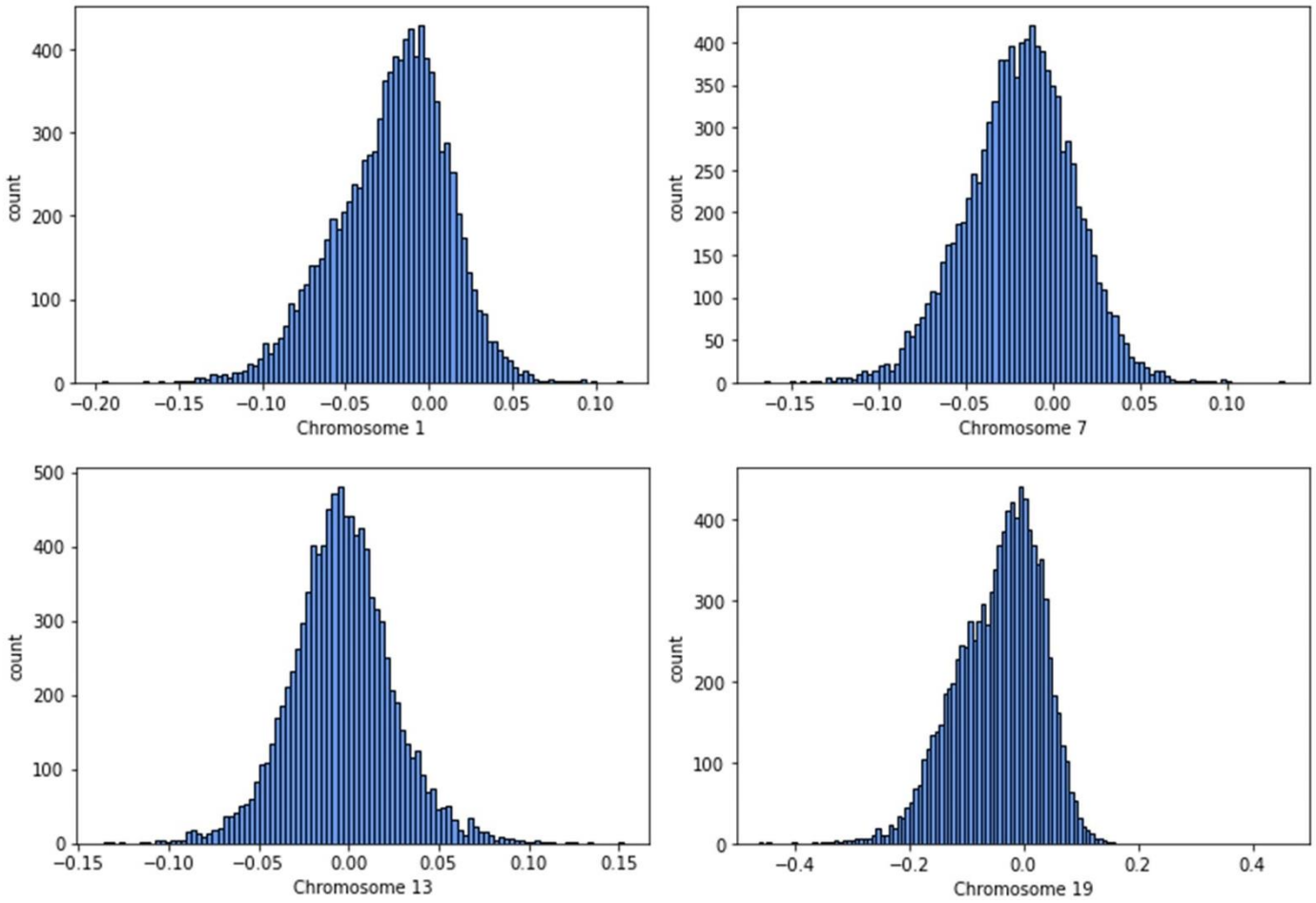

Figure 21: The distribution of relative chromosome lengths obtained from DNA samples in the All of Us dataset for chromosomes 1, 7,13 and 19. These histograms were plotted for average LRR values of 10,000 participants that were randomly selected from 165,127 participants in the All of Us dataset V6.

The analyzed results for all 22 chromosomes of the participants in the controlled tier V6 were saved in a storage bucket associated with our workbench in the All of Us dataset. This storage bucket allows for convenient access and sharing of the analyzed data among registered researchers who have access to the workbench. This collaborative feature facilitates further exploration and investigation of the CSLV analysis in the dataset.

6.1.d: Machine Learning Model Development

Once we calculated the average LRR values within each chromosome or segments of chromosome for all the participants of controlled Tier V6, the CSLV data table was built. Then we explored the ML model development within the environment of Jupyter notebook. For that purpose, the cohort builder in the workbench was used. The Cohort Builder is a custom, point-and-click tool that allows researchers to create, review, and annotate groups of participant data, or cohorts, within the *All of Us* dataset. Once the participants of the interest were selected the dataset builder was used to collect the relevant information about the cohort subjects. This allowed us to query through the participant level data and select the positive-class, participants who have been diagnosed with specific type of cancer and negative class, age matching participants who have not been diagnosed with that specific type of cancer. These data tables are accessible within the workbench and other relevant health information can be linked to each participant of two classes of datasets through the built in feature called concept set within the dataset builder.

The next step in model development involves identifying the shared participants between the CSLV data table and the positive and negative classes for a specific type of cancer. As we have mentioned before each participant is assigned a unique identifier called the person ID, which is used to match individuals with relevant genetic information for the target classes. The last step of data processing before ML model development is under sampling of the control group. In this step, the control group, participants who have not been diagnosed with the specific type of cancer, is under-sampled and age-matched with positive class. This involves randomly selecting participants from the negative class within each age group, while

maintaining a 40:60 percent ratio between the positive class and control. Once the finalized dataset is constructed, the H2O machine learning package, which was previously utilized in aims 1 and 2, is employed to develop a polygenic risk score using a range of different machine learning algorithms. In this step, the H2O package was installed and imported into the Jupyter notebook for each run. To enhance the performance of the developed models, the cloud analysis configuration in the workspace was adjusted as follows: the cloud compute profile was set to CPUs = 16 and RAM = 60 GB, and the worker configuration was set to have 2 workers with CPUs = 4, RAM (GB) = 15, and External Disk (GB) = 150. The finalized dataset was divided into a 20% test set and an 80% training set for each run. The model was developed using the 80% training split. We utilized the H2O AutoML built-in function, which runs through various machine learning algorithms with specific configurations, for each run. The H2O AutoML function was configured with a maximum run time of 900 seconds and 5-fold cross-validation. During each run, the cross-validation predictions were retained on the leaderboard. Additionally, the trained model's performance was evaluated on the unseen test split of the data. All of these steps were done on Jupyter notebook in python programming language.

6.2 Results

Preliminary findings from previous studies suggest the presence of substantial inherited variations across different types of cancer. Moreover, it is evident that the prediction of cancer based solely on germline copy number variations is feasible. [14,122] A similar method has been previously used for assessing the influence of inheritance factors on specific traits using other biobanks. *Toh* et al applied similar approach to develop a poly genic risk score for prediction ovarian cancer by utilizing genetic data in Cancer Genome Atlas (TCGA) project. [14]TCGA was a

project sponsored by the National Cancer Institute. The developed genetic risk score had an AUC of 0.88 which provided an effective means of predicting whether or not a woman will develop an ovarian cancer. In another study conducted by *Toh et al,* a similar strategy was used to predict schizophrenia based on available genetic data in UK Biobank dataset. In their work they were able to develop a model that could distinguish schizophrenia patients from control with an AUC of 0.54. This was an indication of significant genetic correlation, however not very precise. [133]In another study conducted by Ko et al, they employed a similar strategy and successfully developed a machine learning (ML) model for predicting breast cancer using data from the UK Biobank and The Cancer Genome Atlas (TCGA). Their model achieved an impressive AUC (Area Under the Curve) value of 0.83, indicating its strong predictive performance. [134]

For the third aim of this investigation, we focused on developing a polygenic risk score (PRS) to predict the occurrence of different cancer types, namely malignant breast tumors, malignant oral cavity tumors, malignant colorectal tumors, and malignant ovarian tumors. To accomplish this, we utilized the CSLV data table extracted for all 165,127 participants from the controlled tier V6 of the All of Us dataset. The goal of this aim was to overcome the limitations associated with existing risk assessment models and develop a comprehensive PRS that can accurately predict the risk of specific cancer types. By leveraging the CSLV data, which provides detailed information on the genomic structural variability of participants, we aimed to create a more precise and personalized risk scoring system.

6.2.a Genetic Risk Score Model for Determining Risk of Breast Cancer

Breast cancer is the most commonly diagnosed cancer in women worldwide, accounting for 25% of all cancer cases among women. It is estimated that approximately 2.3 million new cases of breast cancer were diagnosed in 2020 alone. [135–137] Due to the advancements in technology and screening methods, the breast cancer death rate continues to decline. However, there is a significant disparity in breast cancer mortality between black and white. This could be attributed to various factors, including disparities in access to healthcare, socioeconomic status, genetic factors and disparities in cancer screening and treatments. These factors contribute to differences in early detection, diagnosis, and timely access to quality healthcare, which can impact the outcomes and survival rates for individuals with breast cancer. By early stage detection of breast cancer, this disease can be curable through mastectomy.

Several predictive models based on inheritance factors, such as the Gail model and Tyrer-Cuzick model, have been developed to aid in the early prediction of breast cancer. The Gail model, a statistical model used for breast cancer risk assessment, incorporates parameters such as age, race, and first-degree relatives to provide a 5-year risk estimate. However, it exhibits modest predictive accuracy with an area under the curve (AUC) of 0.58.[123,124,138] Models that include combinations of SNPs or other genetic factors in their personalized risk calculations have shown improved predictive ability compared to methods that rely solely on family history surveys. This is because family histories are often incomplete and limited to only one or two generations, whereas poly genic risk models can capture a broader range of genetic variations that contribute to disease risk. On the other hand, the Tyrer-Cuzick model

incorporates additional genetic information, including BRCA1 and BRCA2 mutations, along with age, family history, and other factors. This model demonstrates improved predictive ability compared to the Gail model, with an AUC of 0.62 and a 95% confidence interval (CI) ranging from 0.60 to 0.64. An extended version of the Tyrer-Cuzick model, incorporating a 313 variant polygenic risk score, achieves an AUC of 0.64. [139–141] In one of the latest research, polygenic risk score had been calculated based on linear combinations of 313 SNPs to predict breast cancer with an AUC of 0.63. [4]

These personalized risk assessment tools have never been able to achieve higher AUC than 0.65 and their accuracy has been limited by the way that have analyzed the SNPs. The disparity in breast cancer mortality between different races is also partly attributed to the fact that many screening methods and prevention guidelines have been developed based on studies predominantly involving white populations. As a result, there may be limitations in the effectiveness and accessibility of screening and prevention strategies for minority populations, contributing to disparities in breast cancer outcomes. Addressing these disparities requires a more comprehensive and inclusive approach that considers the diverse characteristics of all populations and include SNPs interaction in a novel way.

Developing a personalized risk assessment model for breast cancer using the available data in All of Us offers the potential to overcome limitations observed in previous studies, primarily due to the unique nature of the dataset. We utilized the calculated Chromosomal Scale Length Variation (CSLV) data table from section 6.1.d as the basis for developing our ML model.

Our dataset consisted of two main components: the positive class and the control group. The positive class comprised 7,998 women from the All of Us dataset-controlled Tier V6 who had been diagnosed with malignant breast tumors. On the other hand, the control group was constructed by excluding men and women with any type of cancer diagnosis, resulting in a cohort of 13,794 cancer-free women. From these two separate datasets, we selected participants with available genetic information and created two distinct case-control tables. Specifically, we identified 4,533 cases with microarray genetic data in the positive class and 44,518 cases with microarray genetic data in the control group. **Error! Reference source not found.** represents the racial distribution in positive and control classes.

For model development, in each run, an under-sampled age-matched group of participants from the negative class is selected. This group is then appended to the positive class and used as input for the H2O AutoML function. The purpose of this step is to achieve class balance and ensure that the model is trained on a representative dataset that accounts for an equal age distribution in both the positive and negative classes.

Class	Race Distribution					Sex	Total
	White	Black	Asian/Middle Eastern	Other	Unknown	Female	
Positive (Malignant Breast Tumor)	3342	440	129	64	558	4533	4533
Negative (Control)	17109	13436	1796	1318	13235	44518	44518

Table 5: Racial distribution of participants in positive & negative class of risk assessment model for breast cancer. The positive class includes women with a diagnosis of malignant breast tumor, while the control group comprises cancer-free women

The initial model is developed by utilizing 22 numbers, each corresponding to the average log R ratio (LRR) values of all the SNP markers within each chromosome. These average LRR values capture the overall copy number status or dosage of genetic material at specific loci on the chromosomes. By incorporating these 22 numbers as features in the model, we aim to capture the genetic variations associated with breast cancer risk and improve the accuracy of the risk assessment model. Using the All of Us dataset, we developed a classifier for identifying breast cancer patients, which achieved an AUC of 0.60 with a standard deviation (SD) of 0.003, Figure 22. The Best Classifier identified with the H2o automl package was stacked ensemble.

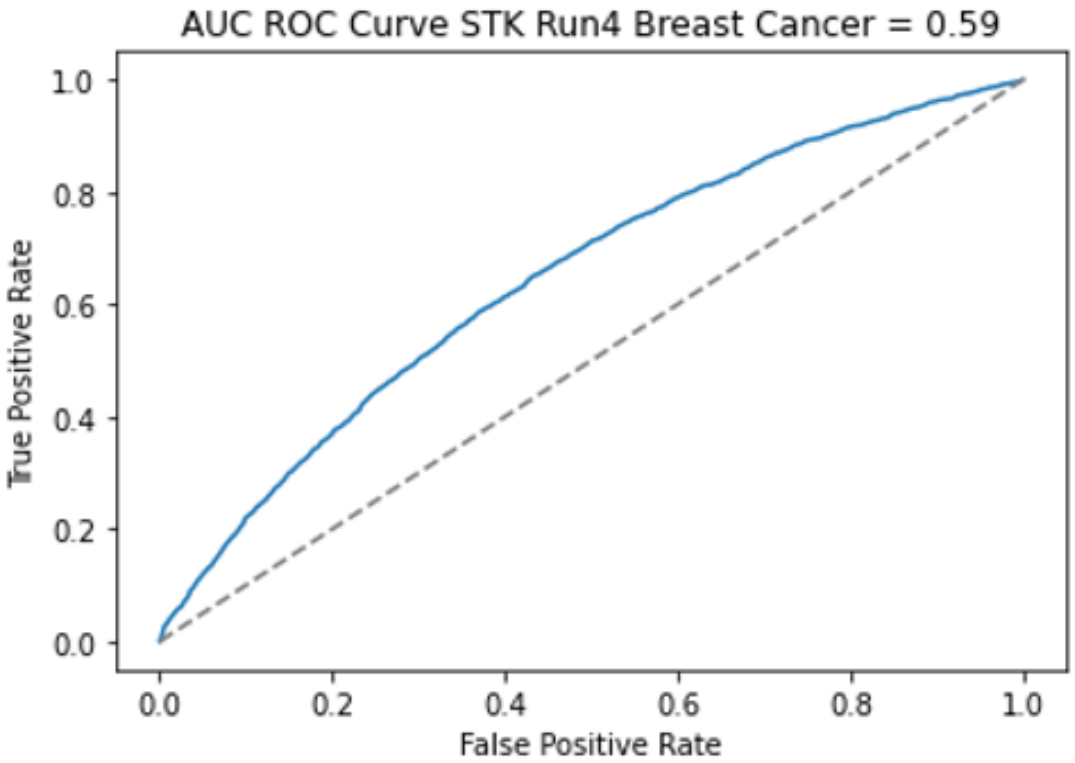

Figure 22: ROC curve of the stacked ensemble model developed using 22 numbers (1split) with each number corresponding to an individual chromosome. The model achieved AUC of 0.59.

The model's performance, as measured by AUC, on the training and cross-validation sets was 0.70 and 0.61, respectively. These results indicate that the model shows promise in predicting breast cancer status comparable with other available polygenic risk scores. [4,139–141]

To improve the model's performance and its predictability compared with other developed polygenic risk scores, we investigated other possible model modifications. As we have described above, for this step average LRR values across entire chromosome was calculated and used as a feature for model development, however we can split chromosome into 2,4, etc. equal segments and take average LRR values across these segments as a predictive feature. By this approach we are able to capture more detailed information about the structural variability of the genome. For instance, chromosome 1 has 146,409 genetic variants, in order to

calculate “4 splits of chromosome 1”, 36,602 sequential values would be averaged to form one value for the first 3 splits and then the last split would be the remaining 36,603 sequential values. This process results in 88 numbers that characterize each case in both classes. By utilizing these 88 numbers instead of the initial 22 numbers, a more comprehensive ML model was developed that incorporates a broader range of genetic information. The evaluation criteria and model’s characteristics remain consistent with the previous approach.

Using the same classes in table 5 with 88 measurements instead of 22, a classifier was trained and tested on unseen split of data with an average AUC of 0.70 with standard deviation of 0.01 on test split, Figure 23. The Stacked Ensemble model, similar to the previous approach, was identified as the best classifier for this optimization.

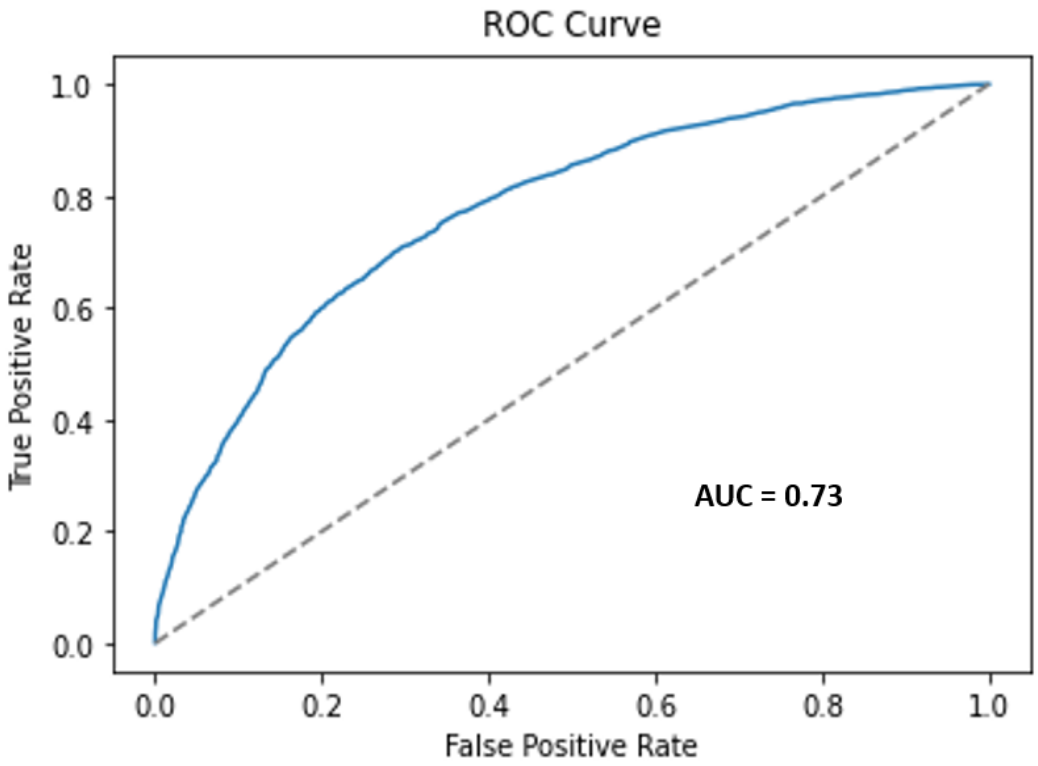

Figure 23: ROC curve of the stacked ensemble model developed using 88 numbers, each number corresponds to a quarter of an individual chromosome obtained by splitting each chromosome into 4 equal segments. The stacked ensemble model achieved an AUC of 0.73,

To compare the performance of the two approaches, the evaluation metrics were calculated at different stages of model development, including training, cross-validation, test, and values were plotted in Figure 24. The metrics include average values of AUC, accuracy, and F1-score calculated from 5 runs for both the 1-split and 4-split models. These metrics provide insights into the performance of the models and allow for a comparison between the two approaches. Comparing the mean AUC for the 4 split model vs 1 split model gave a p value of 9.83×10^{-14} indicating that finer splits significantly improved the predictive ability of the model. It demonstrates how the quality of these predictions increases with finer information on chromosome length variations.

