INTEGRATION OF EPIDEMIOLOGICAL AND CANCER GENOME DATA USING ARTIFICIAL INTELLIGENCE APPROACH IN MIZO POPULATION

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

BRINDHA SENTHIL KUMAR
MZU REGISTRATION NO.: 1807416
Ph. D. REGISTRATION NO.: MZU/Ph. D./1617 of 06.11.2020

DEPARTMENT OF COMPUTER ENGINEERING
SCHOOL OF ENGINEERING AND TECHNOLOGY
FEBRUARY, 2025

INTEGRATION OF EPIDEMIOLOGICAL AND CANCER GENOME DATA USING ARTIFICIAL INTELLIGENCE APPROACH IN MIZO POPULATION

BY

BRINDHA SENTHIL KUMAR

Department of Computer Engineering

SUPERVISOR

Dr. LALHMINGLIANA

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Computer Engineering of Mizoram University, Aizawl.

MIZORAM UNIVERSITY

(A Central University)

Department of Computer Engineering

Tanhril, Aizawl - 796004

Gram: MZU, Phone: 9402322415, Email: lalhmingliana@mzu.edu.in

CERTIFICATE

This is to certify that the thesis entitled **"Integration of Epidemiological and Cancer Genome Data using Artificial Intelligence Approach in Mizo Population"** submitted by **Brindha Senthil Kumar**, Ph.D. Research Scholar for the award of the Degree of Doctor of Philosophy in Computer Engineering is carried out under my supervision and incorporates the student bona-fide research and this has not been submitted for the award of any degree in this or any other university or institute of learning.

Place: Aizawl, Mizoram (**DR. LALHMINGLIANA**)

Date: Supervisor

DECLARATION

MIZORAM UNIVERSITY

FEBRUARY 2025

I **Brindha Senthil Kumar**, hereby declare that the subject matter of this thesis is the record of work done by me, that the contents of this thesis did not form basis of the award of any previous degree to me or to the best of my knowledge to anybody else, and that the thesis has not been submitted by me for any research degree in any other University/Institute.

This is being submitted to the Mizoram University for the Degree of **Doctor of Philosophy in Computer Engineering.**

(**BRINDHA SENTHIL KUMAR**)

Research Scholar

(**DR. V.D.AMBETH KUMAR**)

Head of Department

(**DR. LALHMINGLIANA**)

Supervisor

ACKNOWLEDGEMENT

I consider myself incredibly fortunate to have Dr. Lalhmingliana, Department of Computer Engineering at Mizoram University, as my supervisor. I am profoundly grateful to him for his unwavering support, unconditional confidence, invaluable guidance, scholarly expertise, continuous encouragement, and heartfelt well-wishes throughout my research journey. Without his presence, this work would never have reached completion. Dr. Lalhmingliana has not only provided moral and spiritual support but has also been a pillar of strength in every aspect of my research endeavour.

In addition, I would like to extend my sincerest appreciation to Dr. V.D.Ambeth Kumar, the Head of the Department of Computer Engineering, Mizoram University. I am equally indebted to the esteemed faculty members of the department, including Prof. Ajoy Kumar Khan, Dr. Mayanglambam Sushilatha Devi, Mr. Vanlalhruaia and Mr. Nungleppam Gopil Singh of the Department of Computer Engineering, Mizoram University for their unwavering support and encouragement during my research work. I am also grateful to the non-teaching staff of the Department of Computer Engineering, Mizoram University.

I am thankful to all the Volunteers who took part and made it possible to conduct the study. A special thanks to Doctors and Clinicians from Mizoram State Cancer Institute and Civil Hospital Aizawl.

I am immensely grateful to my fellow lab mates, Puitea Ralte, Zaia and all the research scholars of the Department of Computer Engineering, Mizoram University, who have not only provided moral guidance but have also extended a helping hand whenever needed during my research work.

I am thankful to the Data Collectors and Lab Technicians, Jonathan Lalramhluna, Baby Lalrintluangi, R. Lalengkimi, and T. Lalhriatpuii for collecting the samples, clinical information, documentation, and DNA Isolation. Additionally, I would also like to give special recognition to David K. Zorinsanga, Lalnunmawii and Dintei for their immense help during my research work.

I am also thankful to the SERB, New Delhi (DSTNo:CRG/2022/004700) for the fellowship and to the Department of Biotechnology, New Delhi (DBT-Advanced State Biotech Hub, MZU) for the data support during the research.

I extend my sincere appreciation to my husband, Son, Parents, and maternal uncle whom were always encouraging me to upgrade myself in all aspects of my life and made me person of unique dimensional wisdom, and mother Nature who gave me smooth rides during my entire journey of this program. Also, their boundless care, love, wishes, and moral support have been pillar of strength in my research accomplishments and academic endeavour.

Place: Aizawl **(BRINDHA SENTHIL KUMAR)**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. INTRODUCTION AND REVIEW OF LITERATURE

Machine Learning (ML) models have demonstrated considerable potential in advancing cancer research and clinical practice across various types of cancer, including gastric, breast, and head and neck cancers. These models utilize extensive datasets of patient information, imaging studies, and molecular profiles to enhance diagnosis, prognosis, and treatment planning. In gastric cancer, ML algorithms have been employed to analyze endoscopic images, facilitating the detection of early-stage tumors and prediction of cancer progression (Khosravi et al. 2024). It also had provided a promising approach in medical diagnostics and treatment planning. These advanced computational techniques leverage large datasets of medical images, patient records, and genetic information to accurately categorize different types and stages of gastric cancer (Kumar et al. 2024). Machine learning algorithms, such as support vector machines, random forests, and gradient boosting, have demonstrated significant efficacy in identifying key features and patterns associated with various stages of gastric cancer (Wang et al. 2024). These algorithms can analyze complex combinations of clinical, histopathological, and molecular data to provide more precise and personalized classifications. Gastric cancer classification utilizing machine learning and deep learning (DL) algorithms has expanded beyond traditional methods, incorporating multi-modal data analysis and advanced neural network architectures.

Convolution Neural Networks (CNNs) have exhibited remarkable performance in analyzing medical imaging data, such as endoscopic images and computed tomography scans, to detect and classify gastric cancer lesions and its subtypes with high accuracy (Liu et al. 2023). These advanced neural networks possess the capability to automatically extract relevant features from various imaging modalities, including endoscopic images, computed tomography (CT) scans, and histopathology slides. This automated feature extraction process significantly enhances the accuracy and efficiency of gastric cancer diagnosis, enabling healthcare professionals to make more informed decisions. The ability of deep learning models to identify subtle patterns and biomarkers that may elude human observers is

particularly noteworthy, as it holds the potential for earlier detection and improved prognosis for gastric cancer patients. The integration of ML and DL approaches in gastric cancer classification extends beyond improved diagnostic accuracy (Zhang et al. 2024). Moreover, the application of explainable AI techniques has improved the interpretability of these complex models, allowing clinicians to understand the reasoning behind the classifications and fostering trust in AI-assisted decision-making processes (Binzagr 2024). As these technologies continue to evolve, they hold the potential to revolutionize gastric cancer diagnosis, prognosis, and treatment planning, ultimately leading to improved patient outcomes and more personalized healthcare strategies. This data-driven approach not only enhances the precision of gastric cancer management but also facilitates more efficient clinical trials and accelerated drug discovery processes. As deep learning technologies continue to evolve, their impact on gastric cancer classification and treatment is expected to increase, offering new possibilities for patients and advancing the frontiers of oncological research.

ML models have been developed to interpret mammograms and other imaging modalities, potentially enhancing the accuracy of screening programs and reducing false positives in breast cancer (Wang et al. 2024).ML algorithms have demonstrated promising results in accurately classifying breast cancer tumors as benign or malignant. These algorithms possess the capability to analyze complex patterns in medical imaging data, such as mammograms and ultrasounds, to detect subtle features that may be indicative of cancerous growths (Carriero et al. 2024). By utilizing large datasets of historical patient information and outcomes, machine learning models can continuously enhance their predictive capabilities, potentially facilitating earlier and more accurate diagnoses of breast cancer. Machine learning and Deep learning algorithms had have been exhibiting remarkable potential in breast cancer classification, leveraging their ability to process and interpret vast amounts of medical imaging data. These algorithms can identify intricate patterns and subtle anomalies in mammograms, ultrasounds, and other diagnostic images that may elude human observers. By analyzing pixel-level information, texture patterns, and morphological features, machine learning models can discern characteristics

associated with benign and malignant tumors with increasing accuracy. This capability is particularly valuable in detecting early-stage cancers or ambiguous cases that might be challenging for human radiologists to definitively classify.

The efficacy of machine learning in breast cancer classification lies in its capacity to learn from large-scale datasets and improve over time (Madani et al. 2022). As these algorithms are exposed to more patient data, including historical records, treatment outcomes, and long-term follow-ups, they can refine their predictive models and adapt to new patterns or variations in tumor presentations. This continuous learning process enables machine learning systems to potentially outperform traditional diagnostic methods in terms of sensitivity and specificity (Adam et al. 2023). Moreover, the integration of machine learning into clinical workflows can provide radiologists with valuable second opinions, potentially reducing false positives and negatives, and ultimately leading to more timely and appropriate interventions for patients with breast cancer. Additionally, recurrent neural networks (RNNs) and transformer-based models have been applied to analyze sequential data, such as longitudinal patient records and genomic sequences, to predict cancer progression and treatment outcomes (Lee et al. 2023).

Recent advancements in deep learning have significantly enhanced the accuracy and efficiency of machine learning algorithms for classifying head and neck cancer (Bang et al. 2023). Deep learning models possess the capability to extract complex features from CT scans, MRI images, and histopathological slides, thereby enabling more accurate staging and prognosis of head and neck cancers compared to traditional machine learning approaches (Alabi et al. 2024). Recent advancements in deep learning have substantially improved the field of head and neck cancer classification, offering unprecedented accuracy and efficiency in machine learning algorithms. By leveraging this capability, CNNs enable more accurate staging and prognosis of head and neck cancers compared to traditional machine learning approaches. The application of deep learning in head and neck cancer classification has significant implications for patient care and treatment planning. By improving the accuracy of tumor detection and classification, these advanced algorithms can

potentially lead to earlier diagnoses, more personalized treatment strategies, and improved patient outcomes (Broggi et al. 2024).

As research in this field continues to progress, it is anticipated that deep learning algorithms will play an increasingly crucial role in clinical decision-making processes, potentially transforming the landscape of head and neck cancer management. In head and neck cancer, ML techniques have been applied to analyze CT and MRI scans, aiding in tumor delineation for radiation therapy planning. Furthermore, across all three cancer types, ML models are being investigated to predict treatment outcomes, identify biomarkers, and personalize therapy regimens based on individual patient characteristics and genetic profiles. Machine learning models have been applied to various types of cancer, including gastric, breast, and head and neck cancers.

The field of artificial intelligence, particularly machine learning and deep learning, are advancing at an unprecedented rate, with novel algorithms, techniques, and models being developed and refined continuously. This rapid progress necessitates that researchers and practitioners remain informed of the latest developments by regularly consulting peer-reviewed literature and conducting comprehensive literature reviews. Furthermore, the application of machine learning models in cancer research encompasses a wide range of areas, including early detection, diagnosis, prognosis, treatment planning, and drug discovery. Each of these domains may require distinct types of machine learning approaches, and their efficacy can vary depending on the specific cancer type, dataset characteristics, and research objectives.

Understanding the causes, trends, and distribution of cancer in populations is crucial, and epidemiological risk factors are crucial for cancer prevention and management because they assist researchers, medical professionals, and public health officials in identifying high-risk individuals, putting in place efficient screening programs, and creating public health initiatives. These include inherited traits, exposures to the environment (like carcinogens), dietary and lifestyle choices (like diet and exercise), and infectious agents (Zodinpuii et al. 2022). The incidence of

cancer can be decreased by developing public health programs that target modifiable risk factors, such as obesity, smoking, and inactivity. Guidelines for high-risk populations; cancer screening and treatment strategies can be designed from epidemiological data. By identifying the epidemiological risk factors and causes, researchers can reduce cancer incidence, improve outcomes, and minimize health disparities.

Mutations can lead to cancer by disrupting normal cell functions and promoting uncontrollable cell growth. Certain genetic mutations in oncogeneshave the probability to cause cancer. Mutations can also inactivate tumor suppressor genes, which prevent uncontrolled cell division and help repair DNA damage. Some genetic mutations are inherited from parents and can significantly increase the risk of certain cancers, this can help with early detection and prevention strategies. Most cancer-related mutations also occur in somatic cells and are not passed down to offspring (Chakraborty et al. 2023). These somatic mutations accumulate over time due to environmental factors (like smoking, UV radiation, or exposure to chemicals etc.), aging, or random errors during cell division. Understanding genetic mutations allows for personalized medicine, where treatments are tailored to an individual's specific genetic profile.

Several machine learning models have been developed for cancer classification using medical, laboratory, diet-lifestyle and medical image datasets.ML and DL models have been gaining significance in biological classification and regression to elucidate the underlying risk factors in complex diseases. Signs and symptoms of the same cancer type have been utilized to develop biomarkers (Saba 2020). Machine learning algorithms have been extensively employed in cancer classification, achieving accuracy exceeding 97% (Dias and Torkamani, 2019). Deep learning techniques (Convolution Neural Networks) were developed to classify different cancer types with accuracy exceeding 95% on image data (Yadav et al. 2018). Several recent research publications based on machine learning applications in gastric, breast, and head and neck cancers are summarized in the following paragraphs. Thefeatures from medical check-up data were utilized (age, sex, body

mass index, *Helicobacter pylori*, upper gastroenterology endoscopy findings, and blood parameters) to classify gastric cancer and healthy individuals, producing an area under the curve (AUC) of 89.9% using the XGBoost algorithm (Taninagaet al. 2019). This algorithm utilized features from employees of the Nippon Telegraph and Telephone Corporation (NTT), Tokyo. The cases and controls were screened based on the upper gastrointestinal endoscopy procedure (invasive), with 1144 instances used for the testing dataset and 285 instances reserved for the testing dataset. The feature set of this study did not include dietary habits and environmental factors, despite the well-established association between gastric cancer and diet and lifestyle factors. Additionally, some instances without *H. pylori* infections or gastritis were classified into the gastric cancer class, representing false positives misclassified by the model. Naïve Bayes classifier had that utilized diet and lifestyle factors from the Mizo population. The study included 80 gastric cancer and 160 healthy control instances collected between 2016 and 2019(Brindha et al. 2020). Extra salt, tuibur, alcohol, and smoking were significant features selected by the heatmap feature selection method. These features were used for modeling, which achieved a highest accuracy of 90%, a low false positive rate of 3%, and, given the imbalanced dataset, a precision-recall curve with a micro-average of 80%. This model is considered optimal for identifying early gastric cancer risk factors from diet-lifestyle data. However, epidemiological data alone is insufficient to identify the core cause of gastric cancer at the gene level. Genetic-level changes are primarily attributable to diet and lifestyle habits in various cancer types.

CNN algorithm was developed for gastric cancer classification from endoscopic imagesthat incorporated eleven Inception modules to enhance computational efficiency and expedite the training phase. A total of 1702 early gastric cancer images and 386 non-cancerous lesion images were collected from 4 hospitals in Zhejiang province, Japan (Li et al. 2020). This deep learning model achieved an accuracy of 90.91%. The diagnostic sensitivity of the CNN model was significantly higher than that of human experts. A limitation of this work is that CNN models built on images do not reveal the etiological factors causing gastric cancer. These models can potentially replace radiologists in classifying cancerous and non-

cancerous images after invasive procedures but cannot contribute to minimizing gastric cancer incidence or mortality. Gaussian Naïve Bayes (GNB) and Support Vector Machine (SVM) classifiers had used three breast datasets from the UCI repository and data imputation was performed by the K-Nearest Neighbor algorithm, which achieved the highest accuracy of 96.09% with 5% marginalization of missing values using the GNB model (Anitha et al. 2020). SVM and extra-trees models were combined to develop a breast cancer classifier, which achieved an accuracy of 80.23%, significantly superior to other machine learning models. Feature selection based on extra trees improved accuracy by 7.29% compared to machine learning without the feature selection method (Alfian et al. 2022).Mammogram images, laboratory and demographic characteristics were utilized in a Random Forest classifier, which demonstrated a sensitivity of 95% and specificity of 80% (Rabiel et al. 2022). Parallel benchmarking of deep learning models, in conjunction with natural language processing (NLP) applied to imaging and non-imaging features for breast cancer detection and prognosis, may enable clinicians and researchers to acquire enhanced comprehension as they contemplate the clinical implementation or development of novel models (Hussain et al. 2024).

Electronic health records and mammogram images were integrated to develop a DL model for breast cancer classification, it had utilized 9,611 records for training, 1,055 records for validation, and 2,548 records for testing model performance. The confusion matrix of this model revealed 48% false negatives in mammogram findings (Akselrod-Ballin et al. 2019). Malignancy prediction achieved a Receiver Operating Characteristic curve Area Under the Curve (AUC) of 91%. Incorporating genetic information with clinical data could have enhanced the model's performance, potentially improving clinicians' efficiency in diagnosis and treatment. The algorithm's efficacy could be improved by employing machine learning algorithms for feature selection and implementing deep learning for classification.Jiande Wu and Chindo Hicks had proposed a machine learning approach to differentiate between triple-negative and non-triple-negative breast cancer using RNA-sequence data from The Cancer Genome Atlas (TCGA). SVM effectively classified triple-negative breast cancer samples from 5,502 differentially expressed genes, achieving

a training accuracy of 90% and testing accuracy of 82% when compared to decision tree, K-Nearest Neighbors (KNN), and Naïve Bayes classifiers (Wu and Hicks, 2021). This work focused solely on one breast cancer subtype and did not analyze the diet-lifestyle factors of the samples. Mapping the RNA-Sequence data and lifestyle features could have provided more in-depth knowledge in detecting significant factors causing differential expression in triple-negative and non-triple-negative breast cancer RNA-Sequences. Genome Deep Learning (GDL) model was deployed to classify 12 cancer types from whole exome sequences (WES). The WES cancer sequences were obtained from TCGA, while healthy individual sequences were sourced from the 1000 Genome Project (Sun et al. 2019). Twelve models were constructed to classify between cases and controls of the respective 12 cancer types (Bladder Urothelial, Breast invasive, Colon adenocarcinoma, Glioblastoma multiforme, Kidney renal clear cell, Brain Lower Grade Glioma, Lung squamous cell, Prostate, Ovarian serous cystadenocarcinoma, Thyroid, Skin Cutaneous Melanoma, Uterine Endometrial). The dataset comprised 6,083 cancer samples and 1,991 healthy individual WES. All 12 models achieved an accuracy exceeding 95% in classifying the 12 different cancer types. This method utilized core non-invasive data and could identify mutations with low frequency (disease-causing variants).

Similar approaches could be designed at the population level to identify key mutations and novel mutations in a particular community. Furthermore, integrating WES and epidemiological data could facilitate more precise cancer therapy and cancer risk identification. These approaches may not be sufficient to determine the features responsible for causing breast cancer from genetic and diet-lifestyle perspectives.

SVM model utilizing quantitative ultrasound scan features demonstrated 81% sensitivity and 76% specificity in distinguishing between responders and non-responding patients. This work used higher-order texture-of-texture features to improve predictive model performance (Safakish et al. 2023). Random forest, neural network, elastic net penalized logistic regression, and support vector machine algorithms were employed to predict the primary site of head and neck squamous cell

tumors based on their DNA methylation profiles. The classifiers demonstrated high classification accuracy, with the SVM achieving 89% accuracy (Leitheiser et al. 2022).

A deep learning framework was developed utilizing CT scan images of head and neck cancer. The images were segmented using the fuzzy c-means algorithm. Gray Level Co-Occurrence Matrix was employed to extract features, which yielded 98.8% accuracy (Gupta and Malhi, 2018).SVM classifier achieved an accuracy of 87% utilizing DNA methylation data of head and neck cancer. It predicted the most frequent localizations in the oral cavity, oropharynx, and larynx (Stark et al. 2024). Machine learning models were developed utilizing the Surveillance, Epidemiology, and End Results (SEER) registries database of head and neck squamous cell carcinoma. SVM demonstrated an accuracy of 70%, while the decision tree exhibited an Area Under the Receiver Operating Characteristic (AUROC) of 72%, in comparison to logistic regression and random forest models (Rohatgi et al. 2023).

In Mizoram, since 2003 the Incidence Rate of cancers in both male and female has been following an increasing trend. Stomach cancer is the leading cause of cancer death among male, followed by oesophagus, head & neck and lung. Among female, lung cancer is the leading cause of death followed by cervix, breast, oesophagus, head & neck.It is apparent that Mizoram is experiencing an alarming rise in cancer incidence and mortality rates, leading to the state being dubbed the cancer capital of India(NCDIR, 2014; Zomawia et al. 2023).

Earlier studiesin Mizo population had shown both well-known and novel mutations in cancer-related genes such as *TP53*, *MUC6,* and *ARID1A,* APC, AKT1 and CDH1 in Gastric Cancer (Chakraborty et al. 2023) and *TP53, CACNA1E, IGSF3,RYR1,CDH1*and *FAM155A* insomaticsamples and *ATM, STK11, CDKN2A* in germline Triple Negative Breast Cancer samples (Pachuau et al. 2025).

Consumption of smoked and salted meat, “sa-um” (fermented pork fat), fish and soda as a food additive have been shown to increase the risk of Stomach Cancer in Mizoram (Yadav et al. 2018) and high intake of animal fats, mainly from red meat

was correlated with an increased risk of breast cancer in Mizo population (Thapa et al. 2016).

There is a need of cancer research in integrated aspects including epidemiology, etiology, lifestyle factors and food habits, genetic mutations screening and ML methods for proper diagnosis and prognosis.

The study focuses on the high-risk Mizo population in Northeast India, addressing a critical knowledge gap in cancer research for this specific community. By analyzing point mutations in relation to etiological factors, the research aims to develop more accurate and efficient methods for early cancer detection. This approach not only promises to improve cancer diagnosis but also offers a non-invasive, cost-effective, and rapid alternative to traditional biopsy procedures. The integration of community-specific data will contribute to a better understanding of the cancer burden in the Mizo population and facilitate the extraction of clinical knowledge tailored to their unique genetic and environmental context. Ultimately, this research has the potential to significantly impact cancer prevention and treatment strategies, particularly in resource-limited settings. These models aim to enhance diagnostic accuracy, risk stratification, and treatment planning for cancers by utilizing the complementary information provided by epidemiological and genomic data.

The present study is a seminal work that integrates epidemiological and genomic data to elucidate the etiological factors for human cancer types. Epidemiological and genomic data of various cancer types are available, enabling the integration of these two feature sets. This integration will facilitate the determination of the underlying causes of cancers. It remains a significant challenge for researchers to identify the salient features of a deep learning model that contribute to high accuracy or sensitivity. This research aims to address the knowledge gap between epidemiological and genomic data by enabling early detection of cancer through the identification of significant point mutations in relation to the etiological factors that cause cancer in the high-risk Mizo population of Northeast India. This study will also contribute to the aggregation of data from multiple sources specific to the Mizo

population, which will greatly aid in quantifying the burden of this disease and extracting clinical knowledge specific to the community. Although biopsy is considered the gold standard for cancer diagnosis, it is an invasive, costly, and time-consuming procedure, particularly in remote areas. The present methods will provide non-invasive, time-efficient, and rapid approaches for accurately predicting the disease and its causes at both genetic and diet-lifestyle levels.

The present study also represents a significant advancement in cancer research by integrating epidemiological and genomic data to identify etiological factors for various human cancer types. This integration of diverse datasets allows for a more comprehensive understanding of cancer development and progression. By combining population-level epidemiological data with genetic information, researchers can uncover complex relationships between environmental factors, lifestyle choices, and genetic predispositions that contribute to cancer risk. This approach has the potential to revolutionize cancer prevention and early detection strategies by providing a more nuanced understanding of cancer etiology. This information can be used to implement targeted screening programs, develop personalized prevention strategies, and guide treatment decisions for patients with hereditary cancer syndromes.

Chapter 2. OBJECTIVES

1) Identification of etiological risk factors from cancer epidemiological data by constructing effective **Machine Learning models**.

2) Detection of disease causing **mutations from genomic variants** between cancer types and healthy controls using Machine Learning models.

3) To find the association between **epidemiological factors and genomic traits** by integrating epidemiological and genomic data.

Chapter 3. METHODS

3.1 Epidemiological Data of Gastric Cancer patients and Study Population

The epidemiological data for gastric, breast and head & neck were obtained from Department of Biotechnology, Mizoram University which was collected from Civil Hospital Aizawl and Mizoram State Cancer Institute, Mizoram, Northeast India. Based on the literature, dietary and lifestyle characteristics associated with the etiology of the above-mentioned three cancers were selected for this study. The data utilized in this investigation were collected after obtaining informed consent from cancer patients and healthy individuals.

3.2 Genomic Variant Data

The germline variant dataset was obtained from Mizoram cancer patients in Variant Call Format (VCF) for gastric, breast, and head and neck cancers from Department of Biotechnology, Mizoram University. This study had utilized germline variants of afore-mentioned cancer types, as germline variants play a crucial role in understanding cancer predisposition and inheritance patterns(Cobaleda et al. 2024). The DNA isolated from the patient's blood was subjected to Whole Exome Sequencing (100× coverage) in Illumina NovaSeq 6000 and the quality was checked using FastQC v0.12.1, adapter sequences removed using Trimmomatic v0.39 (Andrews, 2010; Bolger et al. 2014), aligned to GRCh38.p13 reference genome using the BWA-MEM algorithm followed by sorting the aligned reads, marking PCR duplicates, base quality score recalibration (dbSNP138.vcf), variant calling (GATK v4.2.0.0 HaplotypeCaller) and variant quality score recalibration (VQSR) (Li, 2013; Van der Auwera and O'Connor, 2020). The variants were annotated in ANNOVAR using RefGen, ClinVar, 1000 genome, avsnp15, esp3500, gnomad211_exome databases.

The annotated VCF files contained minor allele frequencies of all geographical populations, functional predictions for each single nucleotide polymorphism (SNP): Combined Annotation Dependent Depletion (CADD)

(Rentzsch et al. 2019), Polymorphism Phenotyping v2 (PolyPhen-2) (Adzhubei et al. 2013), Functional Analysis through Hidden Markova Models (FATHMM) (Shihab et al. 2012), Variant Effect Scoring Tool (VEST) (Douville, et al. 2016), Sorting Intolerant From Tolerant (SIFT) (Sim et al. 2012), Likelihood Ratio Test (LRT) (Zeng et. al. 2014), MutationAssessor (Reva et al. 2011), MutationTaster, RadialSVM, and several hand-crafted features (Heterogeneous SNV, homogenous SNV, Amino acid alteration_SNV, exon_splicing SNV, startloss_SNV, stoploss_SN), chromosome number, gene names, reference base, altered base, and positions.

3.3 Python Packages

Python Jupyter notebook nersion 3 platform was utilized to construct data models employing learning packages from scikit-learn (0.20.4) (Pedregosa et al. 2011) and Keras (Chollet et al. 2015). Numpy (1.16.6), pandas (0.24.2), seaborn (0.9.1), scipy (1.2.3), and matplotlib (2.2.5) were imported for interpretation and classification of the datapoints (Hunter, 2007).

3.4 Hardware Configuration

ML and DL models have been developed using GPU Server - HPE PROLIANT DL560 GEN10. CPU – Intel Xeon Gold 5218 x 4 nos, Number of Cores per Processor – 16cores, total 64 cores, Memory – 64GB x 16 nos = 1024GB, GPU - NVIDIA Tesla T4 16GB

3.5 Classification of Gastric Cancer using Epidemiological Features

The dataset comprises 117 gastric cancer patients and 148 healthy controls from the Mizo population. Twenty-nine features were included, encompassing information about dietary habits, lifestyle factors, environmental exposures, and health history (Table 1).

Table 1. Features and Data Labeling

S.No	Features	Data Labelling
1	Gender	1- male, 2-female
2	Age	continuous
3	Marital Status	0 – unmarried, 1- married
4	Weight	Continuous
5	Family history CA	0-no, 1-yes
6	Relatives with CA	0-none, 1-uncle and aunts, 2-father, mother, brother and sister,3-both
7	Cell phone tower	0-no, 1-yes
8	Physical activity/Exercise	0-No, 1- two to three times a week, 2-everyday
9	Spicy foods	0-no, 1-little, 2-average, 3-more frequent
10	Saum(fermented pork fat)	0-no, 1-little, 2-average, 3-more frequent
11	Smoked Vegetables	0-no, 1-little, 2-average, 3-more frequent
12	Smoked Meat	0-no, 1-little, 2-average, 3-more frequent
13	Boiled food	0-no, 1-little, 2-average, 3-more frequent
14	Fried Food	0-no, 1-little, 2-average, 3-more frequent
15	Fruits/Fibres	0-no, 1-little, 2-average, 3-more frequent
16	Reuse of used cooking oil	0-no, 1-yes
17	Cosmetic Use	0-no, 1-occasionaly, 2- regularly
18	Alcohol	0-no, 1-little, 2-average, 3-more frequent
19	Tuibur	0-no, 1-little, 2-average, 3-more frequent
20	Smoking	0-no, 1-yes
21	Started smoking before age 18	0-no, 1- yes
22	Smoking Zozial (unfiltered local cigar)	0-no, 1- Zozial plus other cigars, 2- only Zozial
23	No of cigarettes smoked per day	0-No, 1to– 1 to 10, greater than 10 per day - 2
24	Sadha	0-no, 1-yes
25	Khuva	0-no, 1-yes
26	Are u taking khuva before age 18	0-no, 1-yes
27	Zardha pan	0-no, 1-yes
27	Chewing_Supari	0-no, 1-yes
29	Passive Smoking Exposure	0-no, 1-rarely, 2-continously
30	Class	0-controls, 1-cases

3.5.1 Data Preprocessing

The dataset exhibited missing values, which were subsequently imputed utilizing their respective class median values, as the features were discrete (Emmanuel et al. 2021).This approach helps maintain the overall distribution of the data while addressing the issue of missing values. Using the median as a replacement, the impact of outliers can be minimized, ensuring that the imputed values are representative of the central tendency of the dataset.

3.5.2 Data Distribution

The age distribution in the cases ranged from 40 to 80 years old, while in the controls, the age distribution was 21 to 60 years (Fig 1). There was substantial bias in smoking, alcohol, and other smokeless tobacco product intake between cases and controls, as the control samples included a higher number of female participants, with the majority falling between 23 to 35 years old.

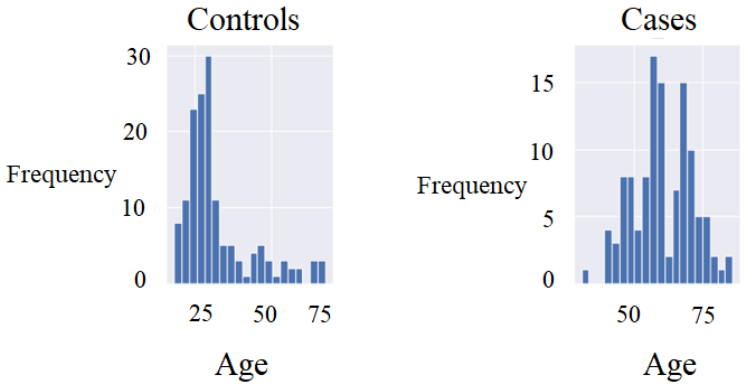


Fig 1. Age distribution byCases and Controls

3.5.3 Feature Scaling

All features in the dataset were normalized using MinMax scaling. The transformed value is calculated by subtracting each feature value from its minimum and dividing by the difference between the maximum and minimum of the feature. Feature scaling can significantly enhance the convergence rate of gradient descent, thereby expediting error minimization during the training of the logistic regression model (Sergey et al. 2015).

$$X_{i\_scale} = \frac{x_i - x_{i\_min}}{x_{i\_max} - x_{i\_min}}$$

where, x_i is the feature value

x_{i_min} is the minimum of the feature x_i

x_{i_max} is the maximum of the feature x_i

x_{i_scale} is the new value of x_i obtained after MinMax feature scaling method

3.5.4 Logistic Regression Model

A logistic regression model was developed to classify healthy individuals and gastric cancer patients based on epidemiological features. The model predicted the outcome of the target variable based on a threshold value. If the outcome exceeded the threshold, the observation was classified as cancer; and if the outcome was below the threshold, the observation was classified as healthy. In this study, the threshold was set at 0.5 for the logistic regression to differentiate between gastric cancer patients and healthy individuals. The logistic regression curve f(t) can be represented as:

$$f(t) = \frac{e^t}{1+e^t} \text{..... eq 1.}$$

Let us assume $t=\theta + \theta_1 x$, substitute in eq. 1

$$f(t) = \frac{e^{\theta+\theta_1 x}}{1+e^{\theta+\theta_1 x}} \text{.....eq 2.}$$

The eq. 2 can be written in logical regression:

$$f(x) = logit(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots)$$

$$\text{i.e.} f(t) = \ln\left(\frac{e^t}{1+e^t}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots \text{ eq. 3}$$

Consider the two probability events: if a probability of event E was *p*, then the probability of not E is *1-p*.

$$\frac{f(t)}{1-f(t)} \text{.....eq. 4}$$

Substitute eq. 4 with f(t) value from eq. 1:

$$f(t) = \frac{\frac{e^t}{1+e^t}}{1-\frac{e^t}{1+e^t}} \text{..... eq.5}$$

Simplify the eq.5 $f(t) = \frac{e^t}{(e^t+1)(\frac{1}{(e^t+1)})}$…..…..eq6

Therefore, $f(t) = logit(e^t)$ from eq 6.

Logistic Regression (LogReg):

$$LogReg(t) = \ln\left(\frac{f(t)}{1-f(t)}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots\ldots\ldots\text{eq7}$$

From equation 7, $\theta_0 + \theta_1 + \theta_2$ were hyperparameter that had to be established during the training phase. These values were crucial for accurate prediction in the testing phase. Fig 2 presented the complete pseudocode of the logistic regression classifier developed for this study.