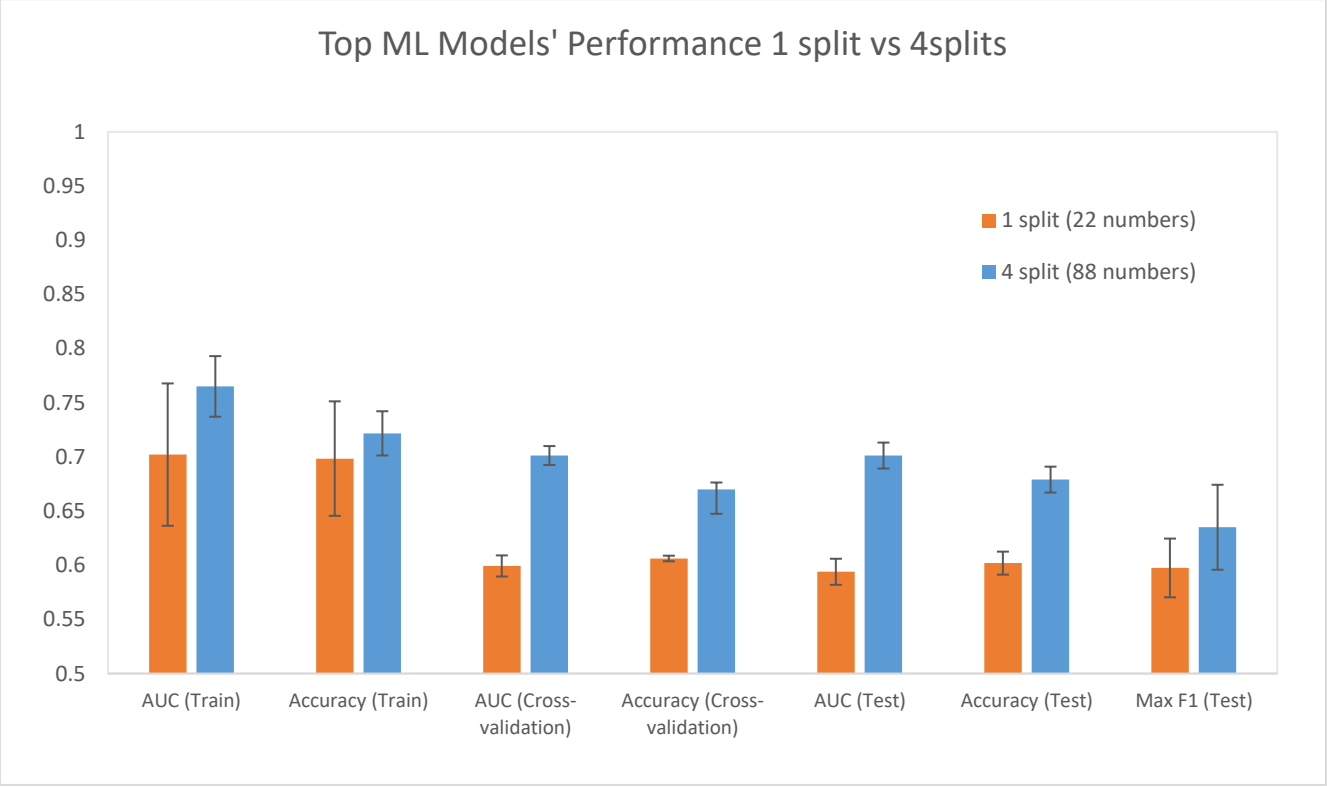

Figure 24: Performance comparison of the developed ML models by 1-split and 4-split approach based on average AUC, accuracy, and F1-score values from 5 runs.

By using H2O automl function, different ML algorithms trained during the training phase and the performance of the developed model compared by plotting the ROC curves on test split, Figure 25, and calculating other performance metrices at different stage of model development, Table 6.

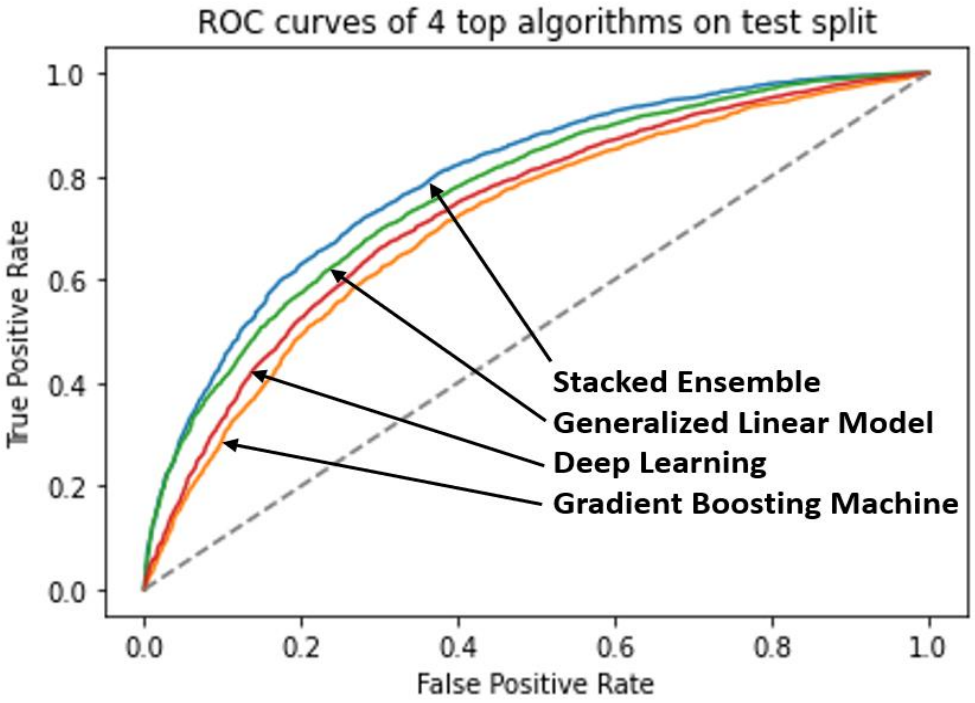

Figure 25: ROC plots of four developed ML models (Stacked Ensemble, Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), and deep learning) for predicting breast cancer based on 88 numbers derived from average LRR values of quarter segments of chromosomes. The ROC plots depict the performance of the models on the test split.

The impact of varying the allocated train time for model development on AutoML function was examined. By extending the time from 900 seconds to 12 hours and 24 hours, the AUC of the top model, stacked ensemble, improved by 1-2%. Notably, the deep learning models, which

initially had lower rankings during the 900-second training, significantly improved and rose to the top positions with the longer training times.

Model	AUC- Train (SD)	Accuracy-Train % (SD)	AUC-Validation % (SD)	Accuracy-Validation % (SD)	AUC-Test (SD)	Accuracy-Test % (SD)	F1 Score -Test % (SD)
Stacked Ensemble	0.76 (0.03)	72% (2.3)	0.70 (0.02)	67% (0.01)	0.70 (0.01)	68% (1.1)	0.64 (0.04)
GLM	0.71 (0.01)	66.5% (3.3)	0.69 (0.01)	67.2 % (1.4)	0.7 (0.01)	67.8% (1.0)	0.62(0.02)
Deep Learning	0.70 (0.03)	66.3% (4.1)	0.65 (0.05)	63% (2.5)	0.65 (0.03)	63%(5.1)	0.6 (0.05)
GBM	**0.69 (0.05)**	**65.3 (1.4)**	**0.55 (0.03)**	**57.2 (1.7)**	**0.56 (0.01)**	**58% (2.4)**	**0.59 (0.02)**

Table 6: Performance metrics of the top 4 machine learning models for predicting breast cancer. Performance metrics were reported as an average of 8 runs along with standard deviation.

Once we identified the best classifier, to stimulate a real- world application, a trained model is tested on unseen test split. The model returns a score for each women in test set. We assessed the accuracy of the model by ranking each woman based on the assigned score and evaluating its performance across quintile. A higher score indicates a higher likelihood of having breast cancer, and we calculated the odds ratio of the model within each quintile. Table 7 displays the

odds ratios calculated from the top quintile for each respective quintile relative to the entire test population, indicating the increased likelihood of having breast cancer associated with higher scores. The top 20% of women (ranked based on the score received from our risk estimate model) in our results had an increase of 9-fold risk over women who scored in the bottom 20%.

Quintile	Number of women with breast cancer	Number of women without breast cancer	Total number of women	Odds Ratio	95% confidence interval
5	310	93	2015	3.47	(2.88, 4.06)
4	221	182	2015	1.45	(1.24, 1.67)
3	167	236	2015	0.80	(0.74, 0.87)
2	113	290	2015	0.56	(0.49, 0.62)
1	89	314	2015	0.39	(0.36, 0.42)

Table 7 : This table represents the odds ratio between the quintile of predicted results from our trained model tested on unseen split of data. The result indicates that the top quintile is 9 times as likely to have an accurate prediction for breast cancer as the bottom quintile.

To understand how the model reaches its results, we analyzed the variable importance to identify the regions that contribute the most to the model's predictions. We focused on the generalized linear (GLM) model, which was ranked second on the AutoML function. As stacked ensemble models do not provide variable importance information, we plotted the relative importance of the most significant variables on the trained GLM model Figure 26. [45]This analysis helped us identify the key variables that had the most influence on the model's predictions.

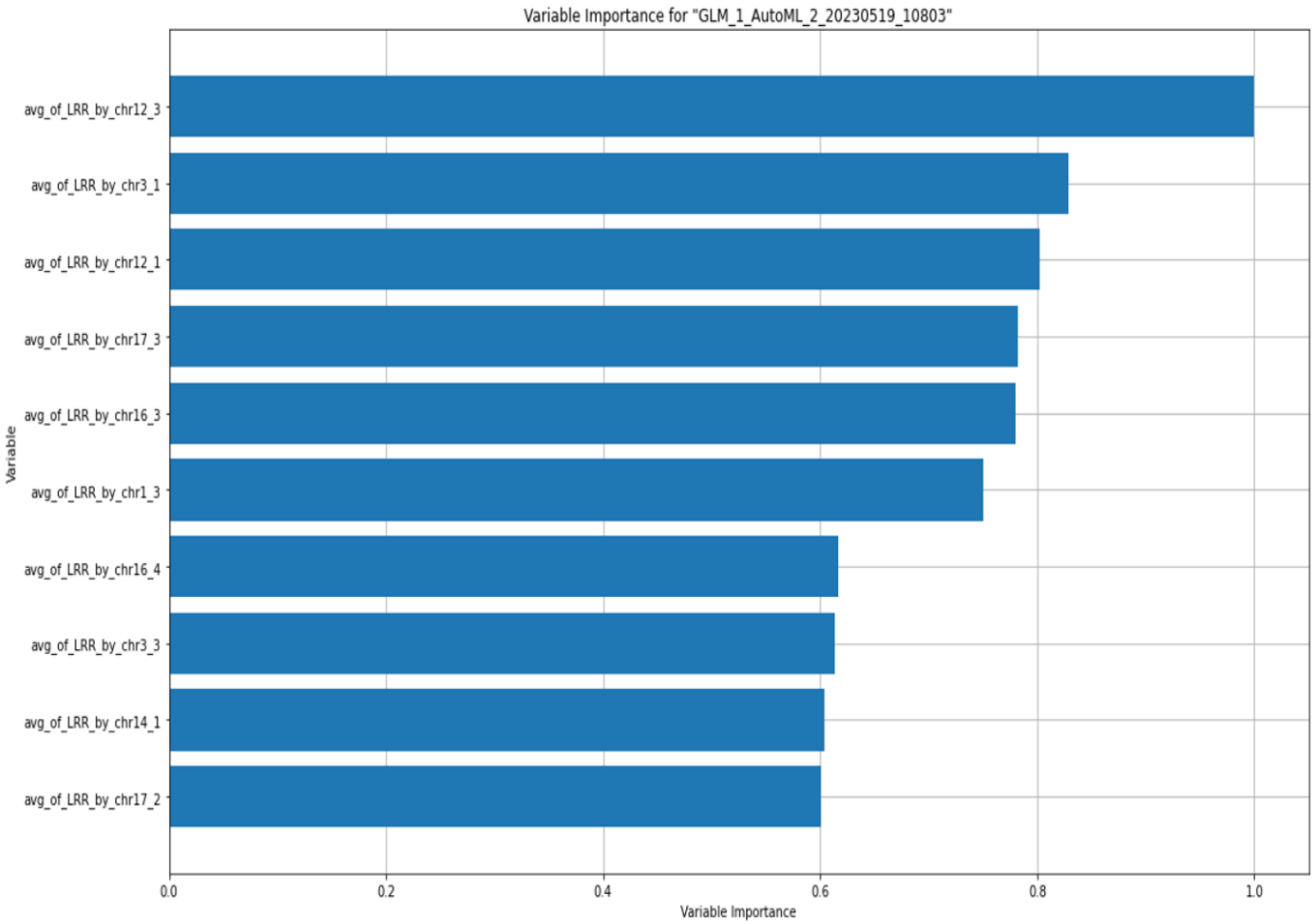

Figure 26 The variable importance plot on breast risk assessment model. The GLM model shows the relative importance of the most important variables in the model.

The SHAP summary plot was used as an additional method to gain insights into the machine learning (ML) model. This analysis specifically focused on tree-based models. Among all the tree based models GBM ranked higher on the leaderboard of h20 AutoML function. Figure 27 displays the SHAP summary plot for the GBM model.

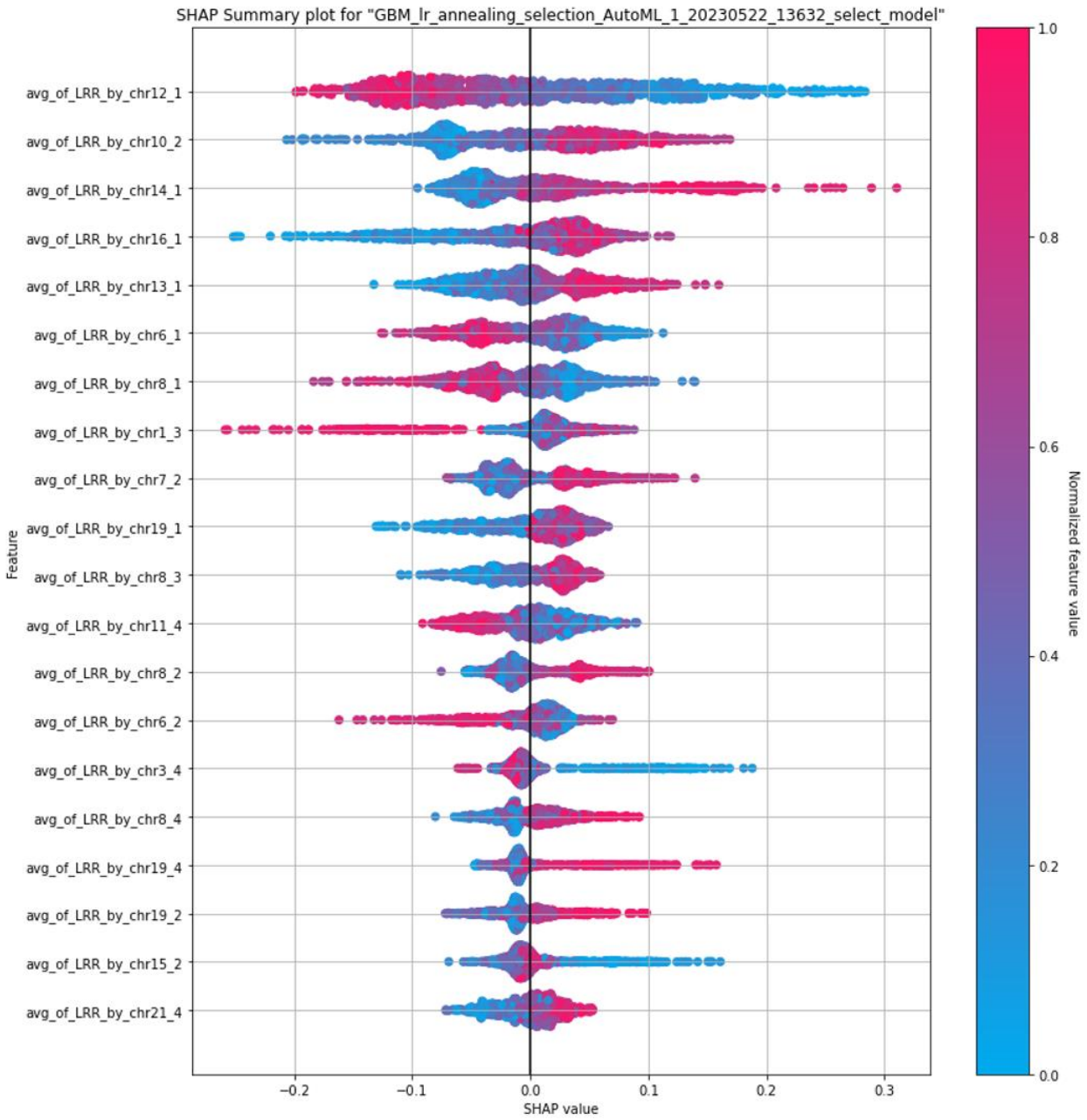

Figure 27: The Shapley additive explanations plot of GBM model for predicting breast cancer.

The analysis of SHAP values and variable importance calculations indicates that there is no single chromosomal region that contributes significantly more than others to the predictions of the model. Instead, the model relies on a combination of multiple chromosomal regions and genetic factors to make accurate predictions. This suggests that the overall genomic landscape and interplay between different regions play a crucial role in determining the risk of breast cancer.

6.2.b: Cross Races Analysis in Developed ML Model for Breast Cancer

We investigated the potential of leveraging the diversity within our datasets by developing a model based on a specific race and testing its performance on individuals from different races. This analysis allowed us to explore the transferability of the model across different racial groups and assess its generalizability. For this reason, we conducted ML model development for predicting breast cancer using different splits of available data for each racial group and tested the models on different racial groups. The positive class in this study, consists of 3,342 patients identifying as white, 440 patients identifying as black or African American, 107 patients identifying as Asian, 22 patients identifying as Middle Eastern, and 2 patients identifying as Native Hawaiian. Additionally, there are 620 patients in the positive class whose race is unknown. To conduct a cross-analysis considering all subpopulations in the positive class, we categorized the cases into three subgroups: white, black or African American, and other races. The "other races" category includes cases with unknown race as well as individuals from Asian, Middle Eastern, and Native Hawaiian backgrounds. Due to the limited number of cases in the Asian, Middle Eastern, and Native Hawaiian groups, they were not analyzed as separate groups in the cross-race analysis.

Once we categorized the positive class into three subpopulations, we aimed to maintain a balanced ratio of 40:60 between the positive and negative classes in the final dataset. To achieve this, we randomly selected an age-matched number of cases from the negative class within each subsection of race for each run. In this study, two different ML models were developed. The first ML model was trained on 80% of the finalized dataset of the white subgroup and tested on the remaining 20% split of the white population. The same trained model also was tested on the black or African American dataset, the other races dataset, and the combined dataset including white, black, and other races. The performance of each model was evaluated by calculating the maximum F1 score, maximum accuracy, and AUC of the ROC curve at each step of testing on each different subcategory of races. These evaluation steps were repeated for each ML model, the recorded values were averaged for 5 separate runs. The average values for evaluation metrics of the 5 runs, along with their corresponding standard deviations, are presented in Figure 28 and Table 8. The model development followed similar criteria as the previous models, and 5-fold cross-validation was employed to mitigate overfitting.