Input:

A set of D records with n attributes, $x_1, x_2, \ldots\ldots x_n$ are features and y is the target variable.

Step1: Data pre-processing

- *Dataset has many missing values, they were imputed using median*
- *Data scaling was done using minmax scalar python library.*

Step2:

- *Allocate features vectors as X and target variable as Y.*
- *Using train_test_split machine learning library split the dataset into 75% for training and 25% for testing*
- *Transpose all the features from the training dataset and reshape the target variable in the training dataset*
- *Transpose all the features from the test dataset and reshape the target variable in the test dataset*

Step3:

- *Define a sigmoid function and pass feature array as parameter.*
- *Define a LogisticRegression Model and pass feature array, target array, learning rate and number of iteration as its parameters.*
  - *Initialize the intercept and slopes of all the features to 0.*
  - *For i in range number of iterations*
    - *Compute $z = W^T X_{tr} + B$*
    - *Compute probabilistic prediction $pp = \sigma(z)$*
    - *Compute cost function $cost = -\frac{1}{m} \sum_{i=1}^{m} Y_{tr} * log(pp) + (1 - Y_{tr}) * log(1 - pp)$*
    - *Apply the Gradient Descent to find the optimal global minima*
      - *Compute the B and W with respect to cost:*

$$dW = \frac{\partial cost}{\partial B} (pp - Y_{tr}) * X_{tr}^T$$

$$dB = \frac{\partial cost}{\partial W} (pp - Y_{tr})$$

      - *From the calculated dW and dB, compute new B and W values with respect to learning rate α*

$$W = W - \alpha * dW^T$$

$$B = B - \alpha * dB$$

Step4: Train the model with step 3set parameters and store the cost

Step5: Define accuracy function by passing features, target values of test data, learning rate, and number of iterations

- *Compute $z = W^T X_{te} + B$*
- *Compute probabilistic prediction $pp = \sigma(z)$*
- *If $pp > 0.5$*
  Store pp values in array pp()
- *Accuracy $= (1 - (\sum_{i=m}^{m} | pp - Y_{te}|)/Y_{te}) * 100$*

Step6: Plot a graph to visualize cost with respect to number of iterations.

Fig 2.Proposed Logistic Regression Algorithm

3.6 Classification of Gastric Cancer using Genomic Features

3.6.1 Gastric Cancer Big Dataset

A total of 192,781 variants were identified. During the quality control process, variants with read depth below the mean value were excluded, and a mean coverage of 93% was maintained. Following quality control measures, 72,471 benign variants and 8,788 pathogenic variants remained (Fig 3). Pathogenic variants were assigned label of 1, while benign variants were assigned label of 0, serving as the class label (dependent variable). The class labels were assigned only if the variant was predicted to be benign or pathogenic by all six prediction tools: MutationTaster, FATHMM_pred, LR_pred, RadialSVM, SIFT_pred, and Polyphen2_HDIV. To minimize noise, avoid biasness in model's performanceand improve the model's generalization to distinguish between cancerous (pathogenic) and non-cancerous (benign) SNPs, columns containing missing values were removed, resulting in 36 features being subjected to data visualization and modeling, as presented in Table 2. Fig 4 depicts the flow chart of the steps from raw input acquisition to the attainment of the desired results. Thirty seven gastric cancer patients and twenty seven healthy individuals' genomic data had been merged for this gastric cancer big data modeling, interpretation and visualization.

3.6.2 Data Preprocessing

The features exhibited varying ranges; consequently, the training dataset was scaled using standard scalar to mitigate misinterpretation by the machine learning models and facilitate the identification of an optimal solution. Missing values in attributes and instances were eliminated to ensure data consistency, prevent bias in data modeling, and significantly reduce errors. The outliers were not imputed with mean or median values, as such imputation could potentially impact the accuracy of clinical findings, drug target effectiveness, patient prognosis, and treatment outcomes.

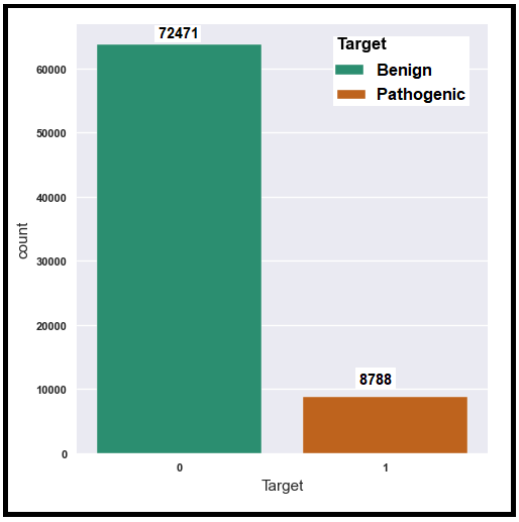


Fig 3. Number of Benign and Pathogenic variants in Gastric Cancer

Table 2. Gastric Cancer Genetic Attributes

Genetic Attributes			
1	Chr	19	AF_sas
2	Start	20	AF_amr
3	End	21	AF_eas
4	Ref	22	AF_nfe
5	Alt	23	AF_fin
6	Genes	24	AF_asj
7	AAChanges	25	AF_oth
8	VEST3_score	26	non_topmed_AF_popmax
9	CADD_raw	27	non_neuro_AF_popmax
10	CADD_phred	28	non_cancer_AF_popmax
11	GERP_RS	29	ExAC_ALL
12	phyloP46way_placental	30	ExAC_AFR
13	AF	31	ExAC_AMR
14	AF_popmax	32	ExAC_EAS
15	AF_male	33	ExAC_NFE
16	AF_female	34	ExAC_OTH
17	AF_raw	35	ExAC_SAS
18	AF_afr	36	Target

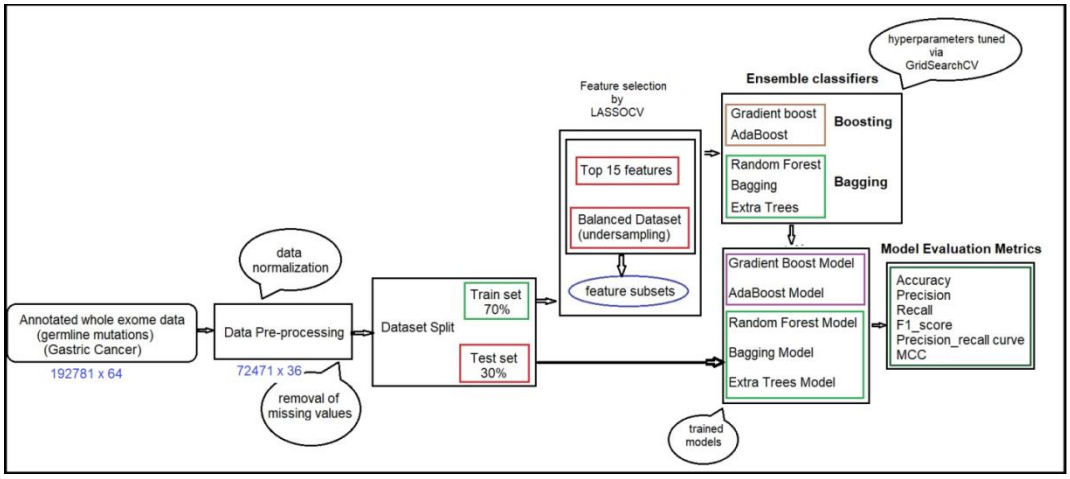


Fig 4.Flow chart of ensemble classifiers on gastric cancer germline mutations.

3.6.3 LASSO Feature Selection

Additionally, feature selection and engineering can help identify the most relevant variables for classification and capture complex relationships within the data, potentially uncovering hidden patterns and improving overall model accuracy. LASSO (Least Absolute Shrinkage and Selection Operator) regression is a powerful technique for feature selection in high-dimensional datasets (Shaon et al. 2024). It adds penalty term to objective function(ordinary least squares), which effectively make some coefficients to zero, thereby automatically performing feature selection. Consequently, the present study utilized LASSO regression with 5-fold cross-validation to identify the key features contributing to variant pathogenicity.

Fig 5 illustrates the significant 16 features that contribute to the target value, while the coefficients of other irrelevant features were reduced to zero. Given that the present dataset was large and highly imbalanced, in order to address noise, decision uncertainty, and reduce computational complexity, this study adopted L1-penalty-based feature selection (LASSO) to extract optimal features (Yu et al. 2021). Table 2 presents the hyperparameter set for feature selection using the extra trees classifier. Additionally, Table 3 displays three significant hyperparameters tuned during 5- fold LASSOCV genetic traits selection. These hyperparameter values had yielded the best average performance across all 5-folds was selected as the optimal

regularization.The final model had been trained with the optimal hyperparameters (Table 3), which identified the most important features by assigning non-zero coefficients to them while shrinking less important features' coefficients to zero (Fig 5).

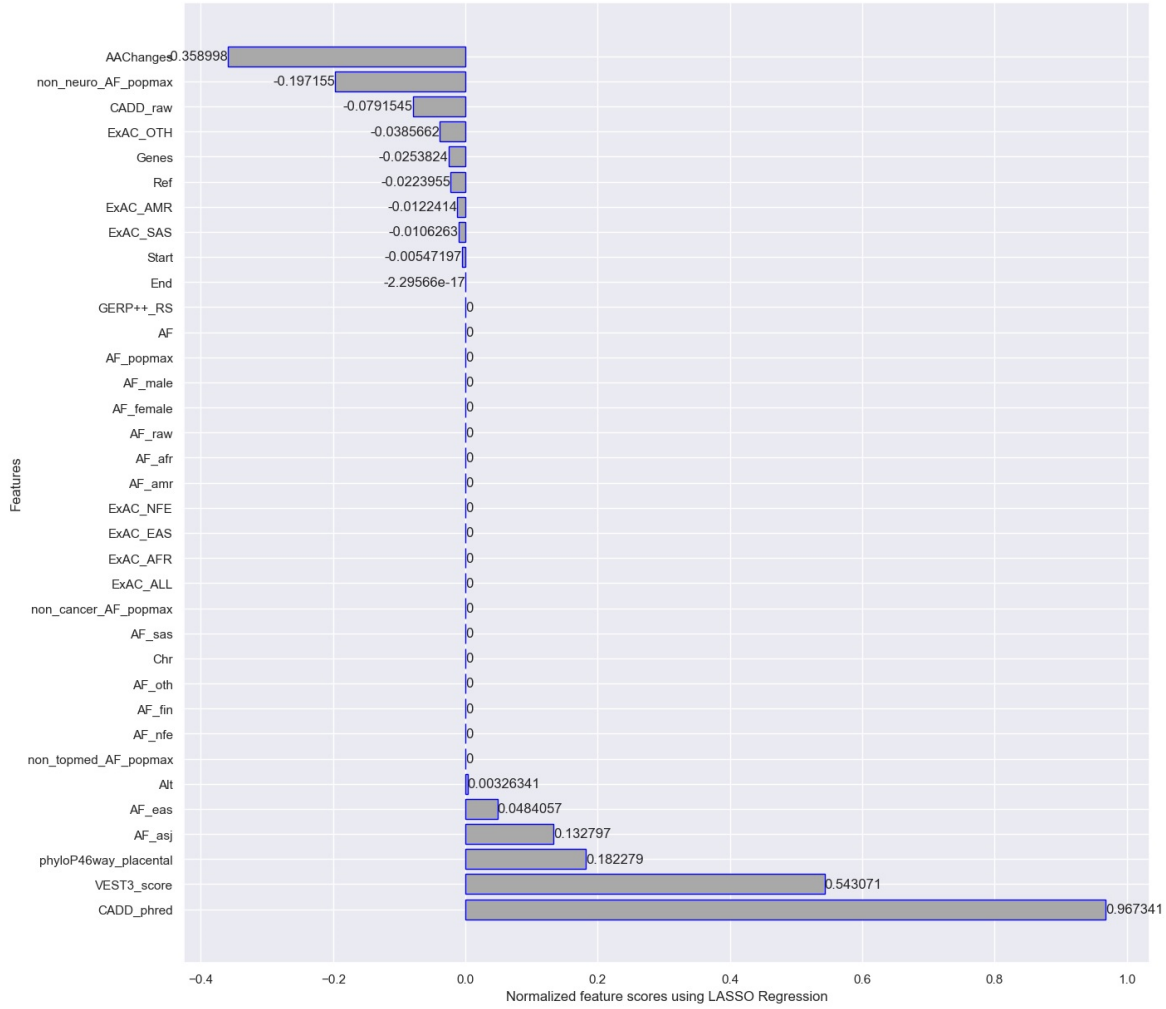


Fig 5. Feature Ranking by 5-Fold LASSOCV

Table 3. Hyperparameter for Feature Selection by 5-fold LASSOCV

Hyperparameter for LASSOCV	Optimal values
cv	5
max_iteration	100000
tolerance	0.000001

3.6.4 Feature Subset

3.6.4.1 Imbalanced Dataset

On utilizing the top 15 genetic traits selected by 5-fold LASSOCV, an imbalanced dataset was constructed with the size and features as illustrated in Fig 4 and Table 4, respectively. In practice, genetic datasets exhibit a significantly higher prevalence of negative observations compared to positive observations (Liu et al. 2022). Beyond the challenge of their substantial dimensionality, the utilization of these highly imbalanced datasets to extract meaningful and significant signatures represents the true potential of Artificial Intelligence.

3.6.4.2 Random undersampling

The dataset presented exhibited significant imbalance, which may result in inadequate learning of the positive minority class. To mitigate poor performance on the minority class and prevent overfitting, a balanced dataset was randomly subsampled (undersampling) and utilized to assess the viability of the features selected by the 5-fold LASSOCV method (Hasanin et al. 2019, Liu et al. 2022). Table 4 clearly demonstrated the balanced feature subsets. Interestingly, feature subsets selected by balanced and imbalanced dataset were same.

Table 4. Feature Subsets

Subsets	**Features**
Top 15 From Imbalanced Dataset	CADD_phred,VEST3_score, phyloP46way_placental, AF_asj, AF_eas, Alt, Start, ExAC_SAS, ExAC_AMR, Ref, Genes, ExAC_OTH, CADD_raw, non_neuro_AF_popmax, AAChanges
Top 15 From undersampled (Balanced Dataset)	CADD_phred,VEST3_score, phyloP46way_placental, AF_asj, AF_eas, Alt, Start, ExAC_SAS, ExAC_AMR, Ref, Genes, ExAC_OTH, CADD_raw, non_neuro_AF_popmax, AAChanges

3.6.5 Dataset split

The dataset was split into 70% training and 30% for testing. The traindataset was utilized to identify optimal hyperparameters using a five-fold GridSearchCV

procedure. Subsequently, the trained model was applied on the test set to evaluate the performance of the models in terms of accuracy, precision, recall, f1_score, Area under the Receiver Operating Characteristic Curve (AUC-ROC), Precision-Recall Area Under the Curve (PR-AUC) of the minority class (pathogenic) and Matthews correlation coefficient (MCC). In this context, MCC and PR-AUC were accorded greater importance in evaluating the model performance on the imbalanced dataset, particularly to assess the performance of the positive minority class. Furthermore, false positives can potentially misdirect drug molecule identification and increase the probability of overlooking true pathogenic variants (Taksler, 2018). As a general principle, PR-AUC is considered one of the primary model evaluation metrics for highly imbalanced big data (Hancock et al. 2023).

3.6.6 Ensemble Models

One approach to minimize false positives in machine learning for cancer classification is to use ensemble methods, which combines multiple models to improve overall accuracy and reduce errors (Liu et al. 2022). Another effective technique is to implement rigorous cross-validation procedures, ensuring the model's performance is consistent across different subsets of data (Arlot and Celisse, 2010).In accordance to the above statements, five ensemble classifiers were selected: Gradient boost, AdaBoost, Random forest, Bagging, and extra trees algorithms.To predict pathogenic and benign variants from gastric cancer big data, five ensemble models were selected: Gradient boost, AdaBoost, Random forest, Bagging, and extra trees classifiers. The models were optimized with hyperparameters obtained through 5-fold cross-validation grid search method on the training datasets using 2 feature subsets (Table 4). All trained models were preserved for future access and practical implementation. Model performances were evaluated on 2 feature subsets (Table 4) based on accuracy, precision, recall, f1_score, MCC, ROC-AUC, and PR-AUC. The best-performing machine learning classifier was selected based on accuracy, precision, recall, and PR-AUC on the imbalanced dataset (Hancock et al. 2023). The optimal classifier was chosen based on accuracy, MCC, and ROC-AUC on the

balanced dataset. The test set was exclusively utilized to assess the performance of hyperparameter-tuned-trained ensemble models.

3.7 Classification of MSI Gastric Cancer using Genomic and Epidemiological Features

3.7.1 Data Source

The study included 79 gastric cancer patients who had been diagnosed with the disease through biopsy and endoscopy. The dataset had eleven major risk factors associated with gastric cancer. This dataset had no data missing or noise, as the questions were answered through in-person interactions with patients to develop appropriate decision-making models. Targeted resequencing of a 60-gene panel specific to gastric cancer was conducted using next-generation sequencing (NGS), and 142 variants (both synonymous and non-synonymous) were identified by the base-by-base variant calling tool (BBB). For this study, somatic mutations were integrated with diet and lifestyle data. The dataset contained 70MSI positive and 72 MSI negative instances.

3.7.2 Data Labeling

The dataset comprised eleven attributes, encompassing both diet and lifestyle data. Data labeling was performed using attributes including: smoked food, alcohol, Khaini/sadha, khuva, tuibur, extra salt, saum, smoking, age, sex, and target (MSI - 1 or non MSI - 0) (Table 5).

Table 5. Diet and Life Style features for MSI Gastric Cancer

Features	**Data Labelling**
Age (years)	(30-85)
Sex	(1 - Male, 2- Female)
Extra salt	(0 - None, 1 - Little to Average, 2 - High/Excess)

Sa-um	(0 - None, 1 - Little to Average, 2 - High/Excess)
Smoked Food	(0 - None, 1 - Little to Average, 2 - High/Excess)
Khaini / Sadha	(0 - None, 1 - Little to Average, 2 - High/Excess)
Tuibur	(0 - None, 1 - Little to Average, 2 - High/Excess)
Alcohol	(0 - Non drinker, 1 - once or twice a month, 2 - more than once a week, 3 - Daily drinker
Smoking	(0 - No, 1 - Yes)
Paan	(0 - No, 1 - Yes)
MSI	(0 – Negative, 1- Positive)

Sa-um- Fermented pork fat; Smoked Food- smoked vegetables and meat;Khaini / Sadha- smokeless tobacco products;Tuibur- smoke filled tobacco water; Paan- betel leaf with lime and raw arecanut; MSI- Microsatellite Instability.

3.7.3 Learning Curves

Learning curves were generated to address underfitting, and feature selection was performed to mitigate overfitting (Viering and Loog, 2021). Underfitting and overfitting were significant problems in data mining that could potentially degrade prediction performance (van Rijn et al. 2015). The primary objective was to select a model capable of performing well on future data and to extract key features associated with MSI-GC. Learning curves were generated for Logistic Regression, Naive Bayes, Multilayer Perceptron, Support Vector Machine, and Random Forest (Fig 6).

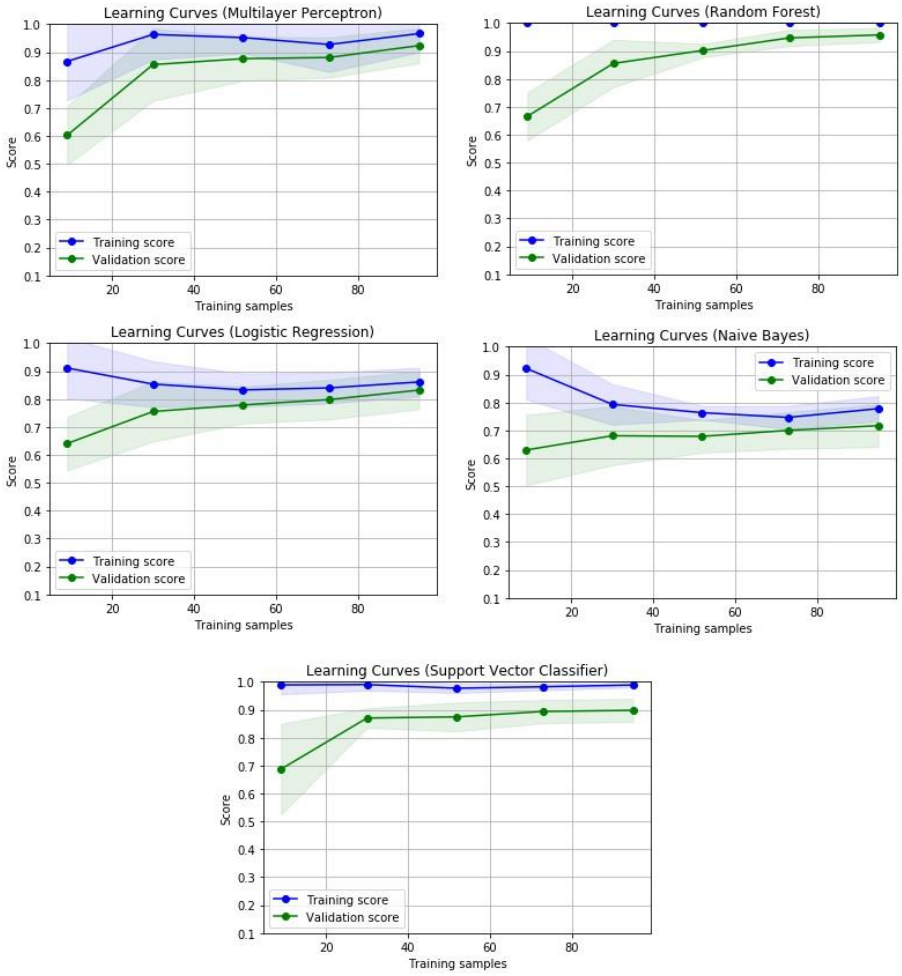


Fig 6.Learning Curves of five supervised learning algorithms for MSI-GC Classification

3.7.4 Feature Selection

Feature selection and redundant data elimination were conducted using three established methods: ridge regression (Friedman et al. 2010), extra trees classifier (Hemphill et al. 2014), and recursive feature elimination methods (Guyon et al. 2002). In thisstudy, lifestyle and food habits were evaluated for their risk for MSI-GC. Consequently, three robust feature selection algorithms were chosen due to their respective advantages: a) Ridge regression method - the ridge penalty prevents nullification of positively and negatively correlated features, thereby plays a role in identifying risk features. Furthermore, the present dataset contains multicollinearity

features, which can be effectively addressed by the ridge regression method (Muthukrishnan and Rohini, 2016), b) Extra trees classifier - This method appropriately ranks correlated features due to the high degree of stochasticity in node splitting during the construction of decision trees and exhibits a low probability of overfitting (Hemphill et al. 2014), and c) Recursive elimination method - This is a notable method for feature extraction in small biological datasets. It iteratively removes the least important feature and ranks them (Chen at al. 2018). The feature extracted by Ridge regression, extra trees classifier and recursive elimination methods were given in Fig 7, Fig 8 and Table 6, respectively.

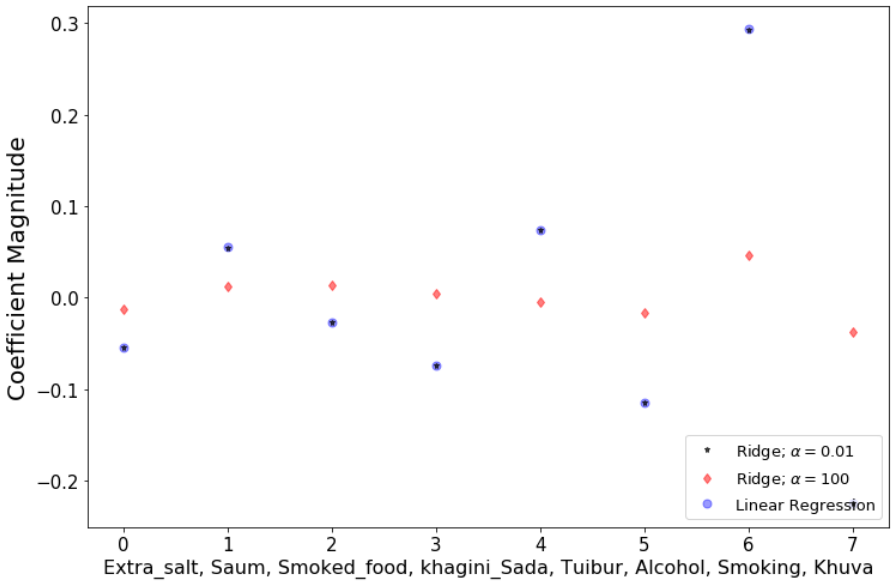


Fig 7.Feature selection by Ridge Regression

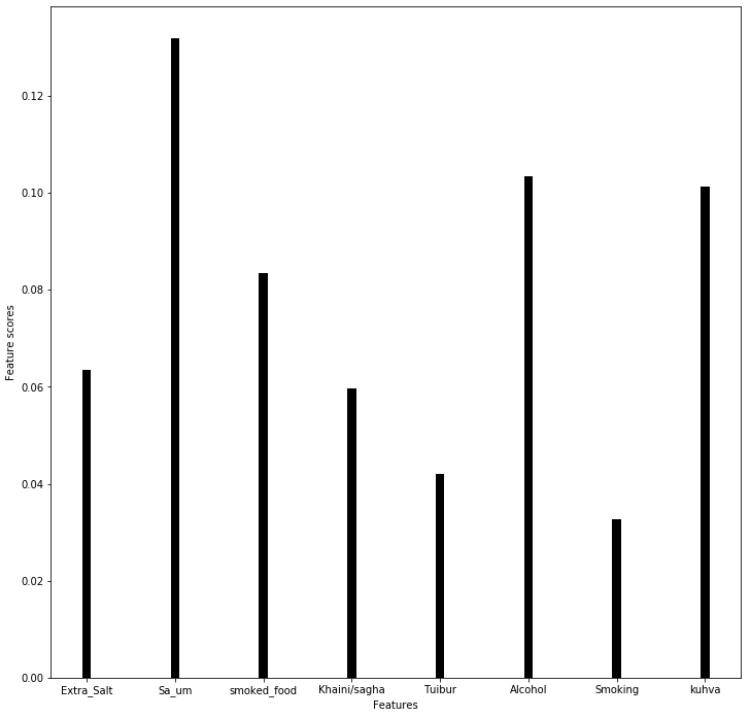


Fig 8.Feature selection by Extra trees classifier

Table 6.Feature Selection by Recursive Elimination Method

Features	**Selected Features**	**Feature Ranking**
Extra Salt	True	1
Saum	False	4
Smoked food	False	2
Khaini / Sadha	True	1
Tuibur	True	1
Alcohol	True	1
Smoking	True	1
Khuva	False	3

3.7.5 Machine Learning Models

The lifestyle-diet dataset was randomly divided in a 70:30 ratio; 95 instances allocated as training and 47 instances as testing set, with random state=0. Accuracy, precision, recall, F1 score, brier score, MCC, and Receiver operating characteristics (ROC) (Brier, 1950; Hajian-Tilaki, 2013; Chicco and Jurman, 2020) were significant

metrics to evaluate the performance of the models. A trade-off between accuracy, precision, recall, brier score, MCC, and ROC was considered to achieve a precise decision-making system. This approach was adopted because focusing solely on accuracy, precision, and recall might lead to misinterpretation of the outcomes. The test data was utilized exclusively to validate the performance of the data models and to obtain unbiased outcomes (Vabalas et al. 2019). The flow chart delineates data pre-processing, model building, validation, optimization, and result evaluation (Fig 9).The models were optimized by the hyperparameters using training dataset (Fig 10 and Fig 11).

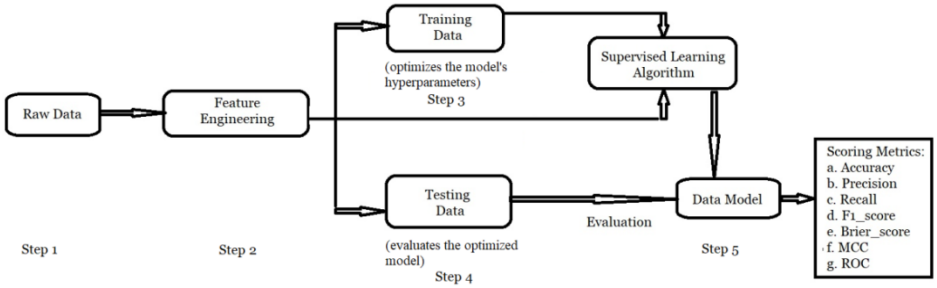


Fig 9. Flow chart of MSI-GC Classifier

```
Random_Forest_optimization
begin
        n_estimators = 100
        criterion = gini
        max_features = sqrt
        bootstrap = true
        random_state = 0
        min_samples_leaf = 2
end
```

Fig 10.Random Forest optimization using hyper parameters

```
Multilayer_perceptron_optimization
begin
        activation_fun = relu
        optimizer = lbfgs
        momentum = 0.9
        learning_rate = 0.001
        random_state = 1
        hidden_layer_neurons(three layers) = 25 neuros/layer
        max_iteration = 100
end
```

Fig 11.Multilayer Perceptron optimization using hyper parameters

3.7.5.1 Step by Step walk through of Flow Chart

Step 1: Raw diet-lifestyle-mutation data were collected and categorized based on MSI status.

Step 2: Feature selection was conducted utilizing Ridge regression.

Step 3: The dataset was partitioned into two-thirds for training and optimization with various hyperparameter configurations to fit and select the model.

Step 4: One-third of the dataset was utilized for testing and evaluating the optimized model.

Step 5: The models were evaluated by comparing the selected performance measures (accuracy, precision, recall, F1 score, ROC, MCC and brier score).

3.7.5.2 Evaluation Parameters

The presented dataset was balanced; therefore, the ROC curve was selected as one of the performance measures, which is considered the gold standard for evaluating biological binary classifiers, particularly for the positive class. MCC was regarded as

another significant evaluation parameter to compare the performance among the classifiers, as the highest MCC value of +1 is only achieved when the majority of positive and negative cases are correctly predicted (Chicco and Jurman, 2020). The Brier score was higher weightage than accuracy, as accuracy can potentially misrepresent the model's true potential on unknown data. Brier score is the mean squared error between expected and predicted values, ranging from 0 to 1, with smaller values indicating higher model performance (Rufibach, 2010).

3.8 Classification of Breast Cancer using Epidemiological Features

3.8.1 Dataset

The dataset consisted of 15 independent and 1 dependent features, with a total of 321 data points (Table 7) out of which 148 were breast cancer cases and 173 were healthy controls.

Table 7.Epidemiological Attributes for Breast Cancer Prediction

Attributes	Data type	Description
Sample Number	Continues	Identification number
How often do you exercise?	Categorical[a]	Daily; 2-4 times a week; Rarely; Never
How often do you take Fruits?	Categorical[a]	Daily; 2-4 times a week; Seasonal; Never
How often do you take Vegetables?	Categorical[a]	Daily; 2-4 times a week; Occasionally; Never
How often do you take Saum?	Categorical[a]	Daily; 2-4 times a week; Occasionally; Never
How often do you take Smoked meat?	Categorical[a]	Daily; 2-4 times a week; Occasionally; Never
How often do you take Smoked vegetables?	Categorical[a]	Daily; 2-4 times a week; Occasionally; Never
Intake of fried or oily Foods	Categorical[a]	Daily; 2-4 times a week; Occasionally; Never
Water intake per day	Discrete	More than 4 litres/day; 3-2 litres/day; 1 litre/day; less than 1 litre/day
Chewing_Pan	Categorical[a]	More than 2 packs/day; 1 pack/day; less than 1 pack/day; Occasionally
Gutkha	Categorical[a]	Yes – number of packs; No

Sahdah	Categorical[a]	Yes – number of packs; No
Tuibur	Categorical[a]	Yes – number of times; No
Smoking	Categorical[a]	Yes – number of packs; No
Alcohol	Discrete	Yes – number of pegs in a week; No
Any Inheritable Disease in the family?	Categorical[a]	First or second degree relatives with any cancers; First or second degree relatives with breast cancer
Class	Discrete	Cancer; Healthy

Categorical[a] data type was converted into numerical data type using one-hot encoding technique during the data preprocessing phase.

3.8.2 Data Preprocessing

The dataset had 7% missing value which may be attributed to patients' compromised physical condition during therapy. Missing values were replaced with their respective group median values after applying one-hot encoding (Das et al. 2019). A heatmap was constructed to determine if there was any multicollinearity between the attributes (Fig 12) which can significantly avoid data redundancies. Fig 13 illustrated the data distribution of cases and controls with respect to all attributes.