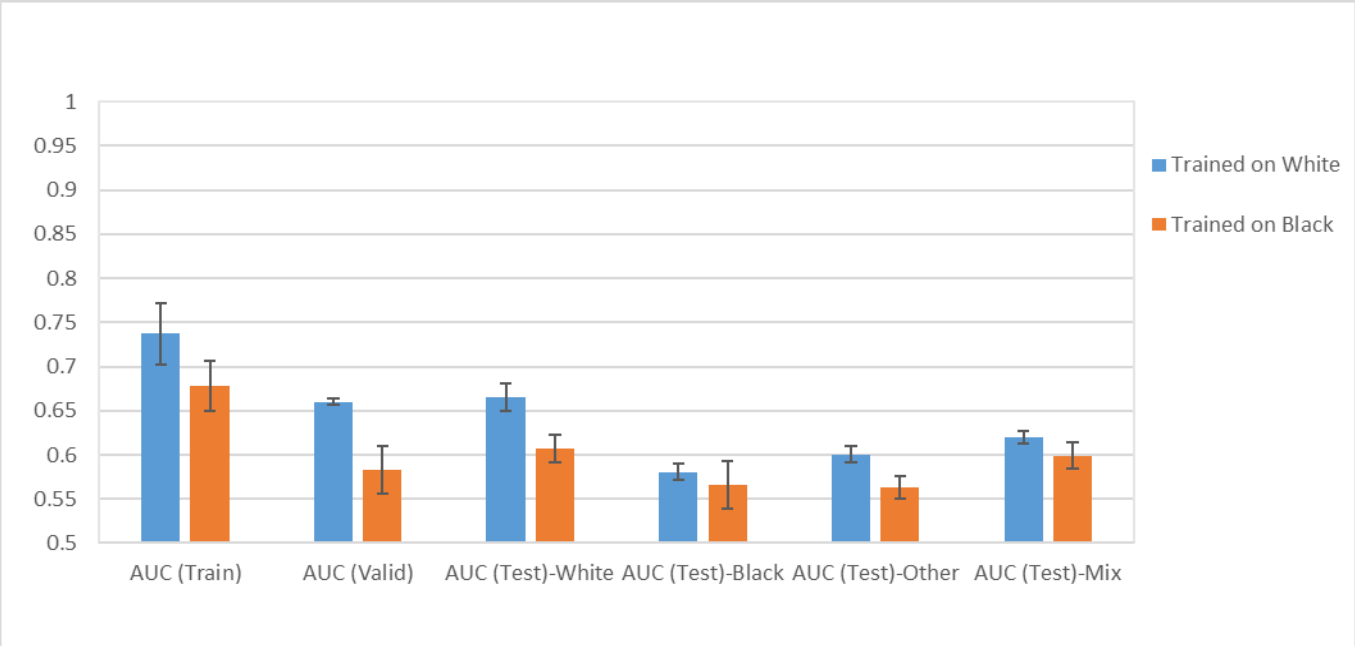

Figure 28: AUC values of two different types of ML models for predicting breast cancer. The first model was trained on the white subgroup, while the second model was trained on the black subgroup. AUC values were recorded for each run at different stages of model development, including training and cross-validation, as well as its performance on different subgroups of races.

Evaluation Model Metrics	Model Trained on White	Model Trained on Black
AUC (Train)	0.74	0.68
Accuracy (Train)	0.70	0.66
AUC (Valid)	0.66	0.58
Accuracy (Valid)	0.65	0.61
Max f1 (Valid)	0.59	0.58
AUC (Test)-White	0.67	0.61
Accuracy (Test) -White	0.65	0.62
Max f1 (Test)-White	0.61	0.57
AUC (Test)-Black	0.58	0.57
Accuracy (Test) -Black	0.61	0.63
Max f1 (Test)-Black	0.57	0.58
AUC (Test)-Other	0.60	0.56
Accuracy (Test) -Other	0.62	0.61
Max f1 (Test)-Other	0.57	0.57
AUC (Test)-Mix	0.62	0.60
Accuracy (Test) -Mix	0.63	0.62
Max f1 (Test)-Mix	0.58	0.57

Table 8: Evaluation metrics of two different ML models. The first model was trained on the white subgroup, while the second model was trained on the black subgroup. The metrics include the AUC, accuracy, and F1 score for each model tested on different races.

ROC curves for the trained ML model on the white subpopulation of the dataset, tested on

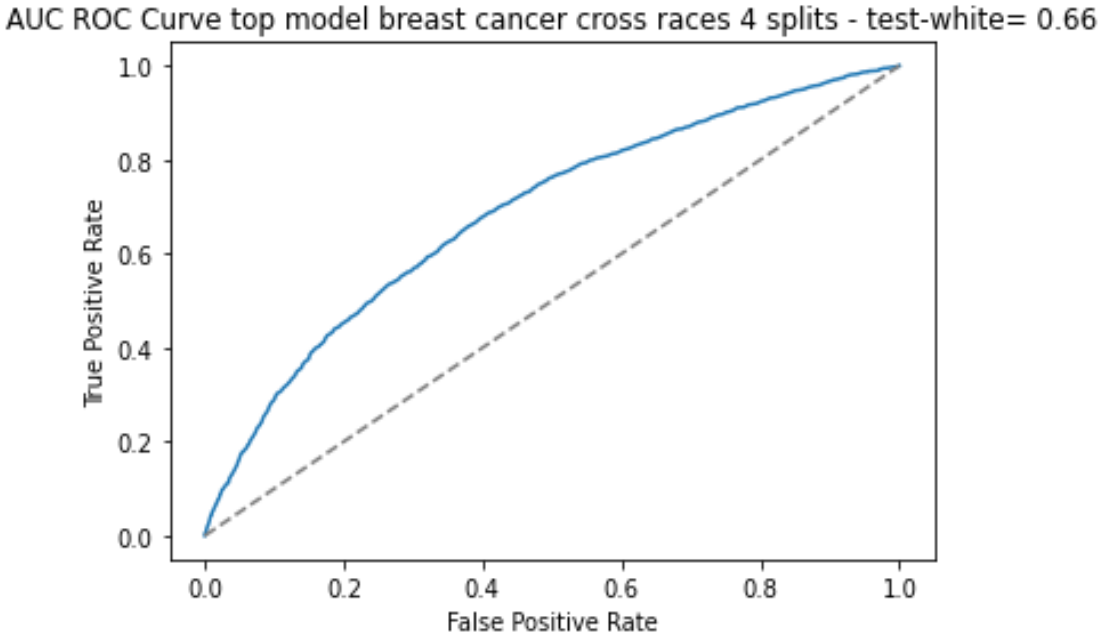

Figure 29: ROC curve of model trained on 80% of the white subpopulation and tested on the remaining 20% test split of the white subpopulation, with an AUC of 0.66.

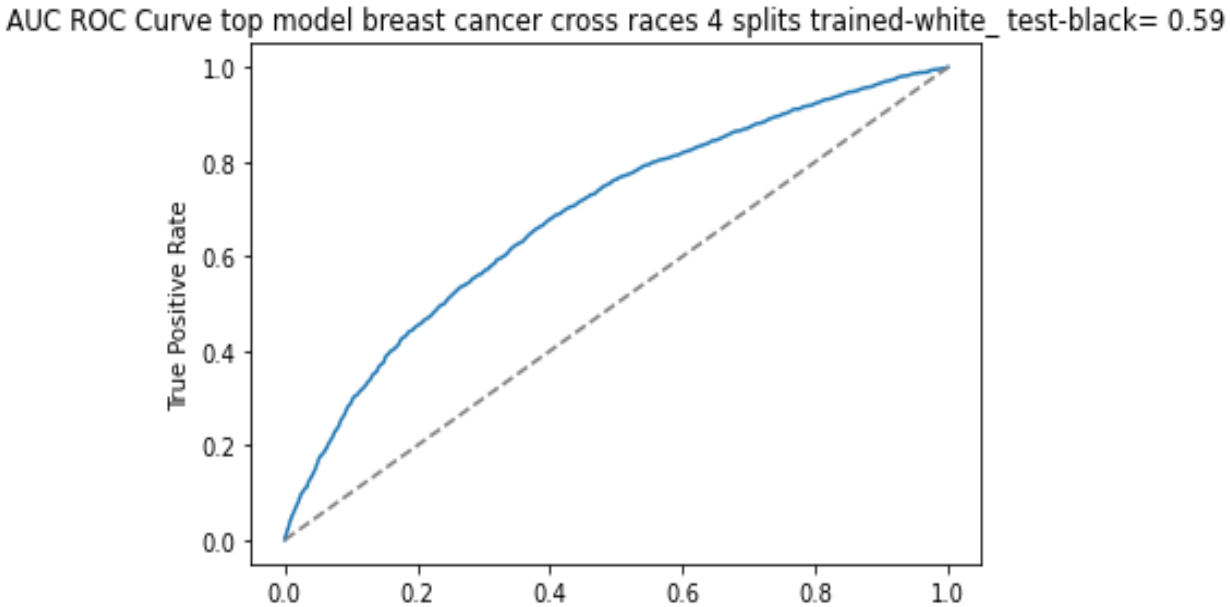

Figure 30: ROC curve of model trained on the white subpopulation and tested on the black subpopulation group with an AUC of 0.59.

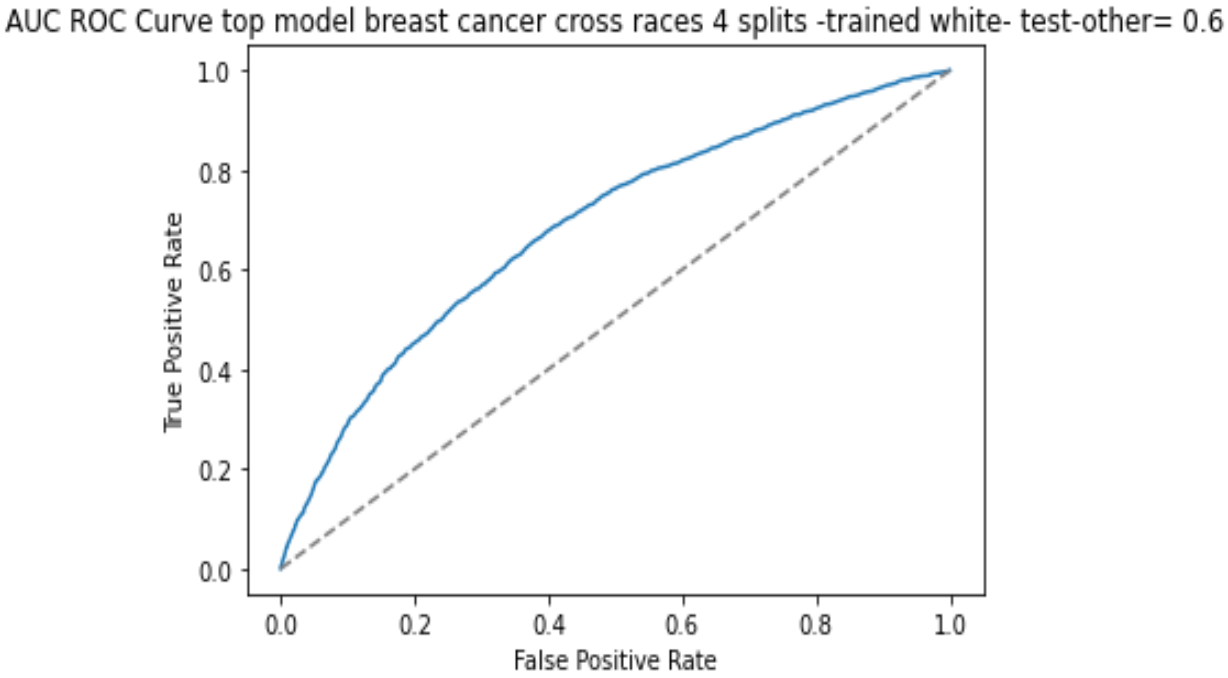

Figure 31: ROC curve of the model trained on the white subpopulation and tested on the other races split of the finalized dataset, demonstrating an AUC of 0.60.

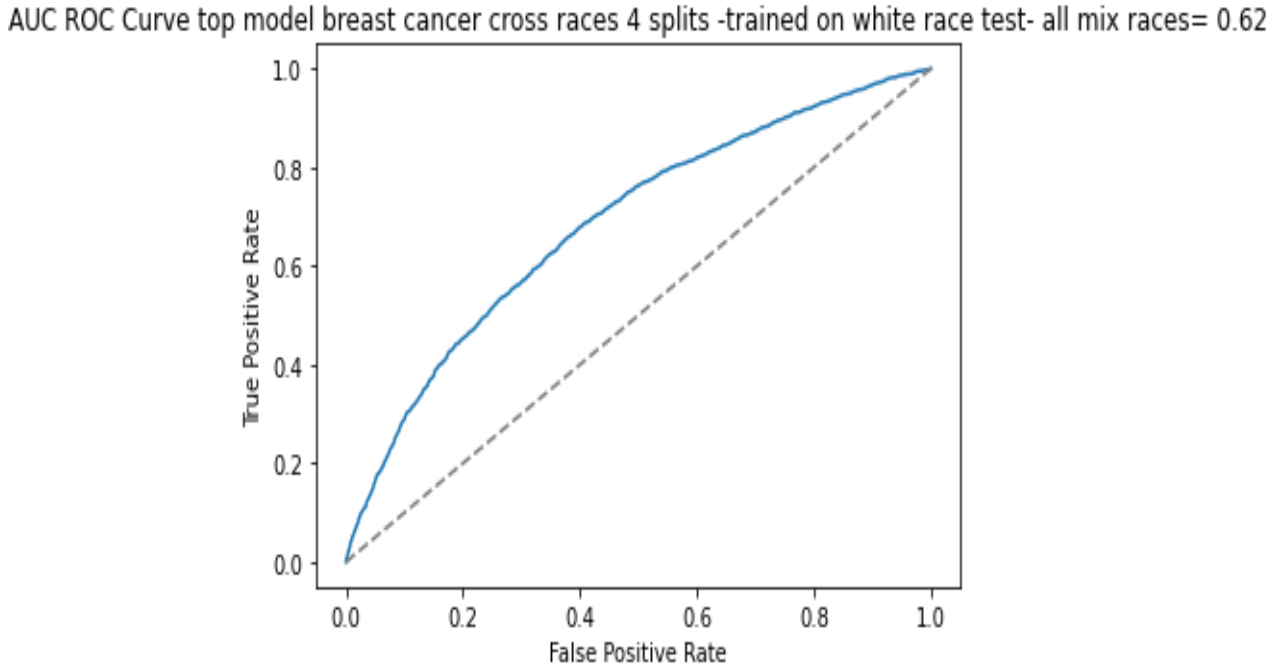

Figure 32: ROC curve of model trained on the white subpopulation and tested on all the combined races, including black, white and other races, with an AUC of 0.62.

remaining 20% unseen split of the white, black, other, and mixed races, are shown in Figure

29,Figure 30,Figure 31and Figure 32 respectively.

Similarly, the ML model trained on 80% of the black or African American subgroup was evaluated by testing its performance on the remaining 20% of the black split, white split, other race subgroups, and the combined dataset of white, black, and other races. The ROC curves of the model, tested on different race subpopulations, are presented in Figure 33, Figure 34, Figure 35, ,Figure 36 along with other evaluation metrics shown in Table 8.

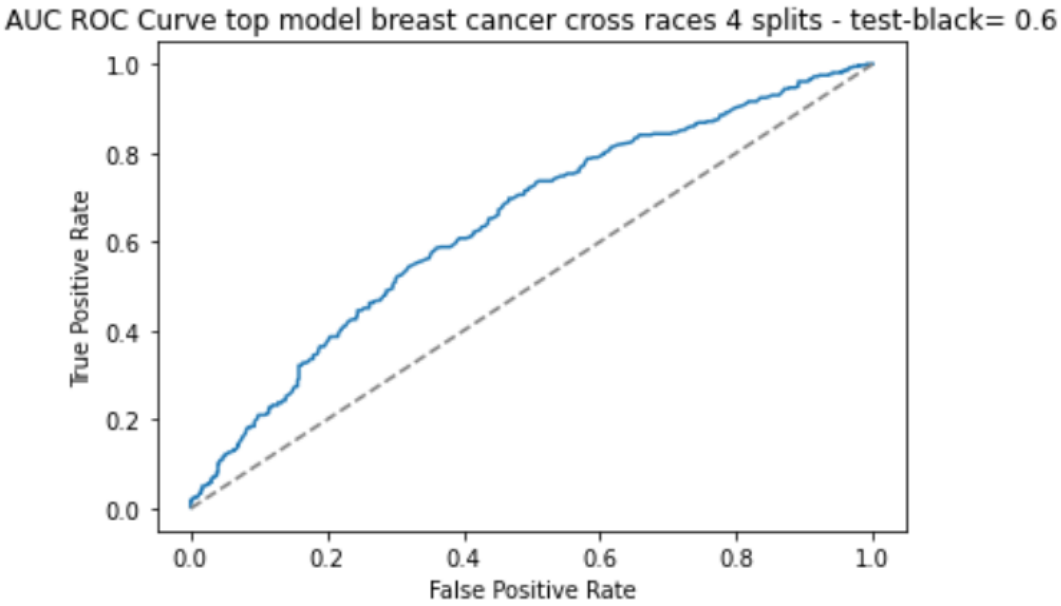

Figure 33: ROC curve of model, trained on 80% of the black subpopulation and tested on the remaining 20% test split of the black subpopulation, with an AUC of 0.60.

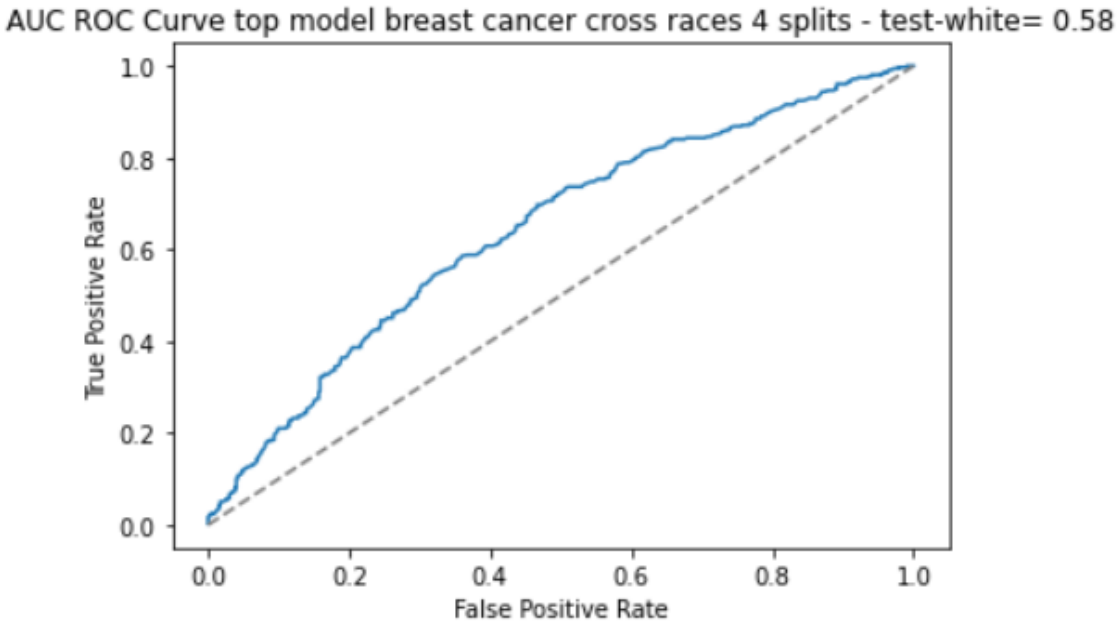

Figure 34: ROC curve of model trained on the black subpopulation and tested on white race, with an AUC of 0.58.

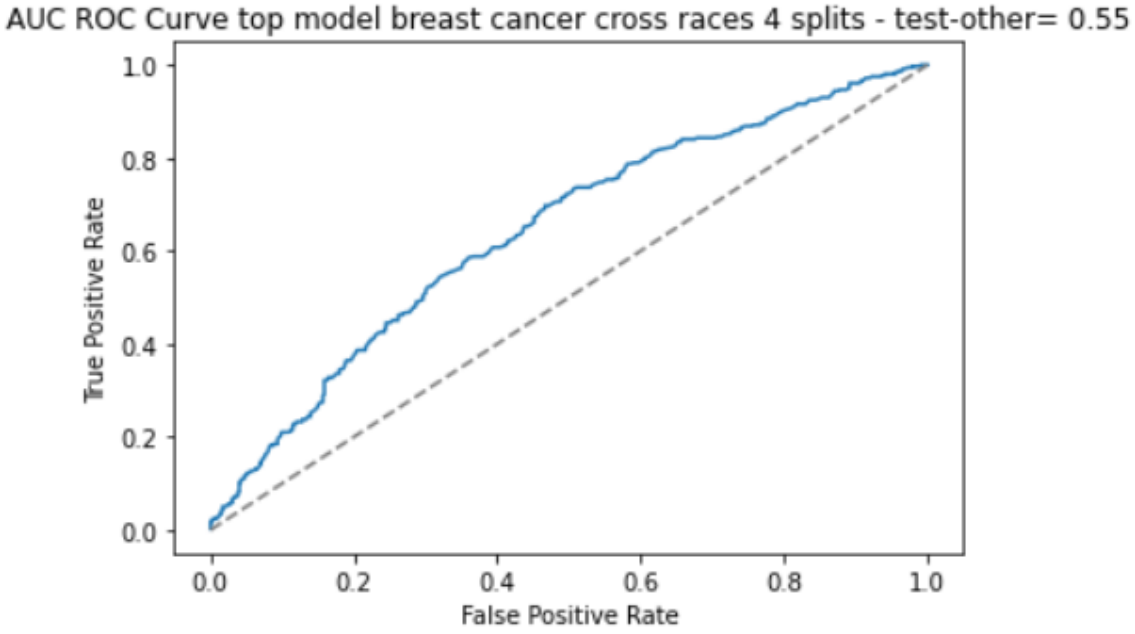

Figure 35: ROC curve of model trained on the black subpopulation and tested on other races, with an AUC of 0.55.

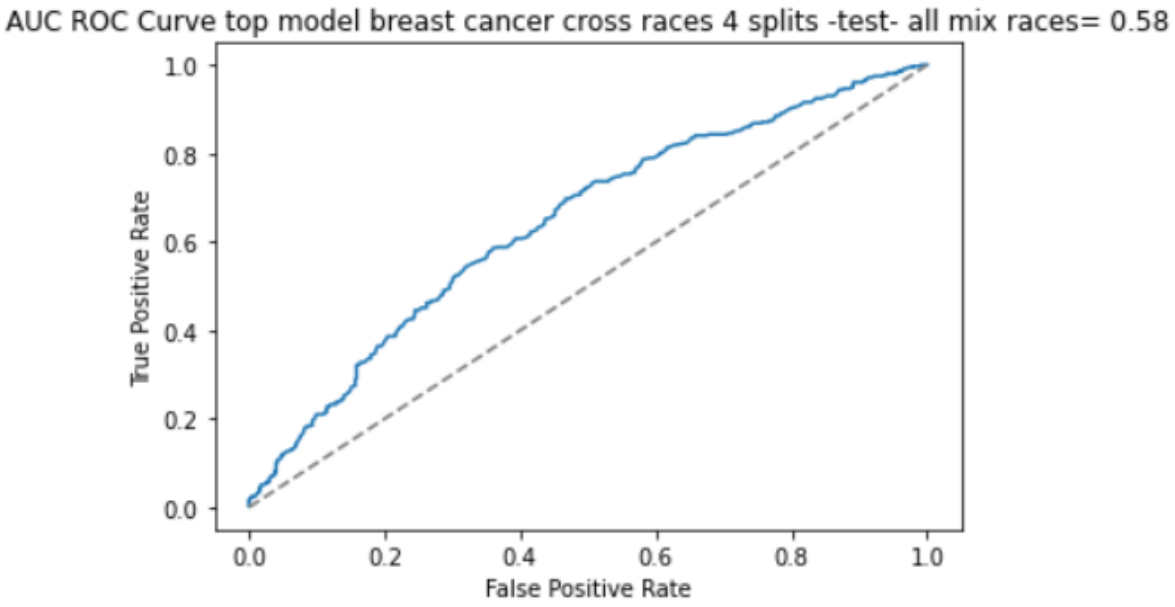

Figure 36: ROC curve of model trained on the black subpopulation and tested all the mixed races, including black, white and other races, with an AUC of 0.58.

Based on the performance analysis of these two trained models on different race subgroups, a notable trend can be observed. Among all the calculated AUC values for different subgroups, the model trained on the white subgroup and tested on the remaining 20% split of the white subgroup exhibited the best performance. During each run of model development on the white split of data, a total of 6,684 cases were used for training and 1,671 cases were used for testing. In contrast, the number of cases in the black subgroup was significantly lower, with a total of 1,100 cases available for each run, of which only 880 cases could be used for training. To further evaluate the differences in model performance, p-values were calculated comparing the AUC values tested on the white subgroup with the AUC values tested on different subgroups. The results showed a significant difference in the model's performance on the white split of the data compared to other racial subgroups. This significant difference could be

attributed to the larger number of cases in the white subgroup, which may contribute to improved model performance. However, when considering a more diverse population, the model's performance decreases significantly, indicating that this approach may be more effective for the white race.

Considering the limitation of the number of cases in the black subgroup, the model trained on the black race exhibited a relatively consistent performance on all other subgroups of races. This may be attributed to the higher diversity observed in the germline DNA of the black race, which could increase the generalizability of the model. Surprisingly the model's performance on white split of population was better than black population which also supports the more diversity within the black group. However, it is important to note that the overall model's performance was lower compared to the model trained on the white race.

To further investigate the potential for model improvement, it would be beneficial to include more data points from the black subgroup. By increasing the sample size and diversity within the black race, it is possible to enhance the model's performance and explore its effectiveness in predicting breast cancer risk within this population.

6.2.c Results Discussion- Breast Cancer

The developed ML model for predicting breast cancer based on CSLV features derived from germline DNA introduces a new approach to breast cancer risk determination. By incorporating machine learning algorithms and considering the epistatic effects of the genome in a nonlinear way on data derived from All of Us study, this model overcomes the limitations of previous models, particularly in terms of generalizability. The developed ML model based on 88 numbers achieved an AUC of 0.70, surpassing other risk determination models for breast cancer,

including the Tyree-Cusick model with an AUC of 0.64 and the polygenic risk score computed from linear combinations of SNPs with an AUC of 0.630. [4,139] This indicates that the developed model has better predictive accuracy and performs more effectively in determining the risk of breast cancer compared to these existing models.

In this approach, 88 numbers were used to characterize each genome by splitting the 22 chromosomes into four equal parts. Further improvements to the model can be explored by splitting the chromosomes into finer segments and increasing the training time. However, understanding the underlying mechanisms of the model can be challenging. Analysis of feature importance and Shapley values reveal that there is no single chromosome region that significantly contributes to the model's predictions.

Another limitation of the proposed model is that it focuses solely on germline DNA and does not incorporate environmental factors in risk determination. However, despite these limitations, the developed model based on CSLV exhibits effective prediction of breast cancer occurrence within one of the most diverse datasets available. In fact, the likelihood of a woman developing breast cancer is 9 times higher if she scores in the top quantile according to the developed model compared to a woman who is ranked in the bottom quantile. This demonstrates the strong predictive power of the model in identifying individuals at higher risk for breast cancer.

6.2.d Genetic Risk Score Model for Determining Risk of Colorectal Cancer

Based on the promising results obtained from the breast cancer risk estimate model, we proceeded to explore the application of a similar strategy for developing a risk assessment model for colorectal cancer. In this sub-chapter, our objective is to investigate the feasibility of predicting the occurrence of colorectal cancer in participants enrolled in the All of Us study. By leveraging the available data and applying machine learning techniques, we aimed to develop a model that could accurately classify individuals as either at high risk or low risk for developing colorectal cancer. [5,21,135]

Colorectal cancer is a prevalent form of cancer diagnosed worldwide and is recognized as the second leading cause of cancer-related deaths in the United States. Since the mid-1990s, there has been a decline in the incidence of colorectal cancer, which can be attributed to the widespread implementation of regular screening programs. However, the incidence and mortality of this cancer is not even across US subpopulations; there is a marked difference in CRC incidences by race and ethnicity. In particular, Black Americans have the highest mortality and incidence rate and Native Americans been ranked second.[142] It's been shown that high screening utilization can potentially eliminate the mortality disparity among Black Americans. The current risk assessment models that have been used to identify the high-risk population for colorectal cancer contain two sections, non-modifiable risk factors and modifiable risk factors. Non-modifiable risk factors include genetic or heritable susceptibility and modifiable risk factors are environmental risk factors such as diet, tobacco use etc. The genetic susceptibility of colorectal cancer (CRC) is strongly influenced by an individual's family history of cancer and the

age at which those cancers occurred. Screening recommendations are often based on these assessments, with individuals at higher risk advised to undergo screening at an earlier age, typically above 50. Indeed, studies have shown that Black and Hispanic Americans with a family history of colorectal cancer (CRC) have lower rates of participation in screening compared to other racial/ethnic groups. Additionally, there is a disparity in knowledge of parental cancer history between Black and White Americans, with Black Americans being less likely to have this information available. Furthermore, the communication of colonic polyp findings within families is less likely to occur among Black Americans. This lack of information or lack of transmittal of information from Black Americans can result in individuals being categorized as average-risk for screening instead of being identified as having a positive family history that warrants earlier screening. [143–146] Current risk assessment tools often do not adequately consider the contribution of racial/ethnic ancestry to heritable susceptibility factors. The limitations mentioned underscore the importance of developing an improved and more accurate risk assessment model for colorectal cancer that adequately incorporates inheritance factors.

Currently there are few colorectal risk estimates models that include germline DNA analysis in their risk estimate calculation, in one study 115 risk variants derived form GWAS and they were used to develop polygenic risk model for East Asians population. The performance of the developed logistic regression model was analyzed by calculating AUC = 0.63. [34] In another study, nine populations of European descent were studied for model development. Binary regression model was developed with an AUC = 0.59 which is based on combined effect of age, gender, family history and genotype of 10 susceptibility loci. [147] Polygenic risk scores have

gained a lot of attention recently new models may continue to emerge as research progresses, however the linear analysis of genetic variants and lack of diversity of datapoints are the main limitations of the current developed models which results in a lower predictability, the maximum AUC that has been achieved is 0.63. [148,149]

Developing a polygenic risk model that incorporates germline DNA analysis in a nonlinear manner based on diverse data points holds great potential for improving the accuracy of risk assessment tools for colorectal cancer and addressing the aforementioned limitations. By leveraging advanced genetic analysis techniques, such as chromosomal structural variations, we can better capture the complex genetic susceptibility to colorectal cancer.

For developing a personalized risk assessment model based on CSLV, we utilized the calculated CSLVs number (section 6.1.d) for each patient in all of us controlled tier V6 dataset and used it as the basis for developing the ML model. We were interested in identifying participants who have been diagnosed with malignant neoplasm or tumor of colon/ rectum and have had microarray genetic data. The control group for this study was constructed by excluding men and women with any type of cancer diagnosis, resulting in a cohort of 78,196 cancer-free women and men in all of us dataset-controlled tier V6. For positive class we identified 1401 cases, and 78,196 cases were identified in control class that have had genetic microarray data. Table 9 represents the racial and sex distribution in both classes of the model that were used for model development.

Class	Race Distribution					Sex			Total
	White	Black	Asian/Middle Eastern	Other	Unknown	Female	Male	Unknown	
Positive (Malignant Colorectal Tumor)	1022	136	25	22	196	728	656	17	1401
Negative (Control)	27949	24810	3116	2210	20111	44518	33678	0	78196

Table 9:Racial and sex distribution of participants in positive & negative class of risk assessment model for colorectal cancer. The positive class includes men and women with a diagnosis of malignant colon-rectum tumor, while the control group comprises cancer-free men and women.

For model development in this section, we employed the same strategy as in the previous section. To achieve a balanced distribution, we selected an under-sampled age-matched group of participants in the control class, ensuring a 40:60 ratio between the positive and negative classes in each run. This approach allows us to account for the imbalanced nature of the dataset and improve the performance and generalization of the developed models.

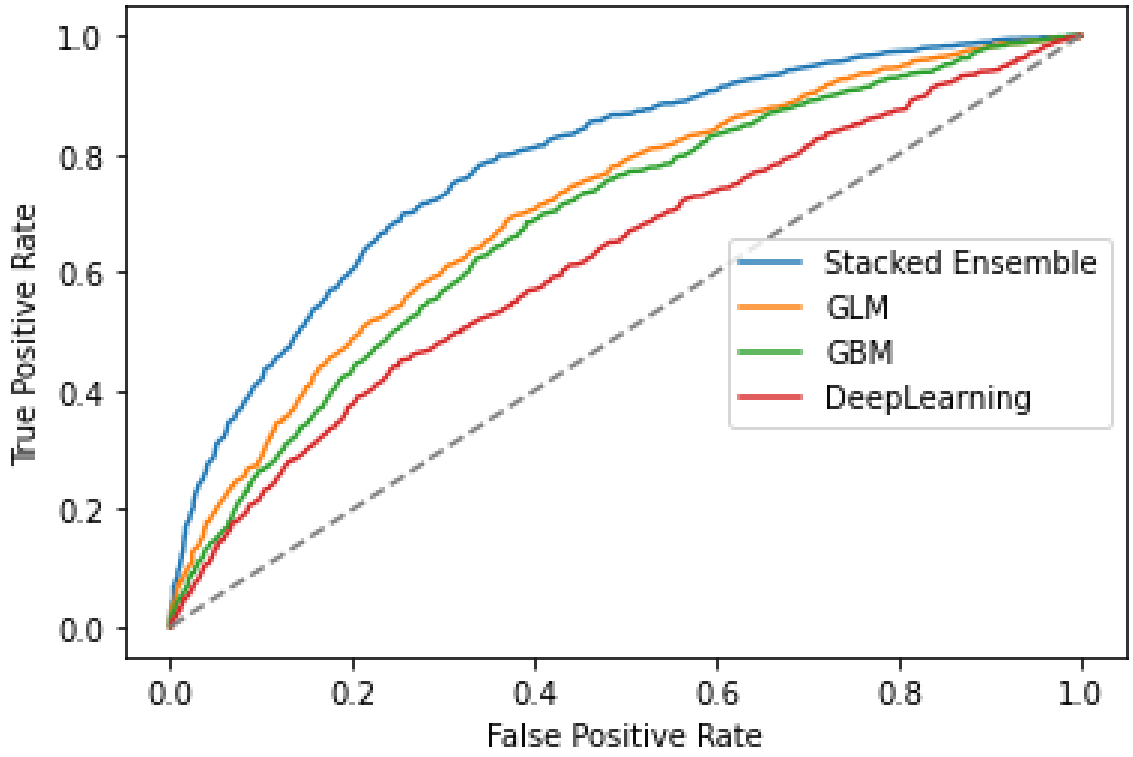

Figure 37: ROC curves of top 4 classifiers for predicting colorectal cancer. Models were trained on 80% of data and tested on the remaining 20% split. The stacked ensemble model performed the best, followed by the GLM and GBM models. The performance of the deep learning models varied and was influenced by the allocated training time, but none of them ranked as the stacked ensemble model in terms of AUC.

Based on the successful results obtained from breast cancer risk determination, we applied the same strategy for model development to predict colorectal cancer. In this approach, we utilized 88 numbers derived from splitting each of the 22 chromosomes into 4 equal segments as predictive features. These numbers were fed into the H2O AutoML function to develop machine learning models for predicting colorectal cancer. The same criteria as the breast cancer risk determination model were implemented for model development.

Figure 37 displays the ROC curve of the top 4 classifiers that were trained on 80% of the dataset and tested against the remaining 20%. Table 10 presents additional evaluation criteria such as accuracy, F1 score, and precision for these models at various stages of development,

Classification Model	AUC- Train (SD)	Accuracy-Train % (SD)	AUC-Validation % (SD)	Accuracy-Validation % (SD)	AUC-Test (SD)	Accuracy-Test % (SD)	F1 Score -Test % (SD)
Stacked Ensemble	0.84 (0.05)	79% (3.5)	0.65 (0.02)	64% (0.3)	0.68 (0.02)	67% (4.2)	0.63 (0.04)
GLM	0.7 (0.02)	65.5 %(0.7)	0.67 (0.02)	64 %(0.7)	0.68 (0.02)	67% (0.01)	0.60 (0.01)

AUC: Area Under Curve; GLM: Generalized Linear Model; Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$; Precision = $\frac{TP}{TP+FP}$; F1 Score = $2 \times \left(\frac{Precesion \times Recall}{Precision+Recall}\right)$
Performance metrics were reported as the average of 5 runs along with standard deviation.

Table 10: Performance metrics of the top 4 machine learning models for predicting colorectal cancer. Performance metrics were reported as the average of 8 runs along with standard deviation.

including training, cross-validation, and testing. These metrics provide a comprehensive analysis of the models' performance and their ability to accurately predict colorectal cancer.

To evaluate the performance of the trained model on unseen data, each case in the test set was assigned a score based on the model's predictions, similar to odds ratio calculation for breast cancer model. The accuracy of the model was then assessed by ranking the scores according to their scores and analyzing its performance across each subsections. A higher score indicated a higher probability of having colorectal cancer. Figure 39 presents the odds ratios calculated based on predictions done by top model, stacked ensemble, for each respective 1/25th subsection.

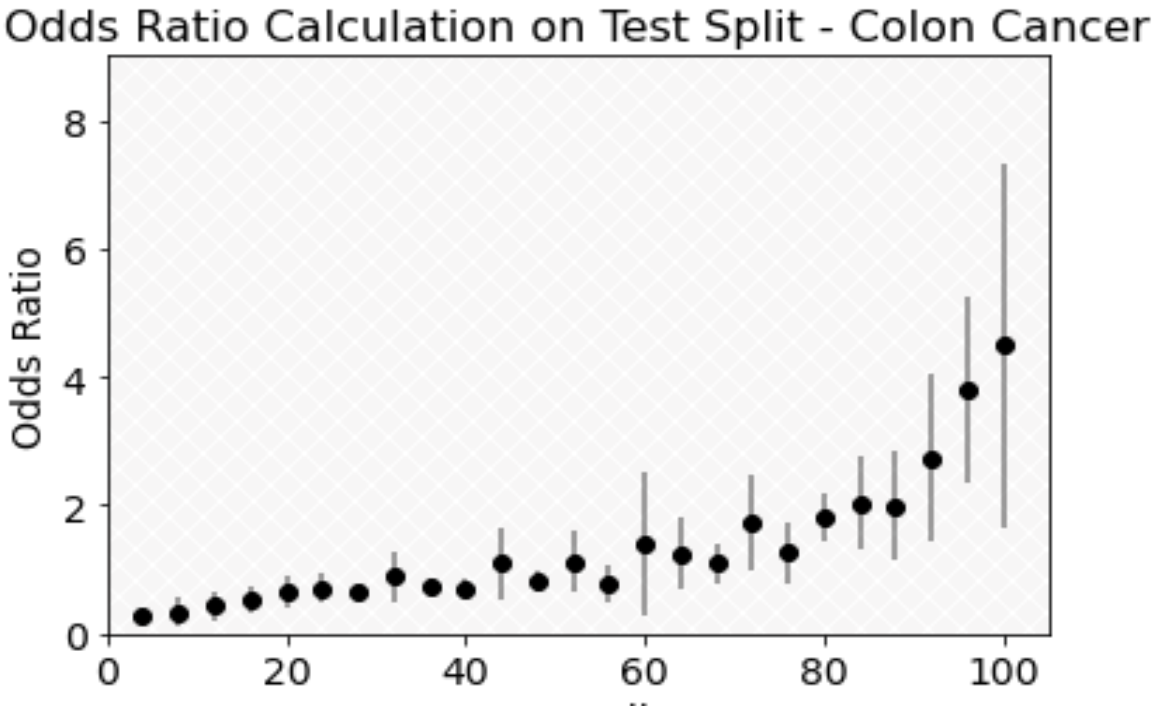

Figure 38: This figure shows that cases ranked higher by the ML model, stacked ensemble, are significantly more likely to have colorectal cancer. This trained model ranked all 700 new cases in the test split based on their likelihood of having a colorectal cancer, based solely on germline DNA CSLV data. This ranking was then split into 25 equal portions, each with about 28 cases. This plot shows the odds ratio of each of the 25 equal partitions along with 95% confidence intervals.