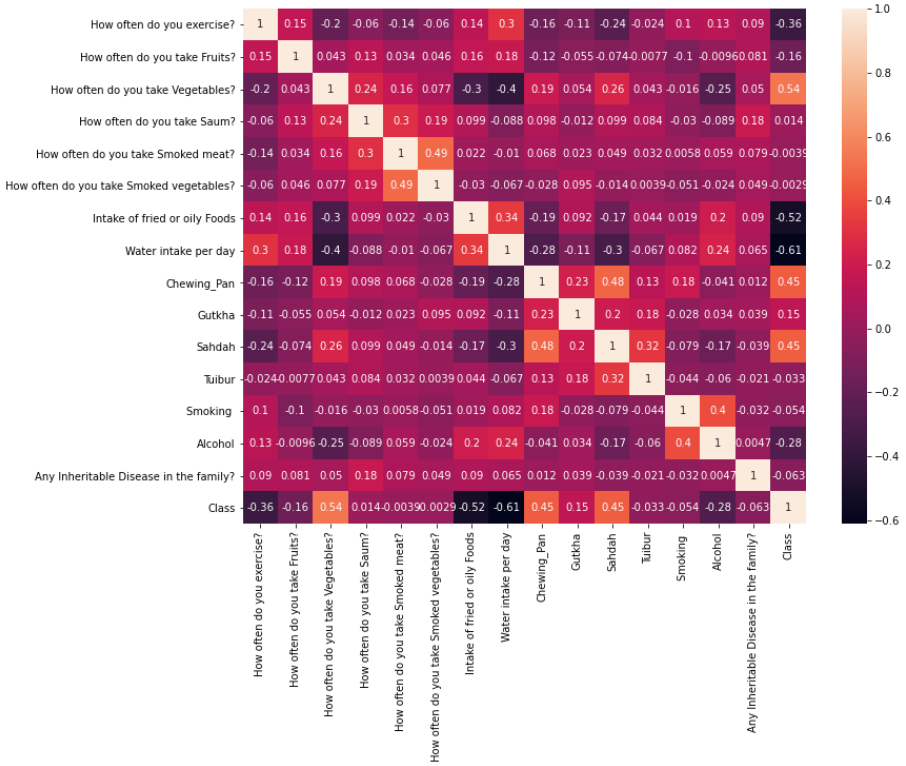


Fig 12. Correlation of features using Heatmap

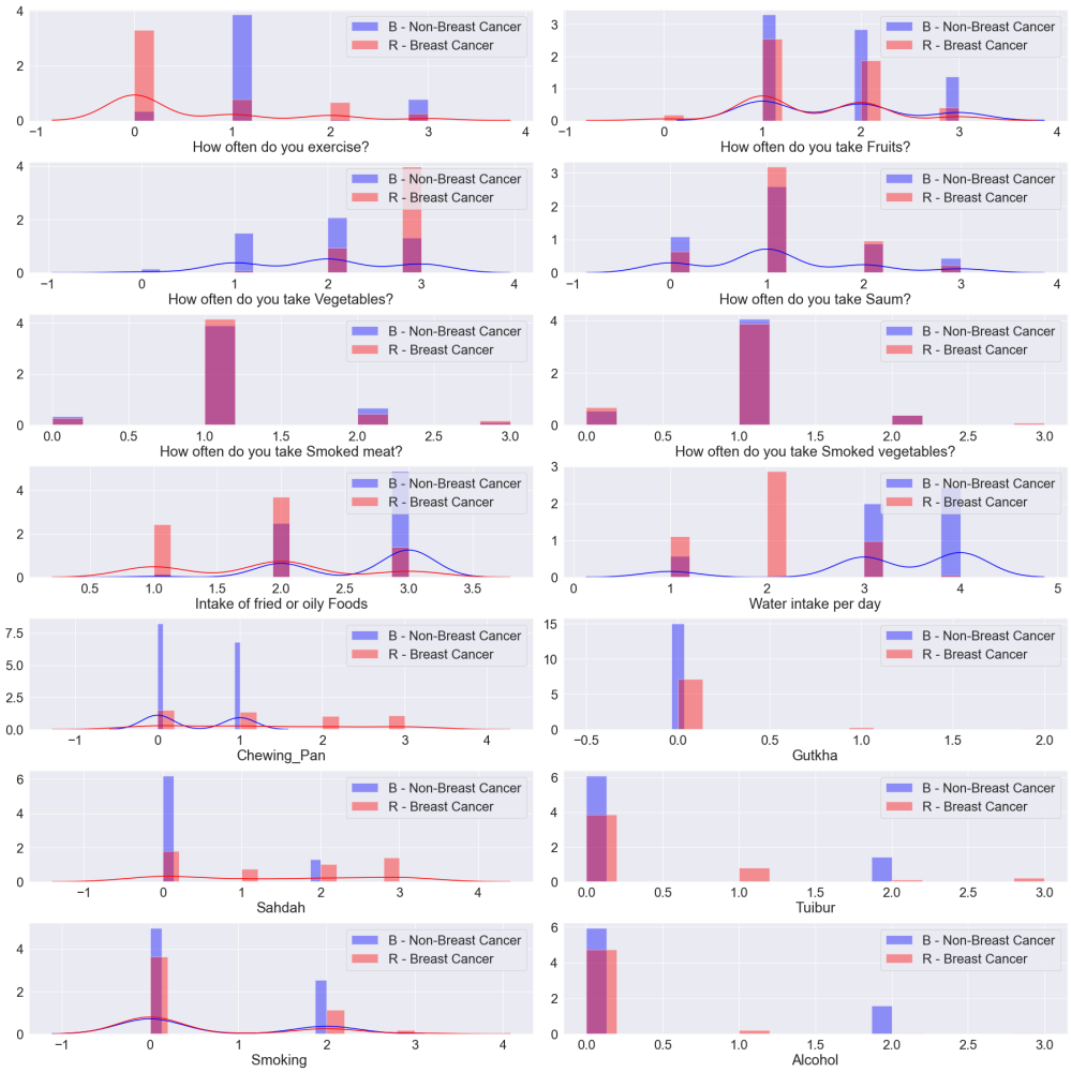


Fig 13.Data distribution of Breast Cancer (Cases) and (Non-Breast Cancer) Healthy Controls

3.8.3 Learning Curves

Learning curves of Gradient Boost, AdaBoost, Bagging,Extra trees, and Random forest classifiers were constructed to evaluate potential data underfitting (Fig 14.). These curves can rule out underfitting or overfitting based on their smooth convergences of training curves and validation curves. The models that exhibited minimal error discrepancy between the two curveswill only be considered to develop machine learning models.

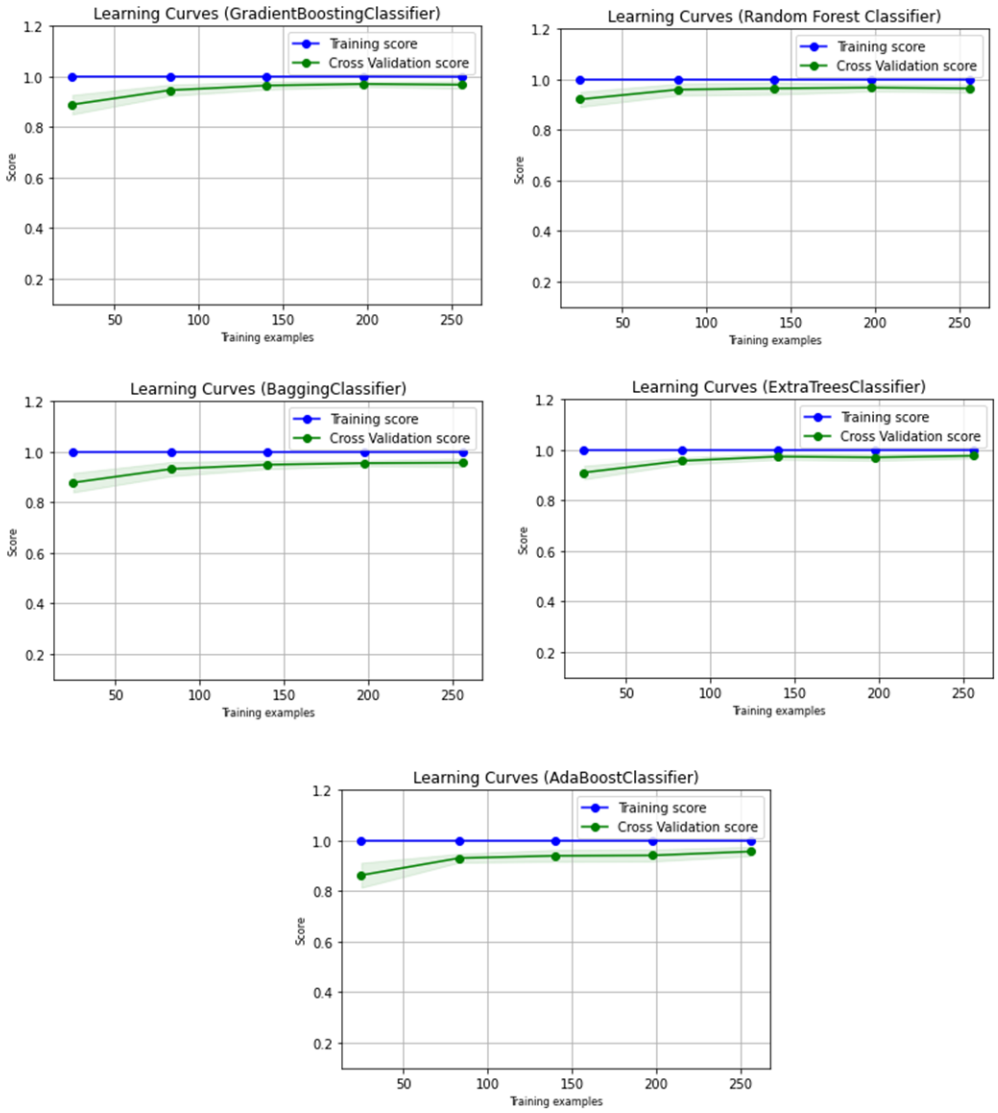


Fig 14. Learning Curves of Ensemble Classifiers for Breast Cancer Classifier

3.8.4 Ensemble Learning Algorithms

Five ensemble learning classifiers were selected and the performance was validated using leave one out cross validation (LOOCV) and 10-fold cross validation (10-fold CV). A flow chart of boosting and bagging classifiers was presented in Fig 15. Classification performance was assessed using repeated stratified k-fold cross-validation with 10 splits and 3 repeats in order to overcome the dataset's imbalance (Teh et al. 2020). Grid search with 10-fold CV were conducted on the entire dataset to determine the optimal hyperparameters for fine-tuning the models (Hosni et al.

2017). Various learning rates were established for boosting classifiers to attain the global minima and expedite convergence (Table 8).

Table 8. Models and their Hyperparameters

Models	No of estimators	Learning rates
Bagging Classifier	200	-
AdaBoost Classifier	100	0.134
Random Forest Classifier	80	-
Extra Trees Classifier	80	-
Gradient Boosting Classifier	100	0.1

3.8.5 FLOW CHART

Bagging Classifiers:

Step 1: Load the dataset

Step 2: Separate the target and independent variables

Step 3: Initialize Leave-One-Out Cross-Validation (LOOCV) and 10-fold Cross-Validation (CV) Step 4: Construct models for bagging classifiers

- *BaggingClassifier(n_estimators = 200, bootstrap = 'True')*
- *RandomForestClassifier(n_estimators = 80, bootstrap = 'True')*
- *ExtraTreesClassifier(n_estimators = 80)*

Step 5: Compute and record mean accuracies of LOOCV and 10-fold CV in lists

Boosting Classifiers:

Step 1: Load the dataset

Step 2: Separate the target and independent variables

Step 3: Initialize Leave-One-Out Cross-Validation (LOOCV) and 10-fold Cross-Validation (CV)

Step 4: Construct models for bagging classifiers

- *AdaBoostClassifier(n_estimators = 100, learning_rate = 0.134)*
- *GradientBoostingClassifier(n_estimators = 80, learning_rate = 0.1, max_features=9)*

Step 5: Calculate and record mean accuracies of LOOCV and 10-fold CV in lists

Step 6: Aggregate all mean accuracies of LOOCV and 10-fold CV

Step 7: Compute the Pearson correlation coefficient for LOOCV and 10-fold CV mean accuracies

Step 8: Visualize the accuracies utilizing scatter and line plots

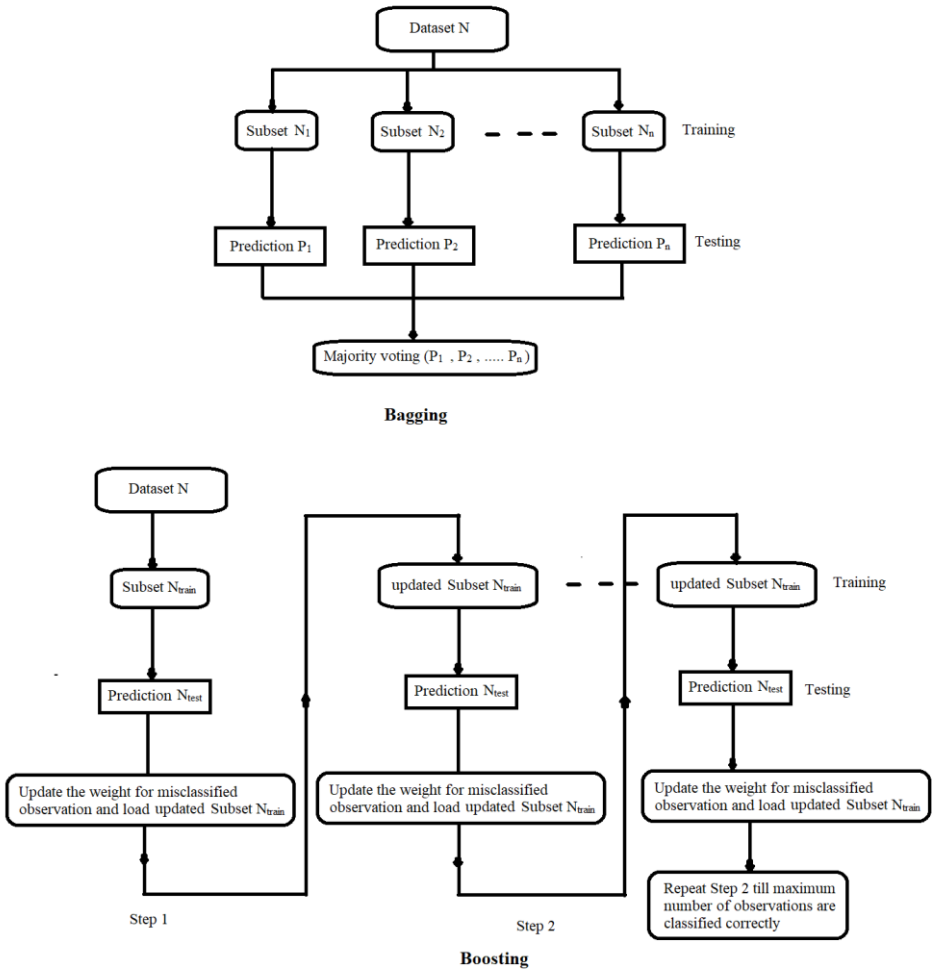


Fig 15. Flow chart of Bagging and Boosting Techniques for Breast Cancer Classification

3.9 Classification of Breast Cancer using Genomic Features

3.9.1 Breast Cancer Big Dataset

Initially, 203,380 variants were identified in breast cancer samples. After quality control and removal of missing values, 1,213 variants were filtered out. In healthy individuals, 4,407,672 variants obtained from the source, in which 755 variants retained following the removal of missing values and quality control in terms of depth and coverage. All variants from breast cancer samples were labeled as 1, while variants from healthy individuals were labeled as 0, as illustrated in Fig 16.

Columns containing missing values were removed, resulting in 42 features being subjected to data visualization and modeling (Table 9).The detailed steps undertaken from the acquisition of raw input to the achievement of the desired outputs was shown in flowchart (Fig 18). Thirty two breast cancer patients and twenty seven healthy individuals' genomic data had been merged for this breast cancer big data modeling, interpretation and visualization.

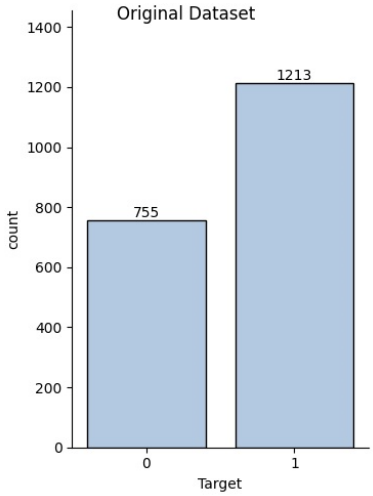


Fig 16. Breast Cancer Dataset with cases and control counts

Table 9. Breast Cancer Genetic Attributes

Genetic Attributes			
1	Chr	23	AF_raw
2	Start	24	AF_afr
3	End	25	controls_AF_popmax
4	Ref	26	AF_sas
5	Alt	27	AF_amr
6	Genes	28	AF_eas
7	AAChanges	29	AF_nfe
8	SIFT_score	30	AF_fin
9	Polyphen2_HDIV_score	31	AF_asj
10	Polyphen2_HVAR_score	32	AF_oth
11	LRT_score	33	non_topmed_AF_popmax
12	MutationTaster_score	34	non_neuro_AF_popmax
13	MutationAssessor_score	35	non_cancer_AF_popmax
14	FATHMM_score	36	ExAC_ALL
15	CADD_raw	37	ExAC_AFR
16	GERP_RS	38	ExAC_AMR
17	SiPhy_29way_logOdds	39	ExAC_EAS
18	phyloP46way_placental	40	ExAC_NFE
19	AF	41	ExAC_OTH
20	AF_popmax	42	ExAC_SAS
21	AF_male	43	Target
22	AF_female		

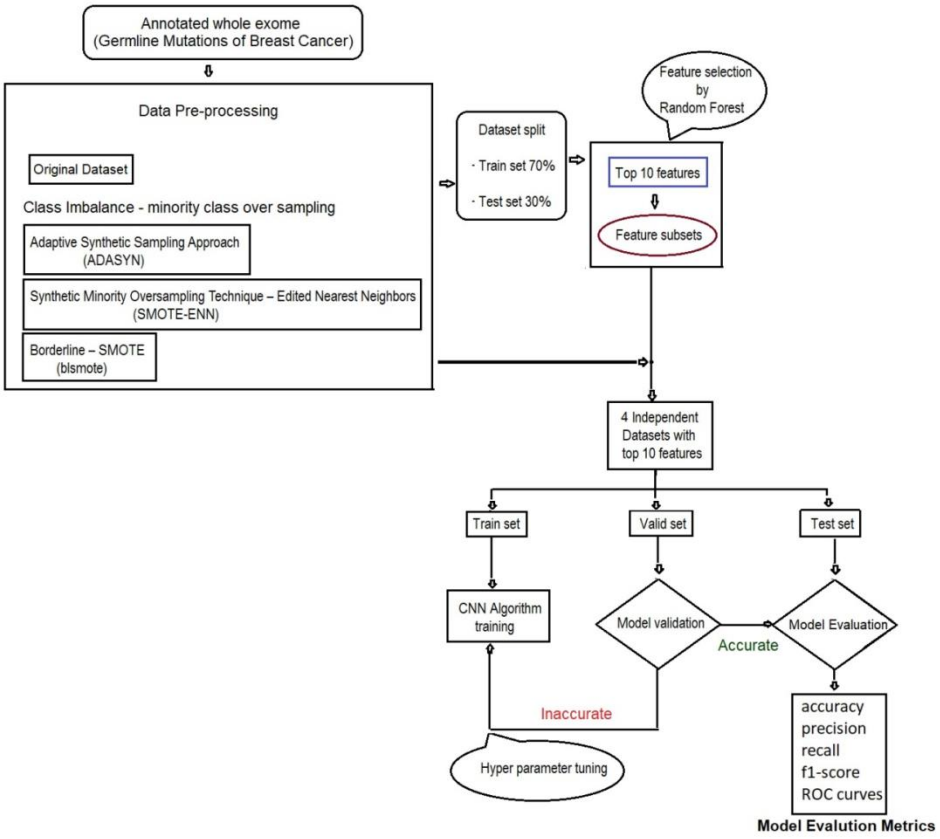


Fig 17.Flow chart for developing CNN classifiers for breast cancer germline mutations.

3.9.2 Oversampling of Minority Class

Through the application of oversampling techniques, researchers and data scientists can artificially augment the representation of minority classes in the dataset. This approach facilitates the creation of a more balanced distribution of classes, enabling machine learning algorithms to learn from a more representative sample. These methods generate synthetic examples of the minority class based on existing ground truth data points. This not only increases the sample size of the underrepresented class but also aids in capturing the underlying patterns and characteristics of the minority instances (Abdulsadig and Rodriguez-Villegas, 2024). Consequently, oversampling can lead to enhanced model performance, increased sensitivity to rare events, and more robust predictions in medical applications, ultimately contributing

to improved diagnostic accuracy and patient care outcomes. In accordance to the above, three methods were employed in this study to addresses the class imbalance issue in an effective manner:

- Adaptive Synthetic Sampling Approach (ADASYN)
- Synthetic Minority Oversampling Technique – Edited Nearest Neighbors (SMOTE-ENN)
- Borderline – SMOTE (blsmote)

Each of the aforementioned methods had generated an independent dataset. In total, four datasets were produced this includes the original raw dataset. The transformation of imbalanced datasets into balanced datasets can mitigate bias towards the minority class and facilitate more accurate feature selection.

3.9.3 Data Preprocessing

Missing values in attributes and instances were eliminated to ensure data consistency, mitigate bias in data modeling, and prevent potential impacts on appropriate clinical findings, drug target effectiveness, patient prognosis, and treatment outcomes. The dataset was divided into training, validation and test sets. The training dataset was scaled using a standard scalar to prevent misinterpretation commonly caused by machine learning models and to facilitate the identification of optimal solutions.

3.9.4 Feature Selection by Random Forest

Random Forest with hyperparameter tuning was employed to identify the key features contributing to variant pathogenicity. This method combines multiple decision trees to reduce variance. The inherent randomness of the model facilitates the evaluation of feature importance and is extensively utilized in datasets with high-dimensional features (Chen et al. 2020).

3.9.5. Feature Subsets

Feature selection in combination with oversampling techniques offers several advantages when generating feature subsets for machine learning tasks (Tsai et al. 2024). By applying feature selection methods to datasets that have undergone oversampling, this can identify the most relevant and informative features while addressing class imbalance issues. This approach helps to reduce dimensionality, mitigate the curse of dimensionality, and improve model performance by focusing on the most discriminative attributes. Ten significant features from the raw dataset (imbalanced), ADASYN dataset, SMOTEENN dataset, and blsmote, were selected to form four independent datasets (Fig 18 to 21).

Although critical evaluation metrics were available for assessing models generated using imbalanced datasets, these models may still exhibit overfitting and limited capacity to effectively learn minority classes. In numerous medical scenarios, the occurrence of certain conditions or diseases may be relatively infrequent compared to the overall population, resulting in an imbalanced representation in datasets. This imbalance can significantly impact the performance and reliability of machine learning models, particularly in tasks such as disease detection, risk prediction, or treatment outcome analysis. To address these issues, the breast cancer big data (imbalanced dataset) was oversampled using three distinct techniques to generate balanced datasets (Yang et al. 2024).

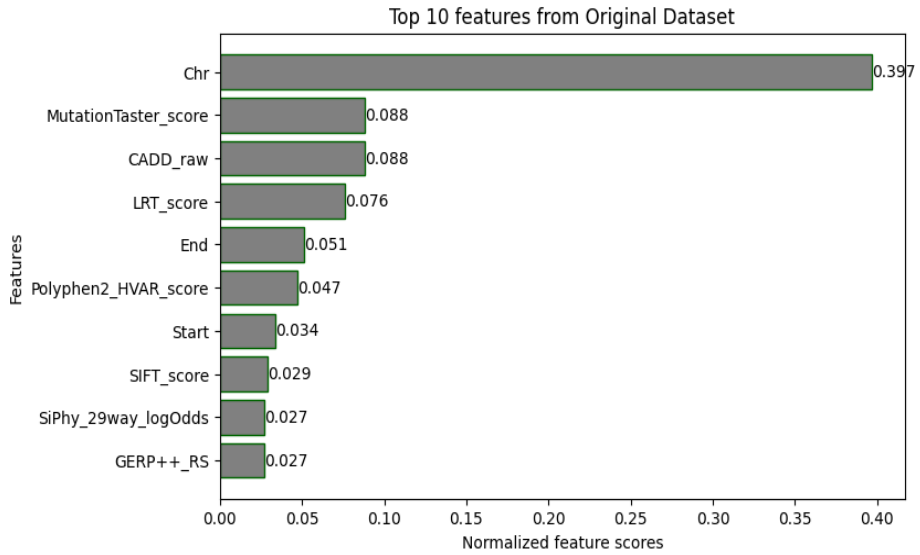


Fig 18. Top 10 features from raw dataset (imbalanced)

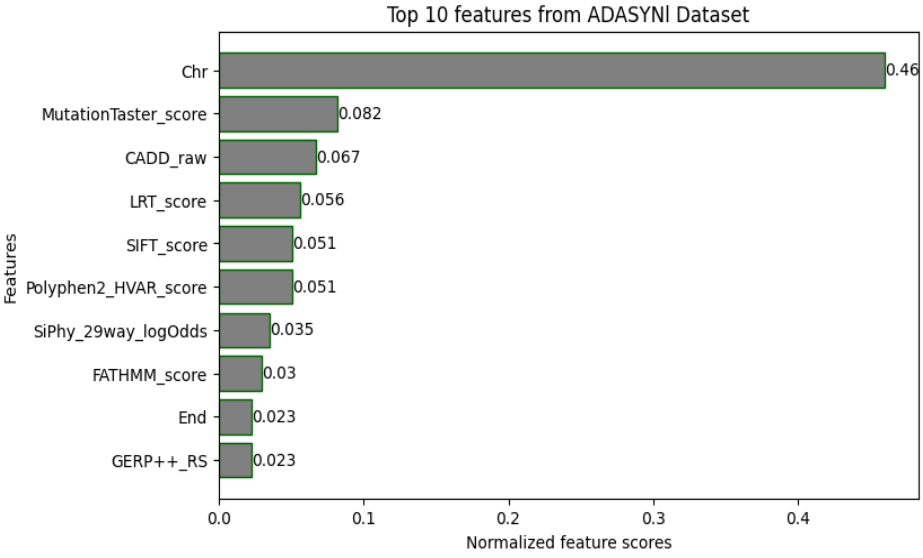


Fig 19. Top 10 features fromADASYN Dataset

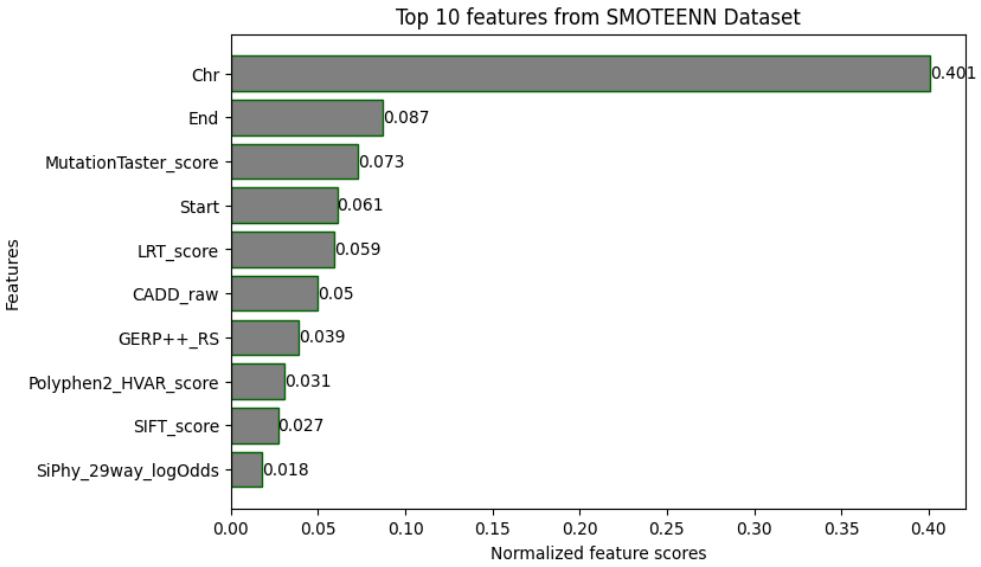


Fig 20. Top 10 features from SMOTEENN Dataset

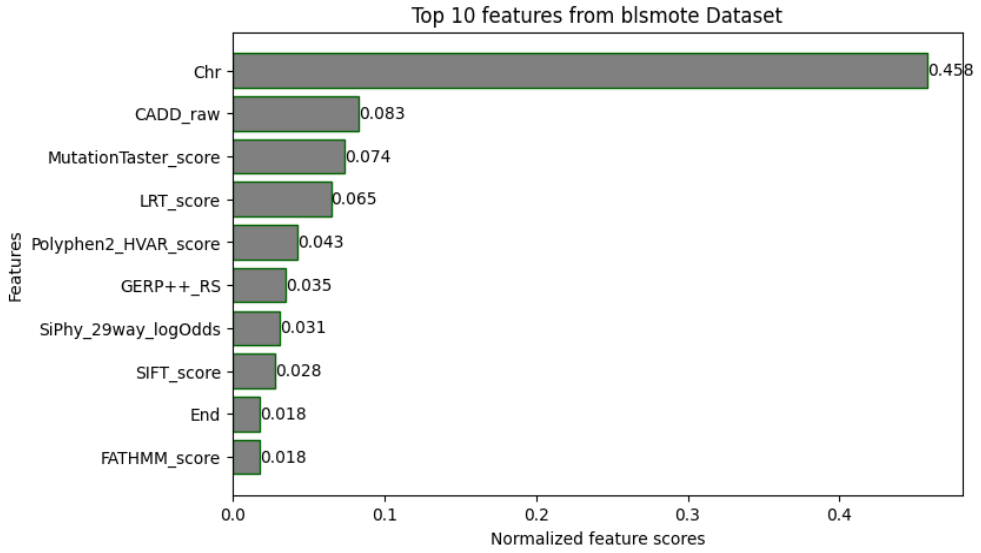


Fig 21. Top 10 features from blsmote Dataset

3.9.6 Convolution Neural Networks Model for Breast Cancer Big Data Classification

The hyperparameters of CNN models werebatch size of 32, 300 epochs, ReLU activation function for input and hidden layers, sigmoid activation function for the output layer, binary cross-entropy loss function, and Adam optimizer. These

hyperparameters were optimized across all four CNN models developed using the raw dataset (imbalanced), ADASYN, SMOTEENN, and BLSMOTE datasets. The dataset was partitioned as train (60%), valid (20%) and test (20%) to evaluate the performance of the CNN models. The CNN models successfully achieved global minima, as evidenced by the smooth and consistent training accuracy, validation accuracy, training loss, and validation loss graphs, which were plotted for the raw dataset (imbalanced), ADASYN, SMOTEENN, and BLSMOTE datasets (Fig 22-25).

The CNN models had three hidden layers which had progressive increase in dropouts rates of 60%, 70% and 80%across the hidden layers to serve multiple purposes. Firstly, it helps in reducing the model's reliance on specific features or neurons, thereby promoting a more robust and generalized learning process. The 60% dropout in the first hidden layer allows for a moderate level of feature retention while still introducing randomness. As the network progresses to deeper layers, the higher dropout rates of 70% and 80% force the model to learn more diverse and resilient representations of the data. This strategy not only combats overfitting but also enhances the model's ability to generalize well to unseen data, potentially improving its performance on validation and test sets (Salehin and Kang, 2023).

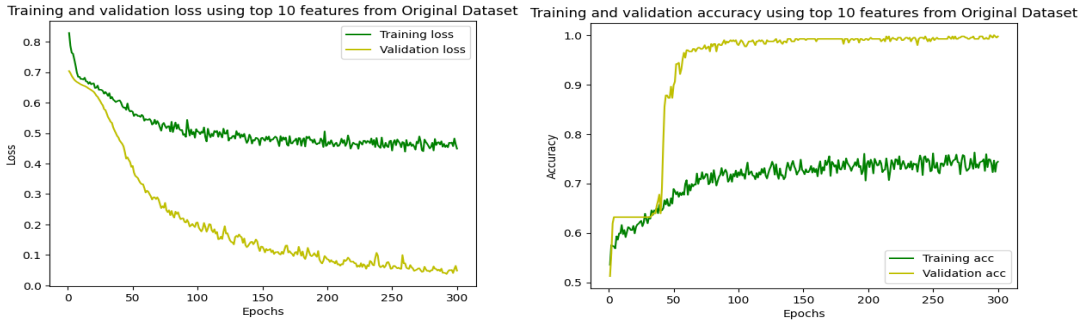


Fig 22. Loss and Accuracy on raw dataset by training and validation datasets

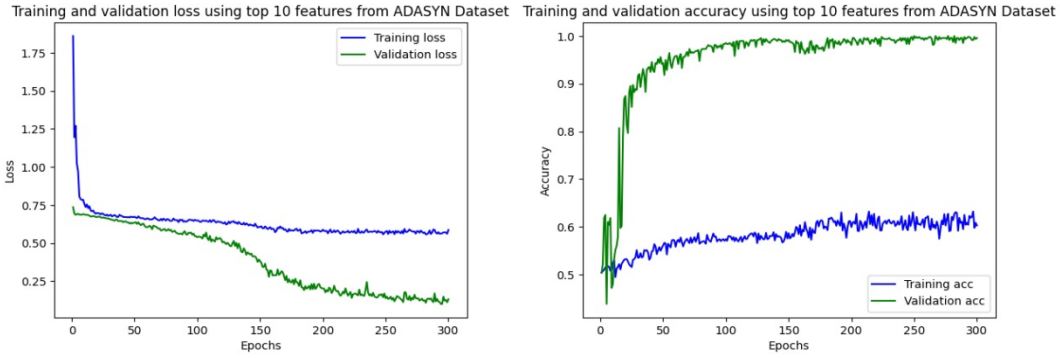


Fig 23. Loss and Accuracy on **ADASYN** Dataset by training and validation datasets

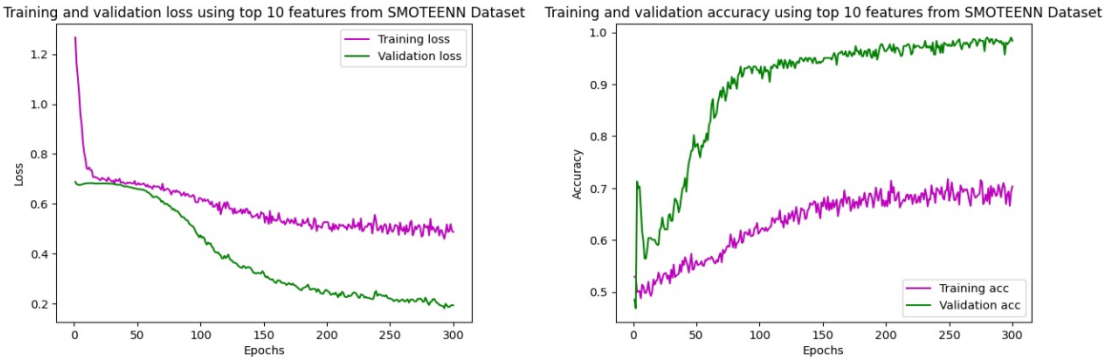


Fig 24. Loss and Accuracy on **SMOTEENN** Dataset by training and validation datasets

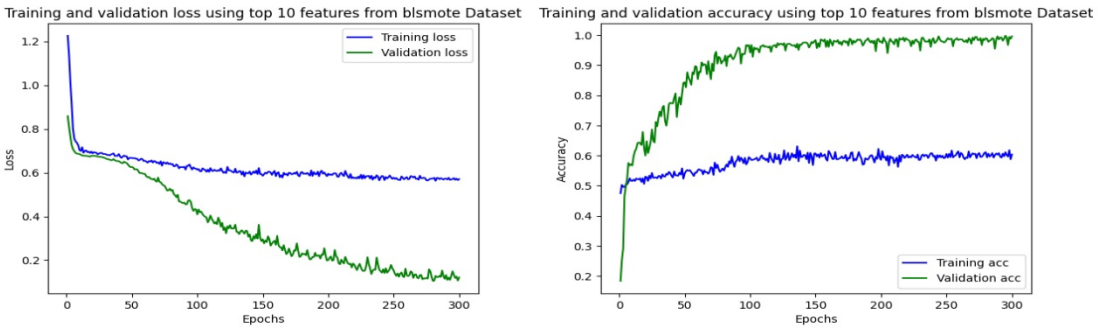


Fig 25. Loss and Accuracy on **blsmote** Dataset by training and validation datasets

3.10 Classification of Head & Neck Cancer using Epidemiological Features

3.10.1 Dataset

The dataset consisted of 14 independent and 1 dependent features, with a total of 300 data points (Fig 26). There were 200 instances for healthy individuals (controls) and 100 instances for head and neck cancer (cases).