Figure 38, demonstrates the increased likelihood of having colorectal cancer associated with higher scores. People who were ranked in the top 10% by the model are 12 times more likely to develop a colorectal cancer compared with the bottom 10% group. These results supply valuable insights into the predictive ability of the model and its effectiveness in identifying individuals at a higher risk of colorectal cancer.

Figure 39,Figure 40 and Figure 41 showcase the variable importance analysis performed on different models within the H2O AutoML framework. Specifically, the GLM model's variable importance is displayed, providing insights into the significance of each variable. The SHAP plot represents the contribution of features in the best decision tree model, XGBoost, shedding light

on their impact on the predictions. Additionally, the variable importance heat map offers a comprehensive overview of the top models on the leaderboard, illustrating the relative importance of variables across multiple models. These figures were generated using the available functions in the H2O AutoML framework.

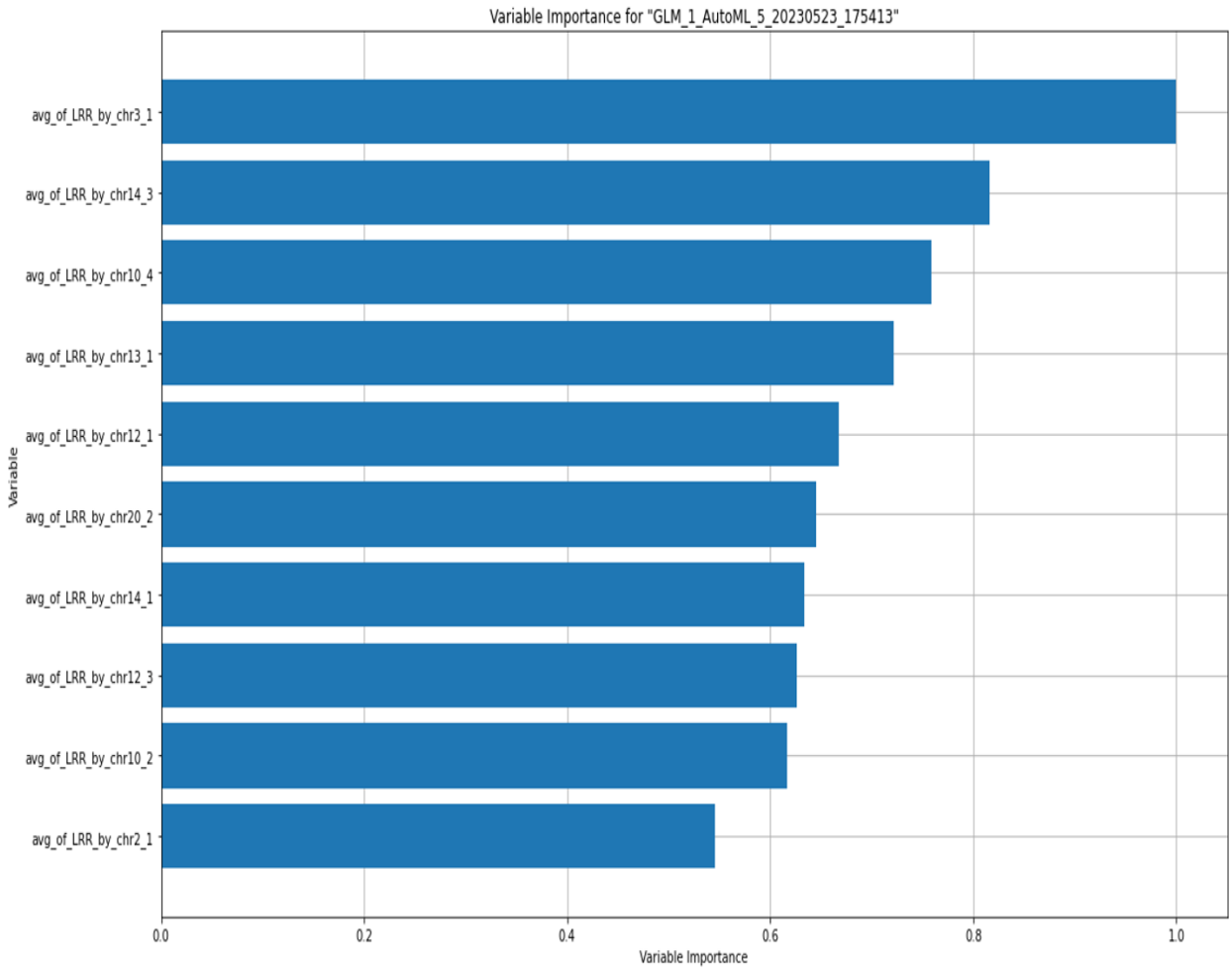

Figure 39: Relative importance of the most significant variables in the model. The variable importance was calculated based on GLM ML model for predicting colorectal cancer.

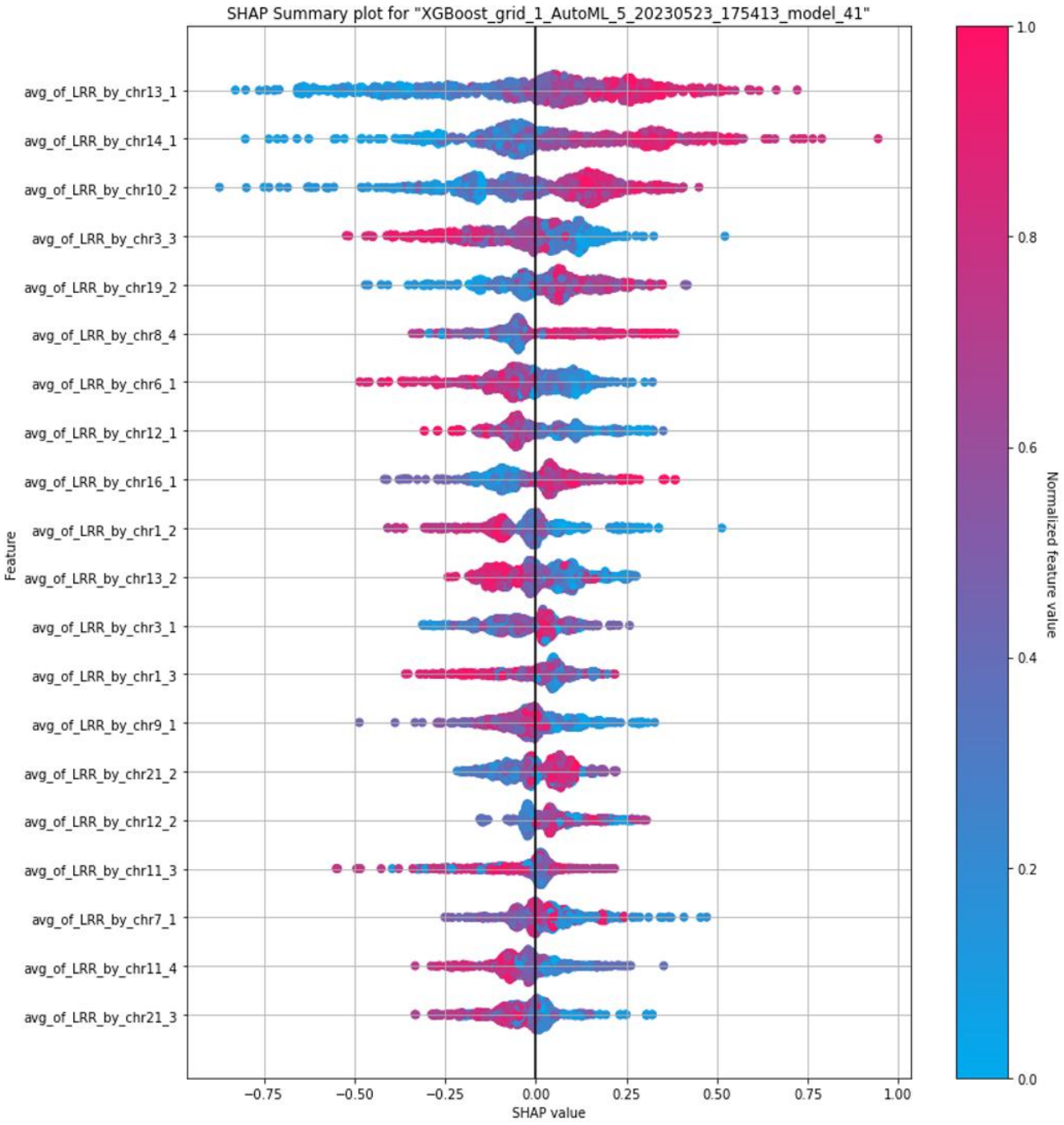

Figure 40: The Shapley additive explanations plot of GBM model for predicting breast cancer.

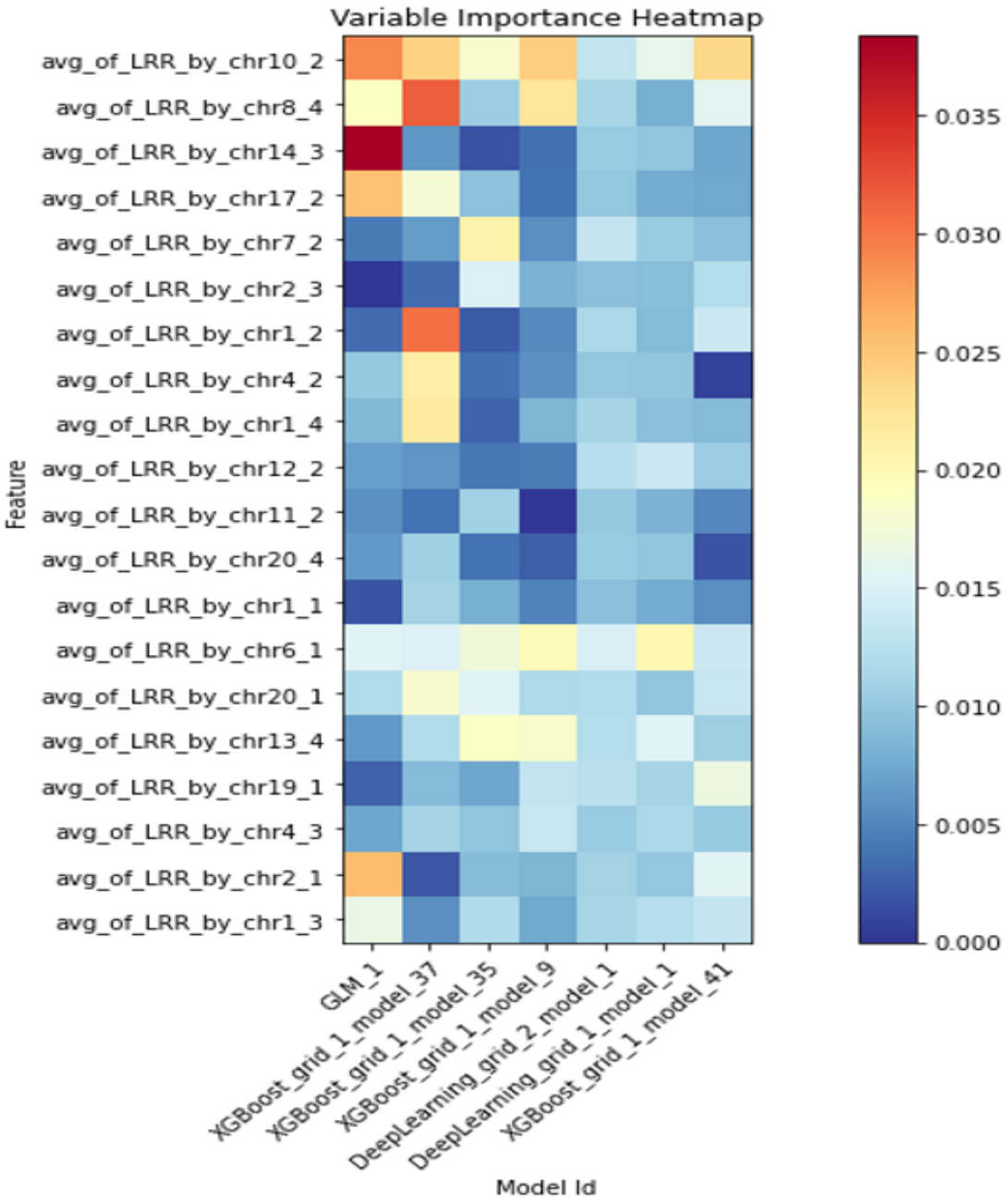

Figure 41: The variable importance heat map for the top machine learning models used in the prediction of colorectal cancer. The heat map supplies a visual representation of the relative importance of different variables across these models.

Similar to breast cancer model analysis, these analysis reveals that there is no specific chromosomal region that significantly outweighs others in terms of its contribution to the predictions made by the model.

Our developed personalized risk estimate model, which incorporates germline DNA analysis to analyze inheritance factors, overcomes the limitations of current risk assessment tools that rely on family history information. This is particularly beneficial for racial and ethnic groups who may have limited knowledge of their parental cancer history. [143]The model achieved an AUC of 0.68 on the unseen split of data, and individuals ranked in the top 10% by the model are 12 times more likely to develop colorectal cancer compared to the bottom 10% group. This outperforms current polygenic risk estimate models with an AUC of 0.63. [34] By incorporating inheritance factors through our novel approach of Chromosomal Scale Length Variability (CSLV), we significantly improve the accuracy of risk estimation for colon cancer. By considering the structural variations and genetic information at the chromosomal level, our model provides a more comprehensive and accurate assessment of individual risk for colon cancer, regardless of the availability of family history data. This ensures that individuals from all backgrounds can benefit from personalized risk estimation and receive appropriate preventive measures and screenings.

6.2.e Genetic Risk Score Model for Determining Risk of Oral Cavity Cancer

In the first aim of this study, our focus was on developing a machine learning model for predicting local regional recurrence of oral tongue squamous cell carcinoma based on sociodemographic and clinical variables. Through the analysis of relative feature importance, we found that several factors, including the number of prior tumors, histology, tumor site, and age, were among the features that exhibited the highest importance in the predictive model. We hypothesized that there will be genetic predisposition associated with tumor occurrence. For this reason, we investigated developing a personalized predictive model purely based on CSLV, to estimate the risk of oral cavity cancer for individuals.

In the past decade, there has been significant research and development of risk estimation models for oral cavity cancers. This area of study has garnered considerable attention due to the rising incidence rates observed across all age groups worldwide. [150–153] Particularly concerning is the increased incidence and mortality rates among young adults below the age of 40 in the European Union and the United States. [154,155]

Unfortunately, the current detection of oral cancer often occurs at advanced stages, leading to high morbidity and mortality rates. Late-stage diagnosis is a significant factor contributing to these outcomes. However, there is hope for improved survival rates, with projections suggesting that implementing reliable risk assessment and screening methods in clinical settings could increase survival rates to 80% to 90% . Early detection and intervention play a crucial role in improving patient outcomes. To facilitate early diagnosis, screening programs targeting high-risk individuals have been developed. These programs employ various techniques such as visual inspections, oral brush biopsy, toluidine blue staining,

chemiluminescence, or fluorescence imaging. These methods help identify potentially abnormal areas in the oral cavity, which can then be further investigated to determine the presence of oral cavity cancers. [156]

The current risk assessment models for oral cavity cancer take into account a range of risk factors to estimate an individual's likelihood of developing the disease. These factors include age, gender, tobacco and alcohol use, betel nut chewing, family history, human papillomavirus (HPV) infection status, and the presence of oral potentially malignant disorders. Nomograms, which are graphical representations of mathematical models, have also been utilized to assess an individual's risk based on these factors.

Identifying the high-risk group for oral cavity cancer is crucial, as early detection and intervention can significantly impact mortality rates. Sankaranarayanan et al, observed that screening using visual inspection conducted by trained health workers led to a remarkable 34% reduction in oral cancer mortality among individuals who were tobacco and/or alcohol users in the intervention group. [157] This highlights the effectiveness of targeted screening programs in identifying individuals at high risk and potentially preventing disease progression and related mortality.

Developing a more comprehensive risk assessment model with the aid of advanced machine learning algorithms can improve the accuracy of the current models significantly. There are several studies that have attempted to develop risk estimate models for oral cavity cancer by utilizing different ML algorithms. However similar to other developed risk estimate models for other types of cancer most of these studies have developed their models based on small population, usually white race, with lack of diversity in their study group. These models

are mainly based on clinical and sociodemographic variables and inheritance factors have not been investigated fully in the developed ML models.[158] As the genetic testing for oral cavity cancer cases is not a routine practice, developing a risk assessment model which includes genetic variables in their model has not been studied yet. [159]

Similar to the previous two ML models in this chapter, we utilized All of Us dataset for model development. Following the same data extraction strategy as previous models, we constructed a dataset consisting of two classes: the positive class and the control group. The positive class comprised participants who had malignant tumors in specific primary sites, including the floor of mouth, lip, oral cavity and pharynx, oropharynx, tonsil, and salivary gland. These primary sites were selected according to the primary sites that were used as local regional sites for developing a prediction model for local regional recurrence of OTSCC.
The control group had similar properties as the control group in the colorectal cancer risk estimate model. Table 11 provides a breakdown of the racial and sex distribution within the positive class and control group used in the construction of the risk estimate model for oral cavity cancer. Model development followed the same strategy as the two previous developed models, colorectal cancer and breast cancer.

Class	**Race distribution**					**Sex distribution**			**Total**
	White	Black	Asian/Middle Eastern	Other	Unknown	Female	Male	Unknown	
Positive (Malignant oral cavity tumor)	1085	63	12	18	110	577	695	16	1288
Negative (Control)	27949	24810	3116	2210	20111	44518	33676	-	78196

Table 11: Sociodemographic, racial and sex distribution within the positive class and control group used for constructing the risk estimate model for oral cavity cancer. The positive class includes men and women with a diagnosis of malignant oral cavity tumor, while the control group comprises cancer-free men and women.

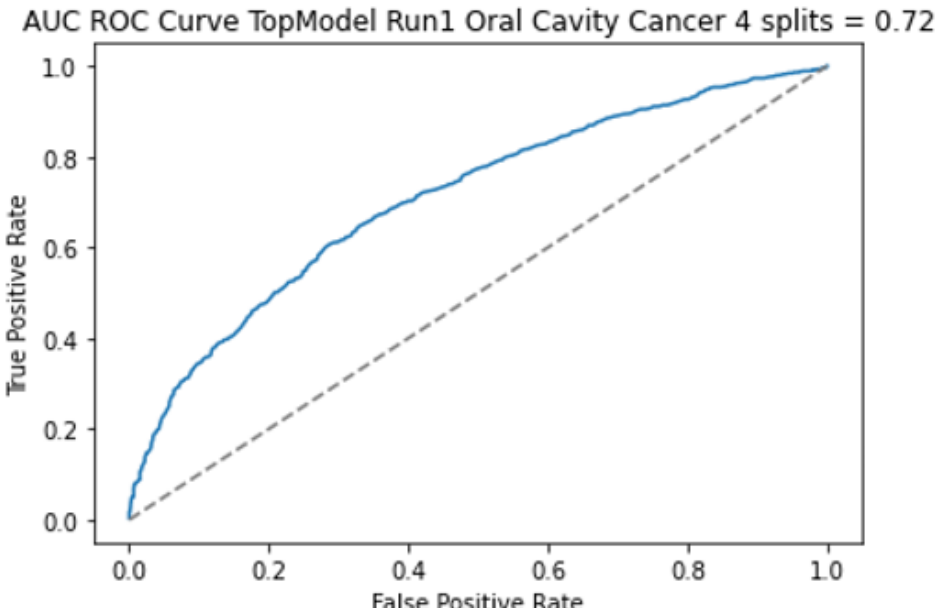

Figure 42: ROC curve of top classier, stacked ensemble model, for predicting oral cavity cancer. The performance of the trained model was evaluated on test split of data and AUC of the ROC curve was calculated. The stacked ensemble model has an AUC of 0.72.

We utilized 88 numbers, each corresponding to equal splits of each chromosome into 4 segments, to build the ML model using the H2O AutoML function. Stacked ensemble model was ranked the best classifier model during the training, cross validation, and test split. Figure 42 displays the ROC curve of the trained stacked ensemble model which was tested on unseen

Classification Model	AUC-Validation % (SD)	Accuracy-Validation % (SD)	AUC-Test (SD)	Accuracy-Test % (SD)	F1 Score -Test % (SD)
Stacked Ensemble	0.69 (0.01)	67% (1.01)	0.69 (0.02)	68% (1.0)	0.60 (0.01)
GLM	0.69 (0.01)	66% (1.2)	0.68 (0.03)	67% (1.3)	0.61 (0.03)
GBM	0.57 (0.01)	61% (1.02)	0.58 (0.02)	60% (1.01)	0.58 (0.002)
Deep Learning	0.61 (0.04)	63% (3.5)	0.62 (0.04)	63% (3.1)	0.59 (0.01)
Performance metrices were reported as the average of 5 runs					

Table 12: Performance metrics of top four ML models for predicting oral cavity cancer.

split of the data with an AUC value. Table 12 and Figure 43 show the evaluation criteria of the top four ML models at different stages of the model development process along with ROC curve of top four classifiers.

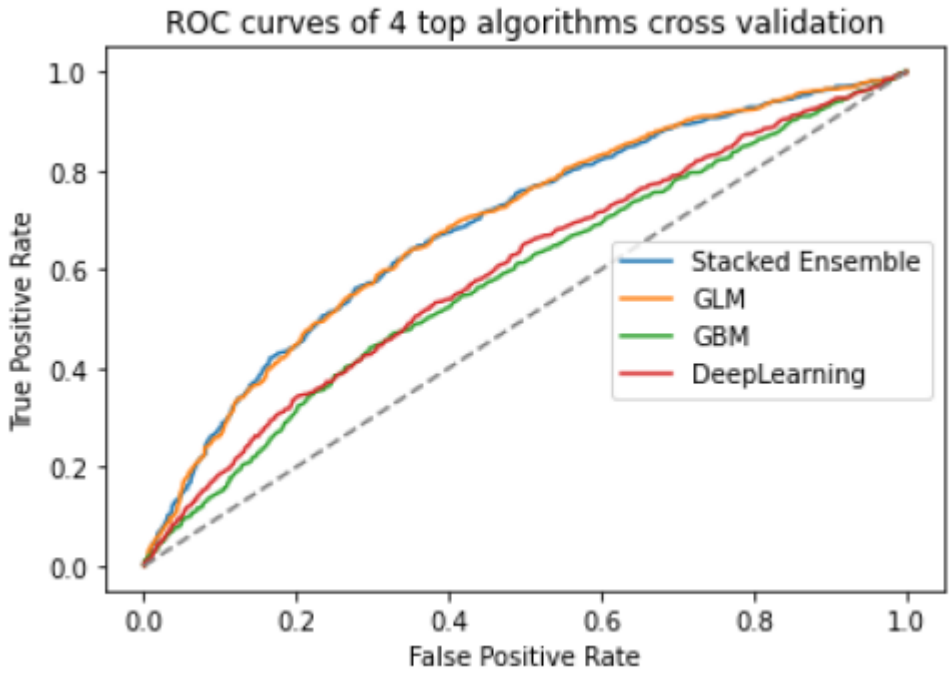

Figure 43 : ROC curve of top 4 ML developed for predicting oral cavity cancer, these plots were generated by testing the trained model on unseen split of data.

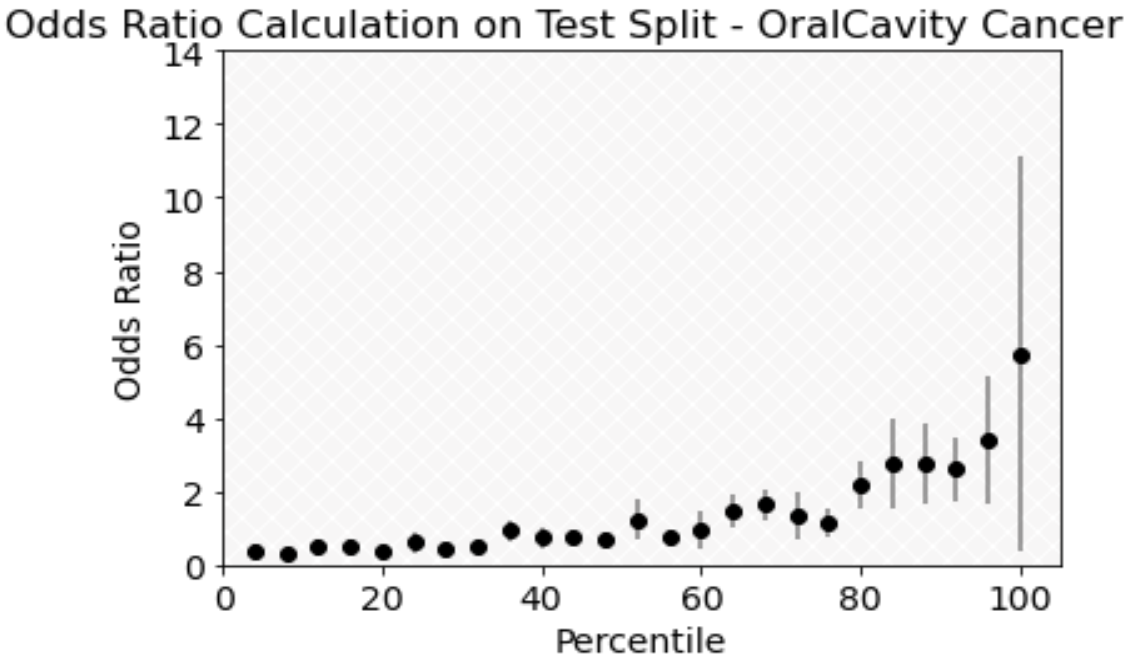

Figure 44: This figure shows that participants ranked higher by the predictive model are significantly more likely to have oral cavity cancer. The predictive model ranked all 645 women and men in the test split of dataset based on their likelihood of having oral cavity cancer, based solely on germ line DNA data. This ranking was then split into 25 equal partitions, each with about 129 participants. This plot shows the odds ratio (relative to entire group) of each of the 25 equal partitions along with the 95% confidence intervals.

By obtaining the scores of the trained model on the unseen test split of the data, we calculated the odds ratio. We ranked men and women on the test split based on their score, divided the rankings into 5 equal quantiles, and calculated the odds ratio relative to the entire group with a 95% confidence interval. The results, presented in Table 13, show that individuals who were ranked higher by the predictive model are 7 times more likely to have oral cavity cancer compared with the bottom quintal. These rankings, determined by the stacked ensemble model and are based solely on germline DNA. Figure 44 displays the average of odds ratio for each of the 25 equal partitions along with their standard deviations of 5 runs.

Quintile	Number of participants with oral cavity cancer	Number of participants without oral cavity cancer	Total number of participants in each quintile	Odds Ratio	95% confidence interval
5	91	38	129	2.86	(2.38-3.33)
4	67	62	129	1.45	(1.38-1.52)
3	58	71	129	0.85	(0.68-1.0)
2	36	93	129	0.60	(0.53-0.66)
1	27	102	129	0.41	(0.29-0.53)

Table 13: The participants in the unseen split of data were ranked by score from lowest to highest by the top trained model into five equal quintiles. This table presents the number of participants with and without oral cavity cancer in each quintile along with the odds ratio compared to the entire group and the 95% confidence interval for the odds ratio.

Figure 45 displays the variable importance for the GLM model, highlighting the relative importance of each variable in predicting oral cavity cancer risk. Figure 46, presents the Shap summary plot for the top tree-based model, XGBoost, showing the impact of each variable on the model's predictions.

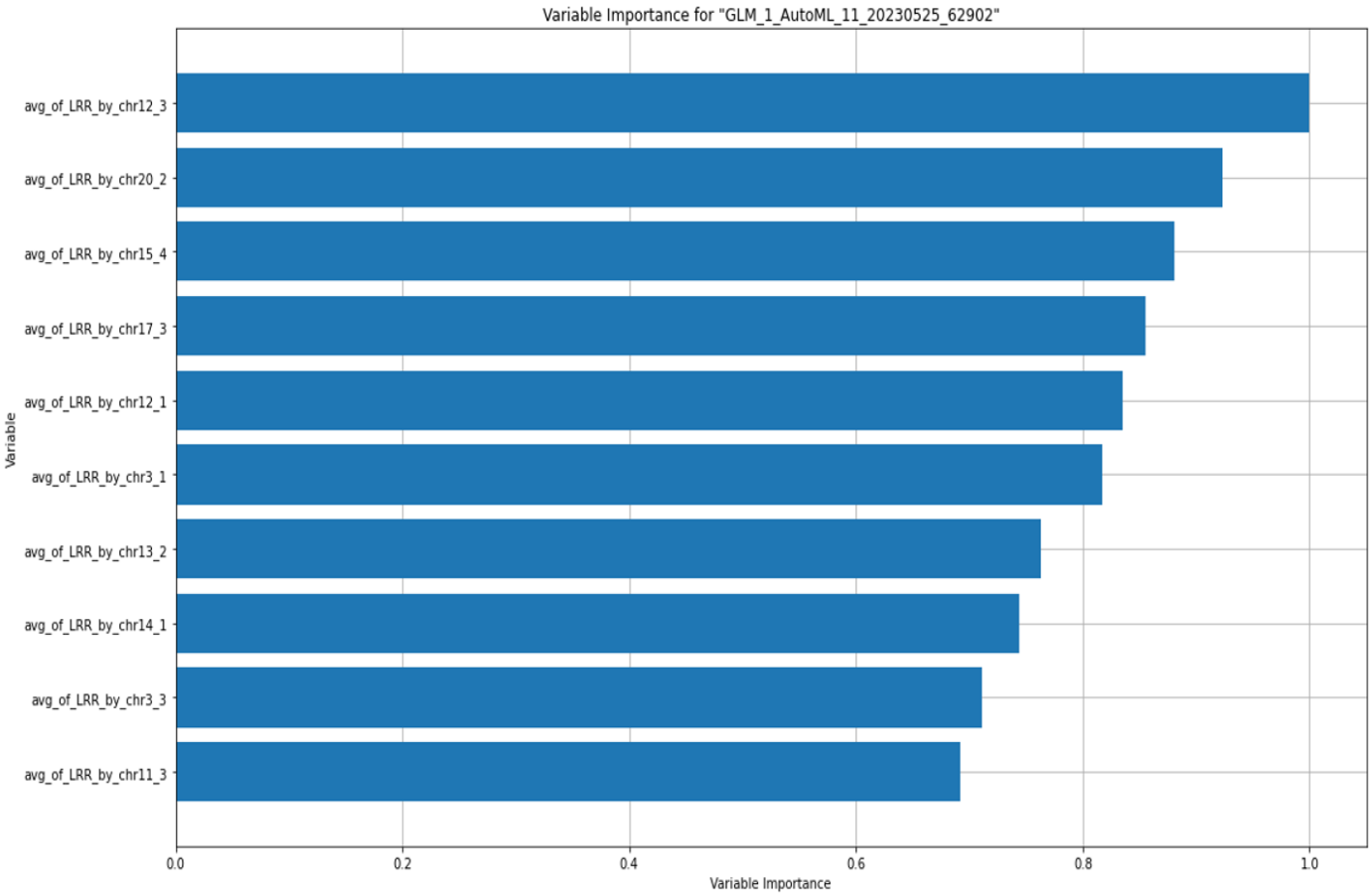

Figure 45: The variable importance plot on oral cavity risk assessment model. The GLM model was ranked second on the leaderboard and been used to show the relative importance of the most important variables in the model.

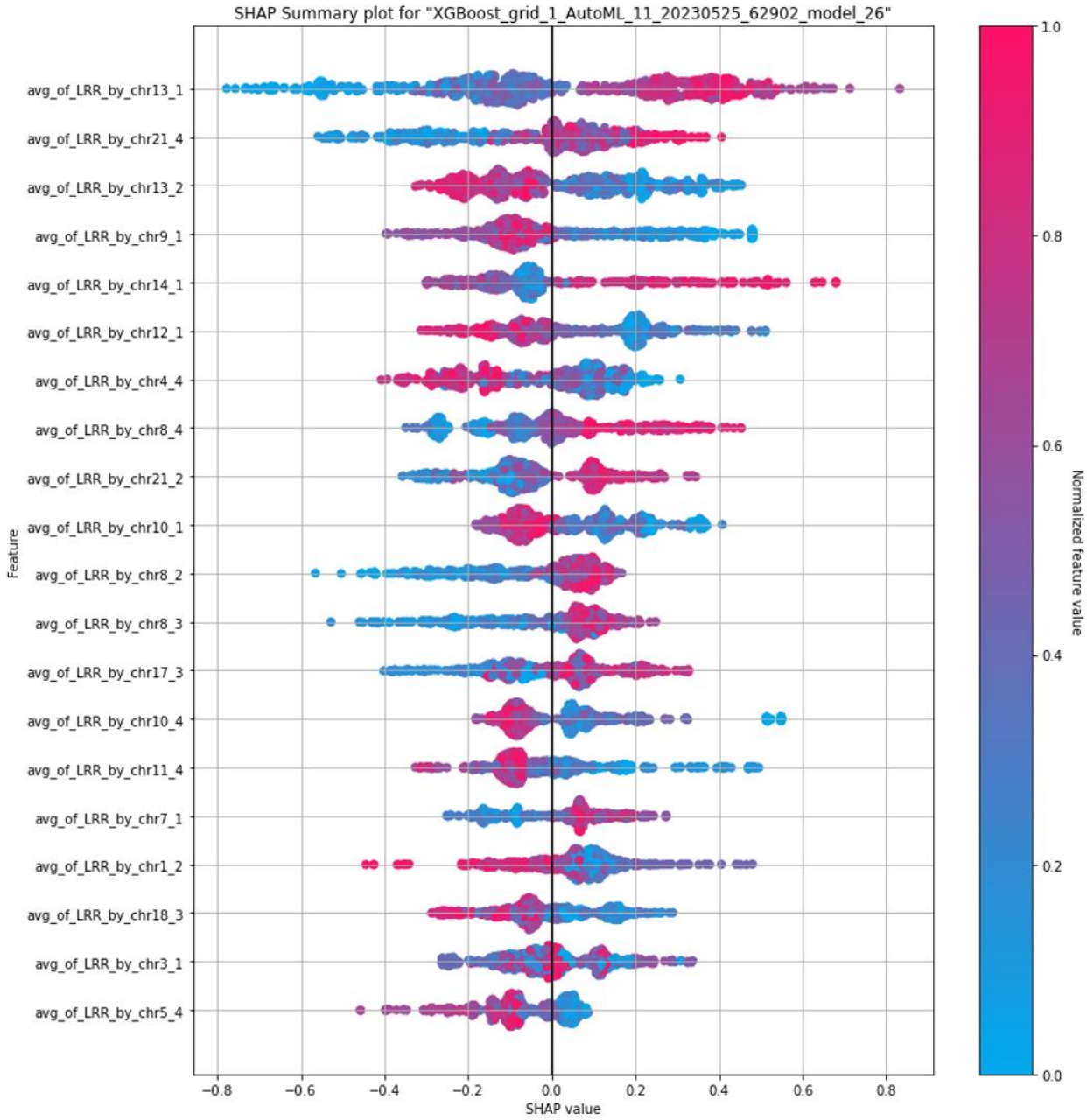

Figure 46: The Shapley additive explanations plot of XGBoost model for predicting oral cavity cancer.

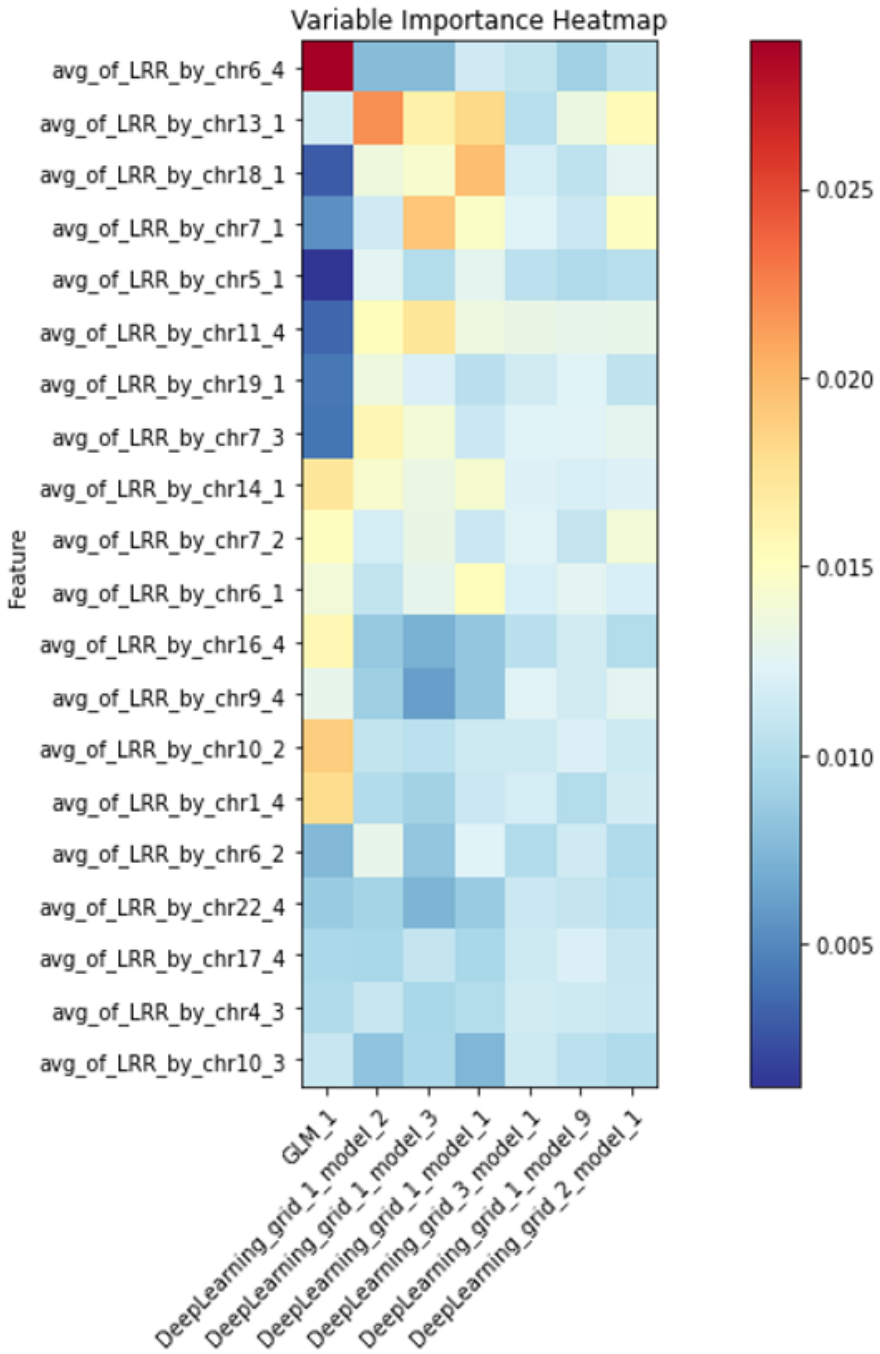

Figure 47: The variable importance heat map for the top machine learning models used in the prediction of oral cavity cancer. The heat map supplies a visual representation of the relative importance of different variables across these models on h2o automl function.