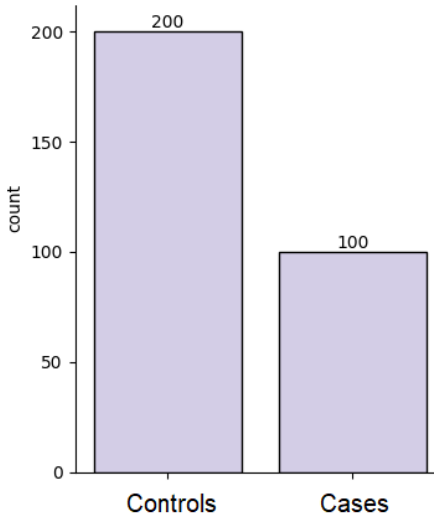


Fig 26.Head and Neck Cancer Dataset with Cases and Controls count

3.10.2 Data Preprocessing

The categorical variables were converted to numerical values using label encoder. The dataset was divided into train and test set with ratio of 3:1, respectively. Further, train set was scaled using StandardScaler and transformed on test set. The missing values and outliers were removed to enhance data integrity and reduce the negative implication on patient care.

3.11 Classification of Head and Neck Cancer using Genomic Features

3.11.1 Head and Neck Cancer Big Dataset

In total, there were 5,141,996 variants identified in head and neck cancer samples. Of these, only 60,824 variants were located in the region of interest (exonic). Following

quality control measures and removal of missing values, 5,784 variants remained. In healthy individuals, 4,407,672 variants were initially identified, with 755 variants retained after quality control procedures and removal of missing values, considering depth and coverage. All variants from head and neck cancer samples were labeled as 1, while variants from healthy individuals were labeled as 0, as illustrated in Fig 27. Columns containing missing values were excluded, resulting in 40 features had been subjected to data visualization and modeling, as presented in Table 10. Fig 28 depicts a flow chart detailing the steps undertaken from raw data acquisition to the attainment of desired results. Ten head & neck cancer patients and twenty seven healthy individuals' genomic data had been merged for this head and neck cancer big data modeling, interpretation and visualization.

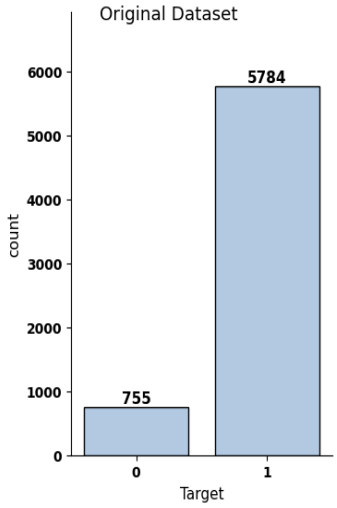


Fig 27.Number of observations by cases (head & neck cancer) and controls (healthy individuals)

Table 10.Head and Neck Cancer Genetic Attributes

Genetic Attributes			
1	Chr	22	controls_AF_popmax
2	Start	23	AF_sas
3	End	24	AF_amr
4	Ref	25	AF_eas
5	Alt	26	AF_nfe
6	Genes	27	AF_fin
7	AAChanges	28	AF_asj
8	SIFT_score	29	AF_oth

9	Polyphen2_HVAR_score	30	non_topmed_AF_popmax
10	LRT_score	31	non_neuro_AF_popmax
11	MutationTaster_score	32	non_cancer_AF_popmax
12	FATHMM_score	33	controls_AF_popmax
13	CADD_raw	34	ExAC_ALL
14	GERP_RS	35	ExAC_AFR
15	SiPhy_29way_logOdds	36	ExAC_AMR
16	AF	37	ExAC_EAS
17	AF_popmax	38	ExAC_NFE
18	AF_male	39	ExAC_OTH
19	AF_female	40	ExAC_SAS
20	AF_raw	41	Target
21	AF_afr		

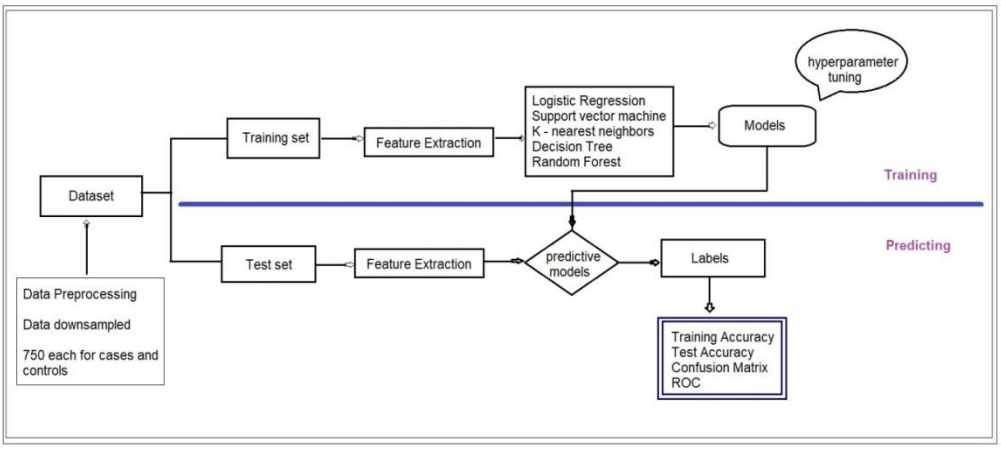


Fig 28. Flow chart of head and neck cancer classification

3.11.2 Random Downsampling

Random downsampling in machine learning is a technique that can address class imbalance issues by reducing the number of samples in the majority class. This approach has the potential to enhance model performance by mitigating algorithmic bias towards the overrepresented class (Lee and Seo, 2022). Furthermore, random downsampling may result in reduced training times and decreased computational resource requirements, as the model operates on a smaller dataset. The presented raw dataset was highly imbalanced as shown in the Fig 27. Thereby, random downsampling was performed to make it a balanced dataset as shown in Fig 29.

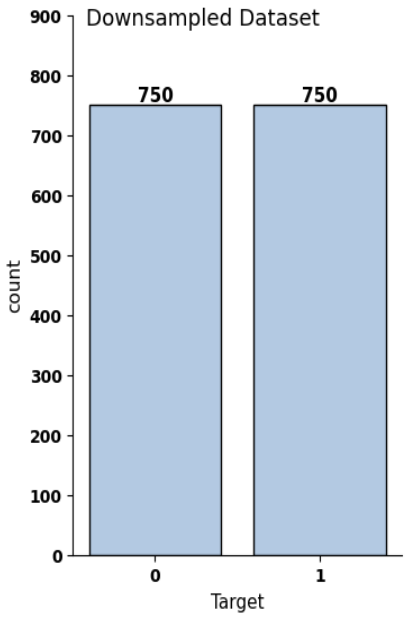


Fig 29. Number of cases (HNC) and controls (healthy individuals) after random downsampling of cases

3.11.3 Data Preprocessing

The dataset was divided into training and test sets. The training dataset was scaled using MinMax scaler. Missing values in attributes and instances were removed to mitigate potential bias in data modeling.

3.11.4 Feature selection using Extra Trees

The Extra Trees classifier, optimized through hyperparameter tuning, was employed to identify the key features contributing to variant pathogenicity. This ensemble method builds multiple decision trees and introduces additional randomness compared to traditional Random Forests. The algorithm's hyperparameters were fine-tuned to optimize its performance, ensuring the most effective identification of key features within the dataset. The optimization process involved adjusting various parameters such as the number of trees, maximum depth, and minimum samples split (Table 11). This meticulous tuning allowed the Extra Trees classifier to better capture the underlying patterns and relationships in the data, ultimately leading to a more robust and accurate selection of important features (Arya et al. 2022). By

employing this optimized approach, the most influential variable that contribute significantly to the model's predictive power can be identified, thereby enhancing the overall understanding of the problem at hand and potentially improving the subsequent analysis or decision-making processes.

Table 11.Extra Trees Hyperparameters Values

Hyperparameters	Values
n_estimators	250
max_depth	6
min_samples_split	15
Bootstrap	True

3.11.5 Multicollinearity Features in Dataset

Heat map was generated using the head and neck big dataset (Fig 30). Heat maps provide a visual representation of the correlation matrix, enabling to expeditiously identify patterns and relationships between variables (Lysdahlgaard, 2023). The color intensity in the heat map indicates the magnitude of correlation, with darker hues typically denoting stronger correlations. This visual methodology facilitates the identification of clusters of highly correlated features, thereby making informed decisions regarding the inclusion or exclusion of variables in their analysis to mitigate multicollinearity issues.

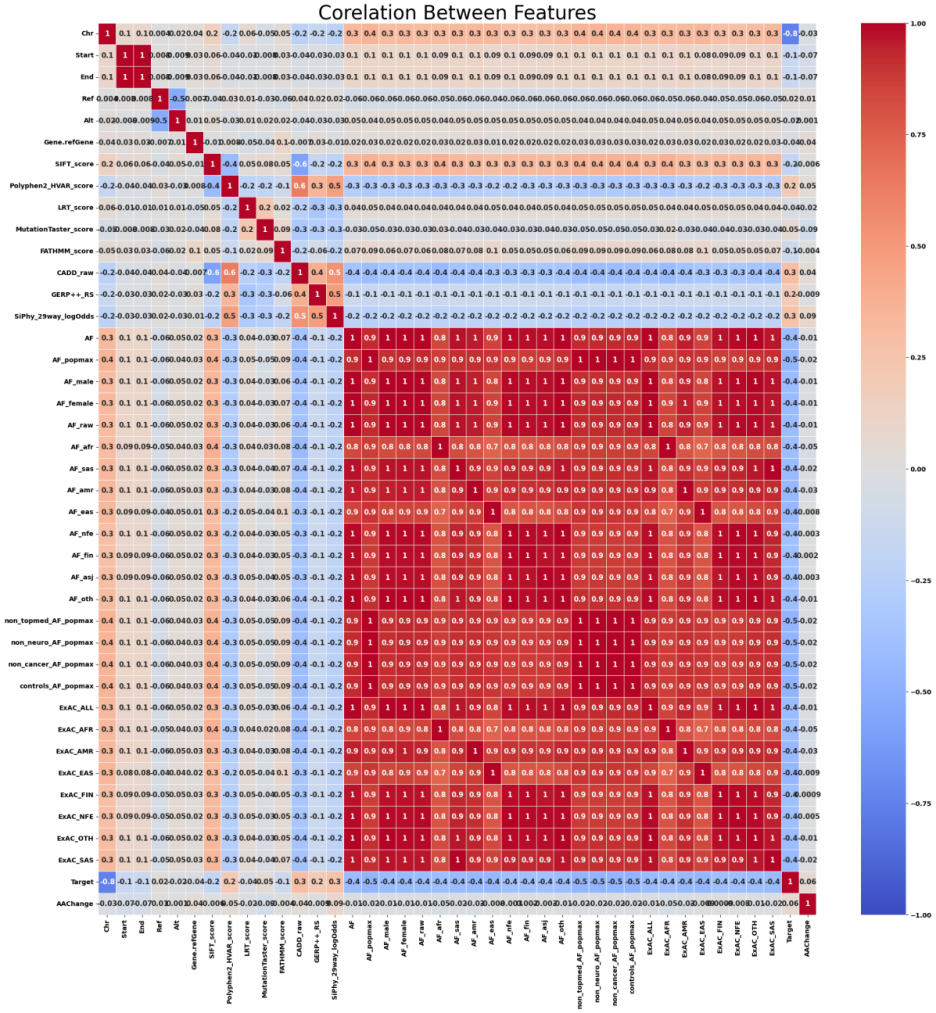


Fig 30. Heat Map of Head and Neck Germline (Big) Dataset

3.11.6 Data Models

Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and logistic regression classifiers were employed to develop models utilizing a large head and neck cancer dataset. The hyperparameters for these models were optimized using the training set, and their efficacy was subsequently evaluated using the reserved test dataset. Table 12, Table 13, Fig 31, Table 14, and Table 15 presented the optimal hyperparameter values established during the training phase for Random Forest, Decision Tree, KNN, SVM, and logistic regression, respectively.

The k value in the KNN algorithm is a critical hyperparameter that significantly influences the model's performance and predictive accuracy. The number of nearest neighbors taken into account while generating predictions for fresh data points is determined by k value.Determining the optimal k value necessitates careful consideration and often involves a trade-off between bias and variance (Fig 31). A small k value may result in overfitting, wherein the model becomes sensitive to noise in the train data and fails to generalize the unseen data. Conversely, a large k value may lead to underfitting, wherein the model oversimplifies the decision boundary and fails to capture significant patterns in the data (Zhang, 2016).

The hyperparameters gamma and C play crucial roles in SVM models, significantly influencing their performance and behavior (Jordaan and Smits, 2002). Gamma, a parameter for non-linear SVMs, determines the influence of a single training example on the decision boundary. A low gamma value results in a more generalized decision surface with a larger influence range, while a high gamma creates a more complex decision boundary that closely fits the training data (Shadeed et al. 2020). The C (regularization parameter) balances between low training error and low model complexity (Table 14). In contrast to a big C value, which seeks for a stronger classification of training points and causes overfitting, a small C value favors a smooth decision border, permitting certain misclassifications.In conjunction, gamma and C significantly impact the SVM model's generalization ability, with their optimal values often determined through techniques such as grid search or cross-validation to achieve optimal performance on unseen data.

Two significant aspects of logistic regression are the hyperparameter penalty l1 and the solver saga, which can substantially influence the model's performance and interpretability. By employing the l1 penalty, logistic regression can generate more interpretable models and mitigate overfitting. The saga solver, conversely, is an optimization algorithm with l1 regularization and a variant of the Stochastic Average Gradient (SAG) method can accommodate both l1 and l2 penalties and is recognized for its rapid convergence properties (Table 15).

Table 12. Random Forest Optimizers

Hyperparameters	Values
Criterion	entropy
max_depth	6
min_samples_leaf	6

Table 13. Decision Tree Optimizers

Hyperparameters	Values
max_depth	5
min_samples_leaf	5

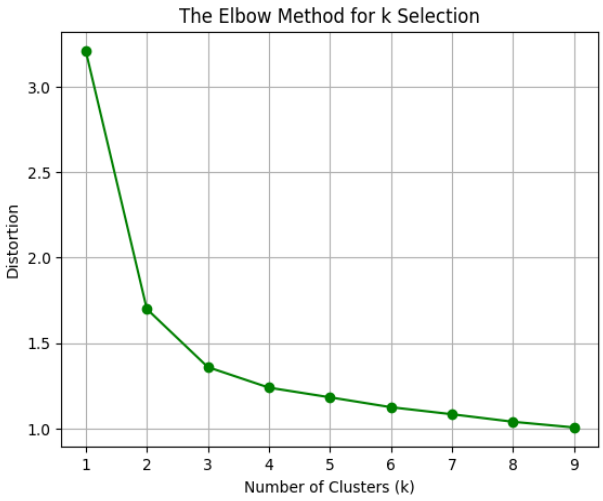


Fig 31. KNN optimizer

Table 14. Support Vector Machine Optimizers

Hyperparameters	Values
C	10
gamma	0.1
Kernel	linear
Probability	True

Table 15.Logistic Regression Optimizers

Hyperparameters	Values
c	0.2
max_iteration	200
penalty	l1
Solver	saga

3.12 Multi-class classification by integrating Gastric, Breast and Head & Neck Cancer features

3.12.1 Data

The dataset comprised five independent features and one dependent feature, with a total of 590 observations across all cancer types (gastric, breast, and head and neck cancer) and healthy individuals (Table 16). The dataset contained 200, 173, 117, and 100 observations from healthy individuals, breast cancer patients, gastric cancer patients, and head and neck cancer patients, respectively (Fig 32).