The variable importance heat map, Figure 47, illustrates the relative importance of variables across multiple models developed for predicting oral cavity cancer.

In recent years, there has been a significant increase in the incidence rate of oral cavity cancer, leading to a focus on identifying high-risk individuals through various assessment tools.[150,151] Several studies have utilized machine learning algorithms, such as artificial neural networks, logistic regression, decision trees, and Fuzzy regression, to develop risk assessment models for predicting oral cavity cancer. Among these models, the top models with the best performance were artificial neural network with 80% sensitivity and voting classifier with an AUC of 0.83. However, these models have limitations, first, as they are developed based on individuals' risk habits and demographic profiles, which inheritance factors had not been fully investigated in their model development. Second, the models were developed based on primarily datasets representing the white race and non-Hispanic and Latino ethnicity.[158–160]

Based on the obtained results, we developed a machine learning model using 88 numbers derived from an individual's germline DNA to predict the risk of oral cavity cancer. The top-performing model demonstrated an AUC of 0.70 and a precision of 85%, indicating promising predictive ability. Upon analyzing the SHAP plot and assessing the importance of different features in each run, we observed that no single feature consistently exhibited high predictability. This suggests that the combination of numbers derived from different splits of each chromosome contributed to the model's development. To further evaluate the model's performance, odds ratios were calculated to compare the likelihood of individuals ranked at the top versus those ranked at the bottom by the model.

To date, there have been no studies analyzing germline DNA in a nonlinear manner to assess the risk of oral cavity cancer in individuals. While the prediction ability of previously

discussed ML models may be higher than our model, we were able to develop a model using a dataset that exhibits greater diversity in its data points. We hypothesize that by integrating germline DNA analysis with other established risk factors, we can enhance the predictability and applicability of future models for determining the risk of oral cavity cancer. These developed models have the potential to identify high-risk patients and tailor screening and prevention strategies based on an individual's specific risk profile.

CHAPTER 7: CONCLUSION, LIMITATIONS & FUTURE DIRECTIONS

Cancer research plays a vital role in understanding and combating one of the most challenging health issues worldwide. Over the years, significant advancements have been made in cancer research, leading to groundbreaking discoveries and novel treatment options. Recent developments in this field have included targeted therapies, immunotherapies, and precision medicine approaches, which have shown promising results in specific cancer types. Additionally, advances in genomics and molecular biology have provided a deeper understanding of the underlying mechanisms driving cancer development and progression. Moreover, research efforts have also focused on early detection methods and personalized risk estimate models, enabling more tailored and effective treatments for individual patients. By the recent advancements in AI and big data in health care we believe human beings can make a significant contribution to early detection and advancing precision medicine.

This investigation demonstrates the potential of utilizing available data from large healthcare databases, such as SEER, UK Biobank, and All of Us datasets, to advance early detection of cancer through advanced machine learning algorithms. For the first aim of this investigation a unique algorithm developed to identify cases of cancer recurrence from the expansive and widely used SEER database. This lays the groundwork for future investigations into various challenges related to cancer recurrence. Using this novel framework, we developed one of the first machine learning-based classification models that can accurately predict 5- and 10-year recurrence in patients with oral tongue squamous cell carcinoma (OTSCC) and breast

cancer. Importantly, this model uses only commonly available demographic and clinical features, making it easily applicable in a clinical setting.

To incorporate hereditary factors in our risk estimation models and study their impact on cancer occurrence, two biobanks have been investigated: UK Biobank and All of Us dataset. These biobanks are invaluable resources, providing diversity and statistical power for our analysis. For this objective, we evaluated a novel approach to incorporate hereditary factors into our model and improve prediction accuracy using ML algorithms. Instead of relying on SNPs and the traditional method of calculating polygenic risk scores, we developed our model based on structural variations in germline DNA. These structural variations could be caused by insertions or deletions of segments genome.

First, we investigated whether we could use the CSLV values corresponding to each chromosome or segment of the chromosome to develop a model capable of accurately classifying breast cancer recurrence from non-recurrence cases based solely on these calculated numbers. We discovered that using the CSLV values, we were unable to accurately classify breast cancer recurrence cases from non-recurrence cases within the available data from the UK Biobank. This limitation may be due to the small number of recurrence cases identified in the dataset and the significant impact of treatment strategies and the stage of first diagnosis on the success of cancer remission. These factors influenced the ability to predict recurrence more than inheritance factors.

To explore the applicability of our unique approach in determining an individual's risk of developing a specific type of cancer, we tested our hypothesis on the recently released data from the All of Us study. The diversity of data points within the All of Us study provides valuable

input that can overcome the limitations of current risk estimate models, which are often developed based on specific race and ethnic groups. By incorporating a more diverse dataset, our approach has the potential to improve the accuracy and inclusivity of risk prediction models for various populations. By calculating CSLV values derived from copy number variations measurements within our associated workbench in the All of Us study, we were able to develop accurate risk estimate models for breast cancer, colorectal cancer, and oral cavity cancer. The models were trained exclusively using inheritance factors and were then tested on a separate, unseen split of data. By calculating Odds ratio and AUC values of the model performance on the test split, we demonstrated the potential of utilizing copy number variations in the form of chromosomal scale length variations to predict complex diseases, particularly cancer. This approach, combined with machine learning techniques, shows great promise in enhancing early detection and advancing precision medicine approaches for cancer. We believe that this approach can be extended to other complex diseases, enabling further investigation into the contributions of genetic variation to individual susceptibility.

However, there are some limitations associated with this study that can be enhanced and improved for future investigations. In our first aim of this study, we utilized avaibale data in SER dataset, due to the limitation of the collected data within the platform, some valubale histopathological and clinical features were not included in our features. Despite these constraints, we were able to develop a model with high predictability for the locoregional recurrence of OTSCC and local recurrence of malignant breast tumor. We believe that incorporating these site-specific variables along with other clinical and sociodemographic variables can only enhance the predictive power of these models.

In the second and third aim of our study, we focused solely on developing risk estimate models based on the calculated CSLV values within segments of the chromosome. However, it is important to note that this approach only accounts for structural variations within the size of the SNP region, specifically deletions and additions of segments of the chromosome that can be detected through CNV analysis. Other structural variations in the genome, such as translocations, cannot be accounted for in this approach.

Additionally, a common limitation associated with all machine learning models, including ours, is the challenge of interpreting the model's performance and understanding how predictions are made. Machine learning models are often considered black boxes, making it difficult to fully comprehend the factors driving their predictions. This limitation highlights the need for further research and development of interpretability techniques in machine learning to enhance the trust and acceptance of these models in the medical field.

We believe that that there is room for further improvement to enhance the applicability and accuracy of the models we have developed. Strategies such as feature engineering, using sub-chromosomes instead of complete chromosomes, and data augmentation can potentially improve the AUC values of the models. Additionally, incorporating SNP data and environmental factors into the models could have a significant impact on their accuracy.

Moreover, to enhance the applicability of the developed models, we can train the model on one dataset and evaluate its performance on a completely different dataset. Each analysis has been conducted on separate datasets and then tested on the same dataset. Future research could further refine this method to improve the predictive ability of the models.

In conclusion, we firmly believe that enhancing the accuracy of cancer risk estimate models and advancing precision medicine can effectively identify individuals at higher risk of cancer and enable personalized screening and intervention strategies. This approach will have a significant clinical impact on early cancer detection and ultimately lead to improve survival rates for individuals who are at risk. By tailoring healthcare interventions based on risk estimates, we can make substantial progress in the fight against cancer and improve patient outcomes.

REFERENCES

1. Toh C, Brody J. Chapter Applications of Machine Learning in Healthcare. In: ; 2021.
2. Etzioni R, Urban N, Ramsey S, et al. The case for early detection. *Nat Rev Cancer*. 2003;3(4):243-252. doi:10.1038/nrc1041
3. Ford PJ, Farah CS. Early detection and diagnosis of oral cancer: Strategies for improvement. *J Cancer Policy*. 2013;1(1-2). doi:10.1016/j.jcpo.2013.04.002
4. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet*. 2019;104(1):21-34. doi:10.1016/j.ajhg.2018.11.002
5. Ping J, Yang Y, Wen W, et al. Developing and validating polygenic risk scores for colorectal cancer risk prediction in East Asians. *Int J Cancer*. 2022;151(10):1726-1736. doi:10.1002/ijc.34194
6. Schatz MC, Langmead B. The DNA data deluge. *IEEE Spectr*. 2013;50(7):28-33. doi:10.1109/MSPEC.2013.6545119
7. Stadler ZK, Thom P, Robson ME, et al. Genome-wide association studies of cancer. *J Clin Oncol*. 2010;28(27):4255-4267. doi:10.1200/JCO.2009.25.7816
8. Yuan Q, Cai T, Hong C, et al. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal Cohort of Patients With Lung Cancer. *JAMA Netw open*. 2021;4(7):e2114723. doi:10.1001/jamanetworkopen.2021.14723
9. Galvan A, Ioannidis JPA, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet*. 2010;26(3):132-141. doi:10.1016/j.tig.2009.12.008
10. Khera A V., Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224. doi:10.1038/s41588-018-0183-z
11. Goode EL, Chenevix-Trench G, Song H, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet*. 2010;42(10):874-879. doi:10.1038/ng.668
12. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi:10.1038/nrg.2016.49
13. Ashley EA. Towards precision medicine. *Nat Rev Genet*. 2016;17(9):507-522. doi:10.1038/nrg.2016.86
14. Toh C, Brody JP. Genetic risk score for ovarian cancer based on chromosomal-scale length variation. *BioData Min*. 2021;14(1):18. doi:10.1186/s13040-021-00253-y
15. Toh C, Brody JP. Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. *Hum Genomics*. 2020;14(1):36. doi:10.1186/s40246-020-00288-y
16. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 2011;12(5):363-376. doi:10.1038/nrg2958

17. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet*. 2007;39(7S):S37-S42. doi:10.1038/ng2080
18. Shaikh TH. Copy Number Variation Disorders. *Curr Genet Med Rep*. 2017;5(4):183-190. doi:10.1007/s40142-017-0129-2
19. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet*. 2015;16(3):172-183. doi:10.1038/nrg3871
20. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013
21. Smith RA, Cokkinides V, von Eschenbach AC, et al. American Cancer Society Guidelines for the Early Detection of Cancer. *CA Cancer J Clin*. 2002;52(1):8-22. doi:10.3322/canjclin.52.1.8
22. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA Cancer J Clin*. 2021;71(1):7-33. doi:10.3322/CAAC.21654
23. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7-33. doi:10.3322/caac.21708
24. 'Richardson LC 'Dowling, N 'Henley, J. An Update on Cancer Deaths in the United States. Published February 2022. Accessed June 21, 2022. https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm#
25. The Genetics of Cancer. National Cancer Institute at the National Institutes of Health. Published August 17, 2022. Accessed August 9, 2023. https://www.cancer.gov/about-cancer/causes-prevention/genetics
26. Durazo A, Cameron LD. Representations of cancer recurrence risk, recurrence worry, and health-protective behaviours: an elaborated, systematic review. *Health Psychol Rev*. 2019;13(4):447-476. doi:10.1080/17437199.2019.1618725
27. Recurrent Cancer: When Cancer Comes Back. National Cancer Institute at the National Institutes of Health. Published December 2, 2020. Accessed June 22, 2022. https://www.cancer.gov/types/recurrent-cancer
28. Dowsett M, Sestak I, Regan MM, et al. JOURNAL OF CLINICAL ONCOLOGY Integration of Clinical Variables for the Prediction of Late Distant Recurrence in Patients With Estrogen Receptor-Positive Breast Cancer Treated With 5 Years of Endocrine Therapy: CTS5. *J Clin Oncol*. 2018;36:1941-1948. doi:10.1200/JCO
29. Keane M, Johnson G. Early diagnosis improves survival in colorectal cancer. *Practitioner*. 2012;256(1753):8-15.
30. Crosby D, Bhatia S, Brindle KM, et al. Early detection of cancer. *Science (80-)*. 2022;375(6586). doi:10.1126/science.aay9040
31. Yang P. *Epidemiology of Lung Cancer Prognosis: Quantity and Quality of Life*. Cancer Epidemiology; 2009.
32. Cancer - Screening and early detection. World Health Organization. Published May 15, 2010. Accessed June 22, 2022. https://www.who.int/europe/news-room/fact-sheets/item/cancer-screening-and-early-detection-of-cancer
33. Kalager Mette, Bretthauer Michael. Improving cancer screening programs. *Science (80-)*. 2020;367(6474):143-144.
34. Ping J, Yang Y, Wen W, et al. Developing and validating polygenic risk scores for

colorectal cancer risk prediction in East Asians. *Int J Cancer*. 2022;151(10):1726-1736. doi:10.1002/ijc.34194
35. Gram EG, Siersma V, Brodersen JB. Long-term psychosocial consequences of false-positive screening mammography: a cohort study with follow-up of 12–14 years in Denmark. *BMJ Open*. 2023;13(4):e072188. doi:10.1136/bmjopen-2023-072188
36. Favaretto M, De Clercq E, Schneble CO, Elger BS. What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLoS One*. 2020;15(2):e0228987. doi:10.1371/journal.pone.0228987
37. Marx V. The big challenges of big data. *Nature*. 2013;498(7453):255-260. doi:10.1038/498255a
38. Collins C, Dennehy D, Conboy K, Mikalef P. Artificial intelligence in information systems research: A systematic literature review and research agenda. *Int J Inf Manage*. 2021;60:102383. doi:10.1016/j.ijinfomgt.2021.102383
39. Alzubi J, Nayyar A, Kumar A. Machine Learning from Theory to Algorithms: An Overview. *J Phys Conf Ser*. 2018;1142:012012. doi:10.1088/1742-6596/1142/1/012012
40. Park C, Took CC, Seong JK. Machine learning in biomedical engineering. *Biomed Eng Lett*. 2018;8(1). doi:10.1007/s13534-018-0058-3
41. Kourou K, Exarchos KP, Papaloukas C, Sakaloglou P, Exarchos T, Fotiadis DI. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. *Comput Struct Biotechnol J*. 2021;19:5546-5555. doi:10.1016/j.csbj.2021.10.006
42. Cunningham P, Cord M, Delany SJ. Supervised Learning. In: *Machine Learning Techniques for Multimedia*. Springer Berlin Heidelberg; :21-49. doi:10.1007/978-3-540-75171-7_2
43. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: *Unsupervised and Semi-Supervised Learning*. Springer; 2019:3-21.
44. Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. Reinforcement Learning, Fast and Slow. *Trends Cogn Sci*. 2019;23(5):408-422. doi:10.1016/j.tics.2019.02.006
45. H2O.ai. H2O - Documentation. Published 2023. Accessed April 28, 2023. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science.html
46. H2O.ai. Gradient Boosting Machine (GBM) - H2O Documentation. Published 2023. Accessed February 12, 2021. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html
47. H2O.ai. Distributed Random Forest (DRF) - H2O Documentation. Published 2023. Accessed February 12, 2021. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html
48. H2O.ai. Generalized Linear Model (GLM) - H2O Documentation. Published 2023. Accessed February 12, 2021. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html
49. H2O.ai. Deep Learning Neural Networks - H2O Documentation. Published 2023. Accessed February 12, 2021. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html%0A

50. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1). doi:10.2202/1544-6115.1309
51. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005
52. Ko C, Brody JP. A genetic risk score for glioblastoma multiforme based on copy number variations. *Cancer Treat Res Commun*. 2021;27:100352. doi:10.1016/j.ctarc.2021.100352
53. Karatza P, Dalakleidi K, Athanasiou M, Nikita KS. Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf*. 2021;2021:2310-2313. doi:10.1109/EMBC46164.2021.9630556
54. Zhou C, Hu J, Wang Y, et al. A machine learning-based predictor for the identification of the recurrence of patients with gastric cancer after operation. *Sci Rep*. 2021;11(1):1571. doi:10.1038/s41598-021-81188-6
55. Lynch CM, Abdollahi B, Fuqua JD, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inform*. 2017;108:1-8. doi:10.1016/j.ijmedinf.2017.09.013
56. Chu CS, Lee NP, Adeoye J, Thomson P, Choi S-W. Machine learning and treatment outcome prediction for oral cancer. *J oral Pathol Med Off Publ Int Assoc Oral Pathol Am Acad Oral Pathol*. 2020;49(10):977-985. doi:10.1111/jop.13089
57. Kim SY, Kim Y-I, Kim HJ, et al. New approach of prediction of recurrence in thyroid cancer patients using machine learning. *Medicine (Baltimore)*. 2021;100(42):e27493. doi:10.1097/MD.0000000000027493
58. Alabi RO, Elmusrati M, Sawazaki-Calone I, et al. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *Int J Med Inform*. 2020;136:104068. doi:10.1016/j.ijmedinf.2019.104068
59. Drozdov I, Kidd M, Nadler B, et al. Predicting neuroendocrine tumor (carcinoid) neoplasia using gene expression profiling and supervised machine learning. *Cancer*. 2009;115(8):1638-1650. doi:10.1002/cncr.24180
60. LG A, AT E. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *J Heal Med Informatics*. 2013;04(02). doi:10.4172/2157-7420.1000124
61. Tseng CJ, Lu CJ, Chang CC, Chen G Den, Cheewakriangkrai C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artif Intell Med*. 2017;78:47-54. doi:10.1016/j.artmed.2017.06.003
62. Qi F, Zhu CJ, Yin L. Predicting breast cancer recurrence using data mining techniques. In: *ICBBT 2010 - 2010 International Conference on Bioinformatics and Biomedical Technology*. ; 2010:310-311. doi:10.1109/ICBBT.2010.5478952
63. Laurie NA, Donovan SL, Shih C-S, et al. Inactivation of the p53 pathway in retinoblastoma. doi:10.1038/nature05194
64. Kotsopoulos J. cancers BRCA Mutations and Breast Cancer Prevention. *Cancers (Basel)*. 2018;10:524. doi:10.3390/cancers10120524
65. Dimaras H, Corson TW, Cobrinik D. Retinoblastoma. *Nat Rev Dis Prim*. 2015;1.

doi:10.1038/nrdp.2015.21
66. Euhus DM, Leitch AM, Huth JF, Peters GN. *Limitations of the Gail Model in the Specialized Breast Cancer Risk Assessment Clinic*. http://cancertrials.nci.nih.gov/forms/CtRisk-
67. Hughes Kevin, Vogel Victor, Roche Constance, Bevers Therese. *Handbook of Breast Cancer Risk-Assessment*. Jones & Bartlett Learning,; 2003.
68. Mukdad L, Heineman TE, Alonso J, Badran KW, Kuan EC, St John MA. Oral tongue squamous cell carcinoma survival as stratified by age and sex: A surveillance, epidemiology, and end results analysis. *Laryngoscope*. 2019;129(9):2076-2081. doi:10.1002/lary.27720
69. Park HS, Lloyd S, Decker RH, Wilson LD, Yu JB. Overview of the Surveillance, Epidemiology, and End Results Database: Evolution, Data Variables, and Quality Assurance. *Curr Probl Cancer*. 2012;36(4):183-190. doi:10.1016/j.currproblcancer.2012.03.007
70. Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology. *Am J Surg Pathol*. 2016;40(12):e94-e102. doi:10.1097/PAS.0000000000000749
71. Islam A, Belhaouari SB, Rehman AU, Bensmail H. KNNOR: An oversampling technique for imbalanced datasets. *Appl Soft Comput*. 2022;115:108288. doi:10.1016/j.asoc.2021.108288
72. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
73. Garc\'\ia-Pedrajas N, Herrera F, Fyfe C, Sánchez JMB, Ali M. *Trends in Applied Intelligent Systems: 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, Cordoba, Spain, June 1-4, 2010, Proceedings, Part II*. Vol 6097. Springer; 2011.
74. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol*. 2009;45(4-5):309-316. doi:10.1016/j.oraloncology.2008.06.002
75. Patel SC, Carpenter WR, Tyree S, et al. Increasing incidence of oral tongue squamous cell carcinoma in young white women, age 18 to 44 years. *J Clin Oncol Off J Am Soc Clin Oncol*. 2011;29(11):1488-1494. doi:10.1200/JCO.2010.31.7883
76. Kim Y-J, Kim JH. Increasing incidence and improving survival of oral tongue squamous cell carcinoma. *Sci Rep*. 2020;10(1):7877. doi:10.1038/s41598-020-64748-0
77. Karadaghy OA, Shew M, New J, Bur AM. Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma. *JAMA Otolaryngol Head Neck Surg*. 2019;145(12):1115-1120. doi:10.1001/jamaoto.2019.0981
78. Berdugo J, Thompson LDR, Purgina B, et al. Measuring Depth of Invasion in Early Squamous Cell Carcinoma of the Oral Tongue: Positive Deep Margin, Extratumoral Perineural Invasion, and Other Challenges. *Head Neck Pathol*. 2019;13(2):154-161. doi:10.1007/s12105-018-0925-3
79. Ermer MA, Kirsch K, Bittermann G, Fretwurst T, Vach K, Metzger MC. Recurrence rate and shift in histopathological differentiation of oral squamous cell carcinoma – A long-term retrospective study over a period of 13.5 years. *J cranio-maxillo-facial Surg Off Publ*

Eur Assoc Cranio-Maxillo-Facial Surg. 2015;43(7):1309-1313. doi:10.1016/j.jcms.2015.05.011
80. Camisasca DR, Silami MANC, Honorato J, Dias FL, de Faria PAS, Lourenço S de QC. Oral squamous cell carcinoma: clinicopathological features in patients with and without recurrence. *ORL J Otorhinolaryngol Relat Spec*. 2011;73(3):170-176. doi:10.1159/000328340
81. An S, Jung E-J, Lee M, et al. Factors Related to Regional Recurrence in Early Stage Squamous Cell Carcinoma of the Oral Tongue. *Clin Exp Otorhinolaryngol*. 2008;1(3):166. doi:10.3342/ceo.2008.1.3.166
82. Sharma P, Shah S V, Taneja C, Patel AM, Patel MD. A prospective study of prognostic factors for recurrence in early oral tongue cancer. *J Clin Diagn Res*. 2013;7(11):2559-2562. doi:10.7860/JCDR/2013/6890.3611
83. Wang B, Zhang S, Yue K, Wang X-D. The recurrence and survival of oral squamous cell carcinoma: a report of 275 cases. *Chin J Cancer*. 2013;32(11):614-618. doi:10.5732/cjc.012.10219
84. Chernock RD. Morphologic features of conventional squamous cell carcinoma of the oropharynx: “keratinizing” and “nonkeratinizing” histologic types as the basis for a consistent classification system. *Head Neck Pathol*. 2012;6 Suppl 1(Suppl 1):S41-7. doi:10.1007/s12105-012-0373-4
85. Safi A-F, Kauke M, Grandoch A, Nickenig H-J, Zöller JE, Kreppel M. Analysis of clinicopathological risk factors for locoregional recurrence of oral squamous cell carcinoma - Retrospective analysis of 517 patients. *J cranio-maxillo-facial Surg Off Publ Eur Assoc Cranio-Maxillo-Facial Surg*. 2017;45(10):1749-1753. doi:10.1016/j.jcms.2017.07.012
86. Vázquez-Mahía I, Seoane J, Varela-Centelles P, Tomás I, Álvarez García A, López Cedrún JL. Predictors for tumor recurrence after primary definitive surgery for oral cancer. *J oral Maxillofac Surg Off J Am Assoc Oral Maxillofac Surg*. 2012;70(7):1724-1732. doi:10.1016/j.joms.2011.06.228
87. Lacko M, Braakhuis BJM, Sturgis EM, et al. Genetic Susceptibility to Head and Neck Squamous Cell Carcinoma. *Int J Radiat Oncol*. 2014;89(1):38-48. doi:https://doi.org/10.1016/j.ijrobp.2013.09.034
88. Copper MP, Jovanovic A, Nauta JJP, et al. Role of Genetic Factors in the Etiology of Squamous Cell Carcinoma of the Head and Neck. *Arch Otolaryngol Neck Surg*. 1995;121(2):157-160. doi:10.1001/archotol.1995.01890020019005
89. Matthias C, Harréus U, Strange R. Influential factors on tumor recurrence in head and neck cancer patients. *Eur Arch Oto-Rhino-Laryngology Head \& Neck*. 2006;263(1):37-42.
90. Heroiu Cataloiu A-D, Danciu CE, Popescu CR. Multiple cancers of the head and neck. *Maedica (Buchar)*. 2013;8(1):80-85.
91. Jerjes W, Upile T, Petrie A, et al. Clinicopathological parameters, recurrence, locoregional and distant metastasis in 115 T1-T2 oral squamous cell carcinoma patients. *Head Neck Oncol*. 2010;2:9. doi:10.1186/1758-3284-2-9
92. Wolfer S, Elstner S, Schultze-Mosgau S. Degree of Keratinization Is an Independent Prognostic Factor in Oral Squamous Cell Carcinoma. *J Oral Maxillofac Surg*.

2018;76(2):444-454. doi:https://doi.org/10.1016/j.joms.2017.06.034
93. Sinha N, Rigby MH, McNeil ML, et al. The histologic risk model is a useful and inexpensive tool to assess risk of recurrence and death in stage I or II squamous cell carcinoma of tongue and floor of mouth. *Mod Pathol*. 2018;31(5):772-779.
94. Brandwein-Gensler M, Teixeira MS, Lewis CM, et al. Oral squamous cell carcinoma: histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival. *Am J Surg Pathol*. 2005;29(2):167-178. doi:10.1097/01.pas.0000149687.90710.21
95. Chaturvedi A, Husain N, Misra S, et al. Validation of the Brandwein Gensler Risk Model in Patients of Oral Cavity Squamous Cell Carcinoma in North India. *Head Neck Pathol*. 2020;14(3):616-622. doi:10.1007/s12105-019-01082-6
96. El-Mofty SK. Histopathologic risk factors in oral and oropharyngeal squamous cell carcinoma variants: an update with special reference to HPV-related carcinomas. *Med Oral Patol Oral Cir Bucal*. 2014;19(4):e377-85. doi:10.4317/medoral.20184
97. O-charoenrat P, Pillai G, Patel S, et al. Tumour thickness predicts cervical nodal metastases and survival in early oral tongue cancer. *Oral Oncol*. 2003;39(4):386-390. doi:https://doi.org/10.1016/S1368-8375(02)00142-2
98. Turnbull C, Rahman N. Genetic Predisposition to Breast Cancer: Past, Present, and Future. *Annu Rev Genomics Hum Genet*. 2008;9(1):321-345. doi:10.1146/annurev.genom.9.081307.164339
99. Kirova YM, Stoppa-Lyonnet D, Savignoni A, Sigal-Zafrani B, Fabre N, Fourquet A. Risk of breast cancer recurrence and contralateral breast cancer in relation to BRCA1 and BRCA2 mutation status following breast-conserving surgery and radiotherapy. *Eur J Cancer*. 2005;41(15):2304-2311. doi:10.1016/j.ejca.2005.02.037
100. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi:10.1038/nrg.2016.49
101. Stankiewicz P, Lupski JR. Structural Variation in the Human Genome and its Role in Disease. *Annu Rev Med*. 2010;61(1):437-455. doi:10.1146/annurev-med-100708-204735
102. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet*. 2004;36(9):949-951. doi:10.1038/ng1416
103. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and Predicting Haploinsufficiency in the Human Genome. *PLoS Genet*. 2010;6(10):e1001154. doi:10.1371/journal.pgen.1001154
104. Sismani C, Koufaris C, Voskarides K. Copy Number Variation in Human Health, Disease and Evolution. In: *Genomic Elements in Health, Disease and Evolution*. ; 2015:129-154.
105. Krepischi ACV, Pearson PL, Rosenberg C. Germline copy number variations and cancer predisposition. *Futur Oncol*. 2012;8(4):441-450. doi:10.2217/fon.12.34
106. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Publ Gr*. Published online 2013. doi:10.1038/ng.2760
107. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun*. 2017;8(1):15183. doi:10.1038/ncomms15183
108. Igo RP, Kinzy TG, Cooke Bailey JN. Genetic Risk Scores. *Curr Protoc Hum Genet*.

2019;104(1). doi:10.1002/cphg.95
109. De La Vega FM, Bustamante CD. Polygenic risk scores: a biased prediction? *Genome Med*. 2018;10(1):100. doi:10.1186/s13073-018-0610-x
110. Madsen BE, Villesen P, Wiuf C. A periodic pattern of SNPs in the human genome. *Genome Res*. 2007;17(10):1414-1419. doi:10.1101/gr.6223207
111. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
112. Mardis ER. Next-Generation Sequencing Platforms. *Annu Rev Anal Chem*. 2013;6(1):287-303. doi:10.1146/annurev-anchem-062012-092628
113. de Araújo Lima L, Wang K. PennCNV in whole-genome sequencing data. *BMC Bioinformatics*. 2017;18(S11):383. doi:10.1186/s12859-017-1802-x
114. Rutkowska L, Pinkier I, Sałacińska K, et al. Identification of New Copy Number Variation and the Evaluation of a CNV Detection Tool for NGS Panel Data in Polish Familial Hypercholesterolemia Patients. *Genes (Basel)*. 2022;13(8):1424. doi:10.3390/genes13081424
115. Allen NE, Sudlow C, Peakman T, Collins R. UK biobank data: Come and get it. *Sci Transl Med*. 2014;6(224). doi:10.1126/scitranslmed.3008601
116. Van Hout C V, Tachmazidou I, Backman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature*. 2020;586:749. doi:10.1038/s41586-020-2853-0
117. Biobank Overview. Accessed November 28, 2022. https://www.ukbiobank.ac.uk/enable-your-research/costs
118. Christopher En-Li Toh. Chromosomal scale length variations as a genetic risk score for predicting complex human diseases in large scale genomic datasets . Published online 2021.
119. Center for Disease Control and Prevention. ICD-10 (Mortality). Published December 2021. Accessed August 16, 2023. https://www.cdc.gov/nchs/icd/icd10.htm
120. Brewster AM, Hortobagyi GN, Broglio KR, et al. Residual Risk of Breast Cancer Recurrence 5 Years After Adjuvant Therapy. *JNCI J Natl Cancer Inst*. 2008;100(16):1179-1183. doi:10.1093/jnci/djn233
121. ISAACS C. New prognostic factors for breast cancer recurrence. *Semin Oncol*. 2001;28(1):53-67. doi:10.1016/S0093-7754(01)90045-4
122. Fatapour Y, Brody JP. Genetic Risk Scores and Missing Heritability in Ovarian Cancer. *Genes (Basel)*. 2023;14(3):762. doi:10.3390/genes14030762
123. Gail MH, Brinton LA, Byar DP, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *JNCI J Natl Cancer Inst*. 1989;81(24):1879-1886. doi:10.1093/jnci/81.24.1879
124. Banegas MP, John EM, Slattery ML, et al. Projecting Individualized Absolute Invasive Breast Cancer Risk in US Hispanic Women. *J Natl Cancer Inst*. 2017;109(2):djw215. doi:10.1093/jnci/djw215
125. NIH National Cancer Institute. The Colorectal Cancer Risk Assessment Tool. Published April 2, 2021. Accessed June 22, 2022. https://ccrisktool.cancer.gov/
126. NIH National Cancer Institute. Lung Cancer Risk Models for Screening (R package: lcrisks).

Published 2023. Accessed July 10, 2023. https://dceg.cancer.gov/tools/risk-assessment/lcrisk
127. Lebrett MB, Crosbie EJ, Smith MJ, Woodward ER, Evans G, Crosbie PAJ. Targeting lung cancer screening to individuals at greatest risk: the role of genetic factors. *J Med Genet*. 2021;58:217-226. doi:10.1136/jmedgenet-2020-107399
128. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun*. 2020;11(1):5131. doi:10.1038/s41467-020-18918-3
129. AllofUs Research Hub. Data Sanpshots . Published February 2022. Accessed June 23, 2022. https://www.researchallofus.org/data-tools/data-snapshots/
130. illumina. Infinium Global Diversity Array-8 Kit. illumina. Published 2023. Accessed August 16, 2023. https://www.illumina.com/products/by-type/microarray-kits/infinium-global-diversity.html
131. All of Us Research Program. All of Us Genomic Quality Report. Published 2023. Accessed August 16, 2023. https://support.researchallofus.org/hc/en-us/articles/4617899955092-All-of-Us-Genomic-Quality-Report-
132. All of Us Research Program. Researcher Workbench . Published 2023. Accessed August 17, 2023. https://www.researchallofus.org/data-tools/workbench/
133. Toh C, Brody JP. A genetic risk score using human chromosomal-scale length variation can predict schizophrenia. *Sci Rep*. 2021;11(1):18866. doi:10.1038/s41598-021-97983-0
134. Ko C, Brody JP. Evaluation of a genetic risk score computed using human chromosomal-scale length variation to predict breast cancer. *Hum Genomics*. 2023;17(1):53. doi:10.1186/s40246-023-00482-8
135. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/caac.21660
136. Krontiras H, Farmer M, Whatley J. Breast Cancer Genetics and Indications for Prophylactic Mastectomy. *Surg Clin North Am*. 2018;98(4):677-685. doi:10.1016/j.suc.2018.03.004
137. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(6):438-451. doi:10.3322/caac.21583
138. Chlebowski RT, Anderson GL, Lane DS, et al. Predicting Risk of Breast Cancer in Postmenopausal Women by Hormone Receptor Status. *JNCI J Natl Cancer Inst*. 2007;99(22):1695-1705. doi:10.1093/jnci/djm224
139. Pal Choudhury P, Brook MN, Hurson AN, et al. Comparative validation of the BOADICEA and Tyrer-Cuzick breast cancer risk models incorporating classical risk factors and polygenic risk in a population-based prospective cohort of women of European ancestry. *Breast Cancer Res*. 2021;23(1):22. doi:10.1186/s13058-021-01399-7
140. McCarthy AM, Guan Z, Welch M, et al. Performance of Breast Cancer Risk-Assessment Models in a Large Mammography Cohort. *JNCI J Natl Cancer Inst*. 2020;112(5):489-497. doi:10.1093/jnci/djz177
141. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med*. 2004;23(7):1111-1130. doi:10.1002/sim.1668

142. Grubbs SS, Polite BN, Carney J, et al. Eliminating Racial Disparities in Colorectal Cancer in the Real World: It Took a Village. *J Clin Oncol*. 2013;31(16):1928-1930. doi:10.1200/JCO.2012.47.8412
143. Carethers JM. Racial and ethnic disparities in colorectal cancer incidence and mortality. In: ; 2021:197-229. doi:10.1016/bs.acr.2021.02.007
144. Perencevich M, Ojha RP, Steyerberg EW, Syngal S. Racial and Ethnic Variations in the Effects of Family History of Colorectal Cancer on Screening Compliance. *Gastroenterology*. 2013;145(4):775-781.e2. doi:10.1053/j.gastro.2013.06.037
145. Kupfer SS, McCaffrey S, Kim KE. Racial and Gender Disparities in Hereditary Colorectal Cancer Risk Assessment: The Role of Family History. *J Cancer Educ*. 2006;21(1, suppl):S32-S36. doi:10.1207/s15430154jce2101s_7
146. Murff HJ. Colonoscopy Screening in African Americans and Whites With Affected First-Degree Relatives. *Arch Intern Med*. 2008;168(6):625. doi:10.1001/archinte.168.6.625
147. Dunlop MG, Tenesa A, Farrington SM, et al. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42 103 individuals. *Gut*. 2013;62(6):871-881. doi:10.1136/gutjnl-2011-300537
148. Smith T, Gunter MJ, Tzoulaki I, Muller DC. The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. *Br J Cancer*. 2018;119(8):1036-1039. doi:10.1038/s41416-018-0282-8
149. Xin J, Chu H, Ben S, et al. Evaluating the effect of multiple genetic risk score models on colorectal cancer risk prediction. *Gene*. 2018;673:174-180. doi:10.1016/j.gene.2018.06.035
150. Brown LM, Check DP, Devesa SS. Oral Cavity and Pharynx Cancer Incidence Trends by Subsite in the United States: Changing Gender Patterns. *J Oncol*. 2012;2012:1-10. doi:10.1155/2012/649498
151. La Vecchia C, Lucchini F, Negri E, Levi F. Trends in oral cancer mortality in Europe. *Oral Oncol*. 2004;40(4):433-439. doi:10.1016/j.oraloncology.2003.09.013
152. Petersen PE. The World Oral Health Report 2003: continuous improvement of oral health in the 21st century - the approach of the WHO Global Oral Health Programme. *Community Dent Oral Epidemiol*. 2003;31:3-24. doi:10.1046/j..2003.com122.x
153. Gupta N, Gupta R, Acharya AK, et al. Changing Trends in oral cancer – a global scenario. *Nepal J Epidemiol*. 2017;6(4):613-619. doi:10.3126/nje.v6i4.17255
154. Moller H. Changing incidence of cancer of the tongue, oral cavity, and pharynx in Denmark. *J Oral Pathol Med*. 1989;18(4):224-229. doi:10.1111/j.1600-0714.1989.tb00767.x
155. Neville BW, Day TA. Oral Cancer and Precancerous Lesions. *CA Cancer J Clin*. 2002;52(4):195-215. doi:10.3322/canjclin.52.4.195
156. Brocklehurst P, Kujan O, O'Malley L, Ogden GR, Shepherd S, Glenny A-M. Screening programmes for the early detection and prevention of oral cancer. *Cochrane Database Syst Rev*. 2013;2021(3). doi:10.1002/14651858.CD004150.pub4
157. Sankaranarayanan R, Ramadas K, Thomas G, et al. Effect of screening on oral cancer mortality in Kerala, India: a cluster-randomised controlled trial. *Lancet*.

2005;365(9475):1927-1933. doi:10.1016/S0140-6736(05)66658-5
158. Speight PM, Elliott AE, Jullien JA, Downer MC, Zakzrewska JM. The use of artificial intelligence to identify people at risk of oral cancer and precancer. *Br Dent J*. 1995;179(10):382-387. doi:10.1038/sj.bdj.4808932
159. Shimpi N, Glurich I, Rostami R, Hegde H, Olson B, Acharya A. Development and Validation of a Non-Invasive, Chairside Oral Cavity Cancer Risk Assessment Prototype Using Machine Learning Approach. *J Pers Med*. 2022;12(4):614. doi:10.3390/jpm12040614
160. Tseng W-T, Chiang W-F, Liu S-Y, Roan J, Lin C-N. The Application of Data Mining Techniques to Oral Cancer Prognosis. *J Med Syst*. 2015;39(5):59. doi:10.1007/s10916-015-0241-3