Table 16.Dataset Description by integrating three cancer types

S.No	Features	Data type
1.	Kuhva	Categorical
2.	Tuibur	Categorical
3.	Alcohol	Categorical
4.	Snuff	Categorical
5.	Smoked food	Categorical

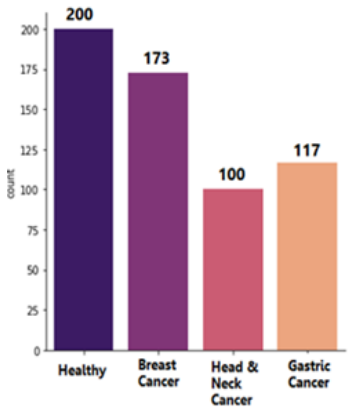


Fig 32.Integrated Datasets and counts by Classes

3.12.2 Data oversampling

The dataset exhibits uneven class distributions; consequently, Synthetic Minority Over-sampling Technique (SMOTE) was employed to oversample gastric, breast, and head & neck cancer data (Fig 33). SMOTE, or Synthetic Minority Over-sampling Technique, is a widely utilized method in machine learning to address class imbalance problems (Elreedy and Atiya 2019). This technique functions by generating synthetic examples of the minority class rather than merely duplicating existing instances. SMOTE operates in the feature space, selecting minority class samples and their k-nearest neighbors to create new synthetic samples along the line segments joining these points. The primary advantage of SMOTE is its capacity to enhance the performance of classifiers on imbalanced datasets without inducing overfitting. SMOTE facilitates the expansion of the decision region of the minority class, rendering it more general and robust (Yang et al. 2024).

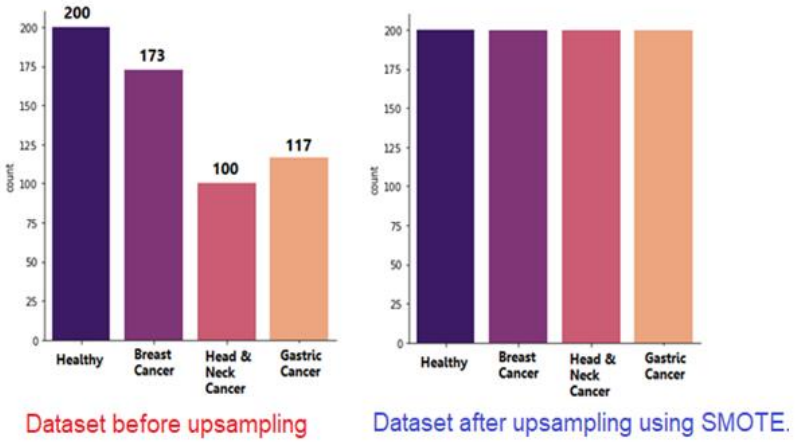


Fig 33. Size of the Dataset before and after upsampling by SMOTE

3.12.3 Feature Selection using Mutual Info Classifier

The integrated dataset was subjected to feature extraction using mutual info classifier. Feature extraction by mutual information (MI) classifier has a robust methodology as it computes the mutual information between each feature and the target variable (Hoque et al. 2014). This method effectively ranks and selects features that are most informative for classification tasks (Fig 34). The procedure typically calculates the MI score for each feature-target pair and subsequently selecting the highest-scoring features. Furthermore, MI-based feature extraction is non-linear and capable of capturing complex relationships between variables that linear methods may fail to detect.

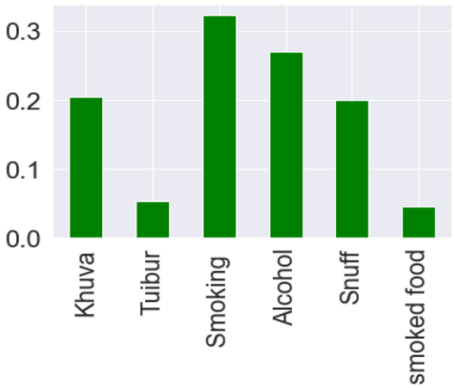


Fig 34.Feature selection using mutual info classifier

3.12.4 Multilayer Perceptron Model

Multilayer perceptrons (MLPs) offer several advantages for multiclass classification tasks. Their capacity to learn complex, non-linear decision boundaries renders them well-suited for handling datasets with intricate relationships between features and multiple classes (El-Hassani et al. 2024). MLPs can automatically extract hierarchical representations of the input data through their hidden layers, enabling them to capture subtle patterns and interactions that may not be apparent in the raw features. MLPs can efficiently scale to accommodate many output classes by simply adjusting the number of neurons in the output layer. Furthermore, MLPs can be readily adapted to incorporate various loss functions and regularization techniques, allowing for fine-tuning of the model's performance and generalization capabilities. This part of the work had presented an optimized multilayer perceptron model that was specific to this integrated dataset. The detailed steps of the algorithm were presented as a workflow in Fig 35.

Step 1: Load the dataset
Step 2: Split the dataset into 70% train and 30% test with random_state=3
Step 3: Build Multi-layer perceptron model for 4-class classification
- *MLPClassifier(activation='relu', learning_rate=0.01, momentum=0.98, max_iter=500, Solver='lbfgs', hidden_layer_sizes=h*

Step 4: h = [24,16,8], [16,8,6], [16,8,4], [16,8]
Step 5: Calculate accuracy for above four different hidden layer set
Step 6: Opted the model which gives good accuracy with low computational cost.

Fig 35. Workflow of an Optimized Multilayer Perceptron Model

3.12.5 Ensemble Data Models on Big Data for three cancer types

The dataset consisted of variants that has coverage of 100% from gastric (8368 instances), breast (4067 instances) and head & neck (4920 instances) cancer types. The pathogenic and benign categories were assisgned if and only if all the five tools (MutationTaster, FATHMM, LRT, SIFT, and Polyphen) either predicted a variant to

be pathogenic or benign, respectively. A balanced dataset was developed with 1886 variants in each of the above cancers. Preliminary data preprocessing were done by scaling the values and transforming categorical variables into numerical variables using label encoder. AdaBoost, Random forest, Extra trees and Bagging classifiers were used to develop best classification models. These models were fine tuned using 5-fold GridSearchCV to identify the global minima and to minimize the errors in the final classification results on the test dataset.

Chapter 4. RESULTS

4.1 Gastric Cancer Classification from Epidemiological Data

The proposed logistic regression model demonstrated a promising accuracy of 98.5% by achieving the global minimum with an intercept of 0.22 and a learning rate of 0.000915, as illustrated in Fig 36. Saum (fermented pork fat), the use of aluminum cookware, plastic utensils for food storage, smoking of Zozial (infiltrated local cigar), smoking before the age of 18,passive smoking, sadha (smokeless tobacco product), tuibur (aqueous tobacco extract), chewing betel nuts (kuhva, zardhapan, supari), residing or working in proximity to cell phone towers, and family history of cancer were identified as epidemiological risk factors for gastric cancer in the Mizoram population.

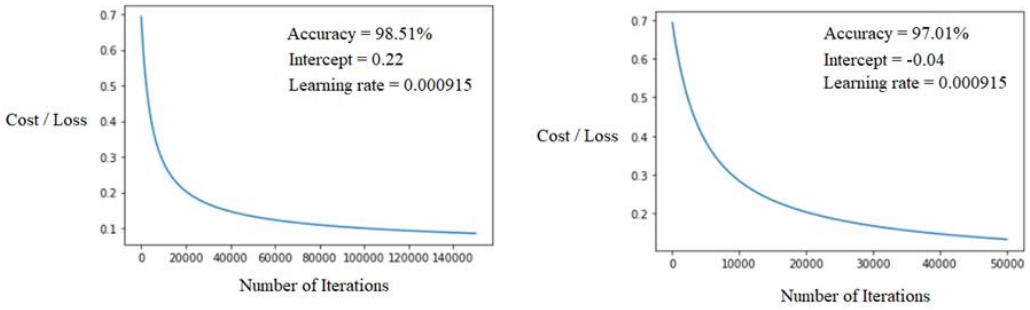


Fig 36. Cost vs. Number of iterations by Logistic Regression

4.2 Gastric Cancer Classification from Big Data

4.2.1 Feature Selection

LASSO model was experimentally tuned based on 5-fold CV with hyperparameters: max iteration was 100000 and tolerance was 0.000001, which had found fifteen significant features for causing pathogenicity in gastric cancer. AAChanges, CADD_raw, non_neuro_AF_popmax, ExAC_AMR, Ref, Genes, ExAC_OTH, ExAC_SAS, and Start were found to be negatively correlated; whereas CADD_phred, VEST3_score, phyloP46way_placental, AF_asj, AF_eas, and Alt were positively correlated(Fig 5 and Table 3).

4.2.2 Hyperparameter Optimization of Ensemble Models

Features selected by the 5-fold LASSOCV were evaluated using five ensemble models. The models were optimized utilizing identical hyperparameters for both balanced and imbalanced datasets. Gradient Boosting was optimized with a learning rate of 0.01 and 1000 estimators. AdaBoost achieved the desired results with a learning rate of 0.05 and 100 estimators. The bagging classifiers employed max_depth as a crucial hyperparameter, which generally controls model overfitting. Random Forest's optimizers consisted of 50 estimators and a max depth of 50, while Extra Trees utilized the same number of estimators but had a max_depth of 500. The Bagging Classifier employed decision trees as its weak learner and was optimized with 1000 iterations (Table 17).

These models demonstrated optimal performance on both training and testing datasets using balanced and imbalanced datasets. None of the models exhibited overfitting issues, as evaluation metrics consistently improved from training to testing. The performance of the ensemble models was presented in tabular form and graphical representation (Table 18, Fig 37 - 40).

4.2.3 Extra Trees Classifier

The Extra Trees classifier applied to the balanced dataset demonstrated the highest and most effective performance across all metrics, with 96.06% accuracy, 95.74% precision, 96.33% recall, 96.0F1-score 4, ROC_AUC of 99%, MCC of 0.92, and Brier score of 0.04 (Table 18, Fig 37 and 40). Its performance on the imbalanced dataset exhibited the highest accuracy at 97.57% and lowest Brier score at 0.02 (Table 18 and Fig 39). It was also observed that there were marginally uneven values across all metrics: precision 95.62%, recall 90%, F1-score 92.73%, PR-AUC 98%, and MCC of 0.91 (Table 18, Fig 38 and 39). These results indicated a more robust generalization capability for both positive and negative classes on the balanced dataset.

4.2.4 Random Forest Classifier

Random forest had become the second-best performing model on a balanced dataset showed accuracy 94.39%, precision 93.67%, recall 95.11%, F1-score 94.39%, ROC-AUC 98%, MCC 0.89, and brier score 0.06 (Table 18, Fig 37 and 40). This model could better identify most pathogenic variants. When comparing these values with the imbalanced dataset metrics, there were significant decreases in recall (84.71%) and F1-score (88.75%) (Table 18). Although the accuracy of the imbalanced dataset was 96.31%, which was found to be superior to that of the balanced dataset, this parameter alone cannot be considered significant in terms of assessing the model's performance (Table 18, Fig 36 and 39).

4.2.5 Bagging Classifier

On the balanced dataset, it achieved an accuracy of 93.94%, precision of 93.62%, recall of 94.19%, F1-score of 93.90%, ROC-AUC of 98%, MCC of 0.88, and Brier score of 0.06 (Table 18, Fig 37 and 40). The classifier on the imbalanced dataset demonstrated slightly higher margins in evaluation metrics, with an accuracy of 97.06%, which was comparable to the extra trees accuracy on the imbalanced dataset. Precision, recall, F1-score, and ROC exhibited a minor decrease of 93.52%, 89.12%, 91.27%, and 96.1%, respectively (Table 18, Fig 36 and 39). The results obtained using the balanced dataset with extra trees indicated a similar outcome for the bagging classifier.

4.2.6 Boosting Models

AdaBoost and Gradient boosting algorithms demonstrated moderate performance in determining the pathogenicity of the variants (Table 18, Fig 37-40). Precision and recall scores on the balanced dataset were notably more favorable compared to values obtained from imbalanced datasets (Table 18 and Fig 37-40). Gradient boost on the imbalanced dataset yielded a precision of 86.12% and recall of 71.18%, whereas on the balanced dataset, it achieved a precision of 90.30% and recall of 91.13% (Table 18 and Fig 37-40). Conversely, AdaBoost with the

imbalanced dataset produced a precision of 83.46% and recall of 66.76%, while performing more effectively on the balanced dataset with a precision of 88.33% and recall of 85.63% (Table 18 and Fig 37-40).In the aforementioned results of bagging classifiers and boosting classifiers, all models demonstrated satisfactory performance on the balanced dataset; consequently, accuracy, MCC, and Brier score potentially can led to misinterpretation on the imbalanced dataset.

4.2.7 Highly Mutated Genes and Genes Functional Connectivity in Gastric Cancer

MADD, CAPRIN2, KANK1, SUN1, NRCAM, and USP36 had more than 100 mutations in gastric cancer germline dataset (Fig 41).The MADD gene family has been implicated in various cellular processes associated with gastric cancer progression. Overexpression of certain MADD genes can result in increased cell proliferation and resistance to apoptosis in gastric cancer cells.CAPRIN2 has been implicated in various cellular processes that may contribute to cancer development and progression. CAPRIN2 can modulate cell proliferation, apoptosis, and migration, which are fundamental characteristics of cancer cells. The SUN1 gene plays a crucial role in maintaining nuclear envelope integrity and cellular stability and reduced SUN1 expression was associated with enhanced cancer progression and unfavorable prognosis across diverse malignancies.

The graphical representation had shown the interaction of all genes collected from the present gastric cancer germline data (Fig 42). There was evidence of hotspots of gastric cancer genes which were exhibiting independent interaction between nodes (genes). These were experimentally proven functional connectives explored by string database.

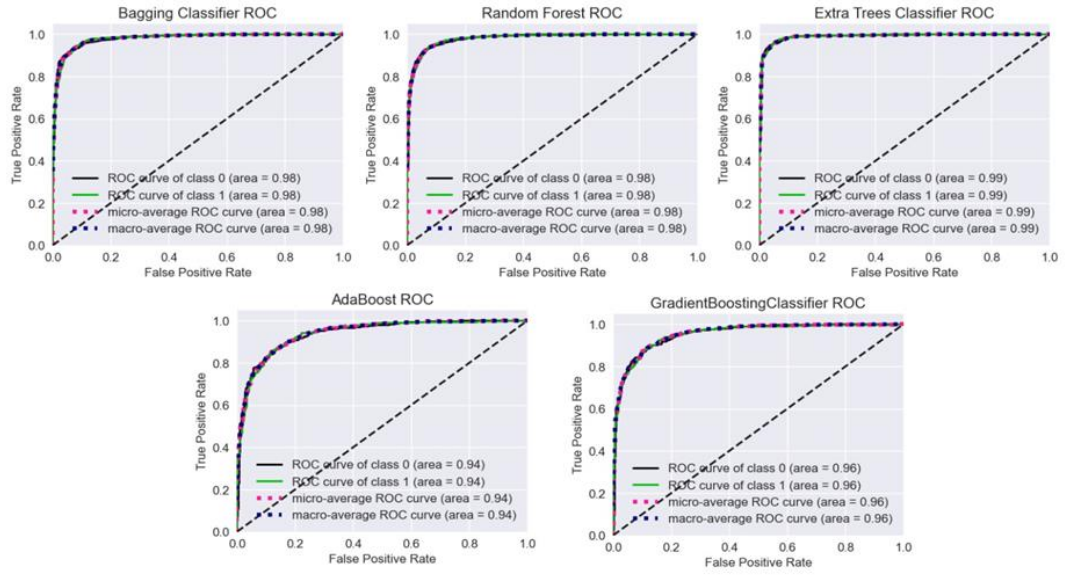
Table 17. Hyperparameter Optimization forEnsemble Models

Feature subsets	**Models**	**GridSearchCV Hyperparameters**
Top 15 From Imbalanced Dataset	Gradient boost	learning_rate=0.01, n_estimators=1000
	AdaBoost	learning_rate=0.05, n_estimators=100
	Random Forest	max_depth=50, n_estimators=50
	Bagging	estimator=DecisionTree(), n_estimators =1000
	Extra Trees	max_depth = 500, n_estimators=50
Top 15 From undersampled (Balanced Dataset)	Gradient boost	learning_rate=0.01, n_estimators=1000
	AdaBoost	learning_rate=0.05, n_estimators=100
	Random Forest	max_depth=50, n_estimators=50
	Bagging	estimator=DecisionTree(), n_estimators=1000
	Extra Trees	max_depth=500, n_estimators=50

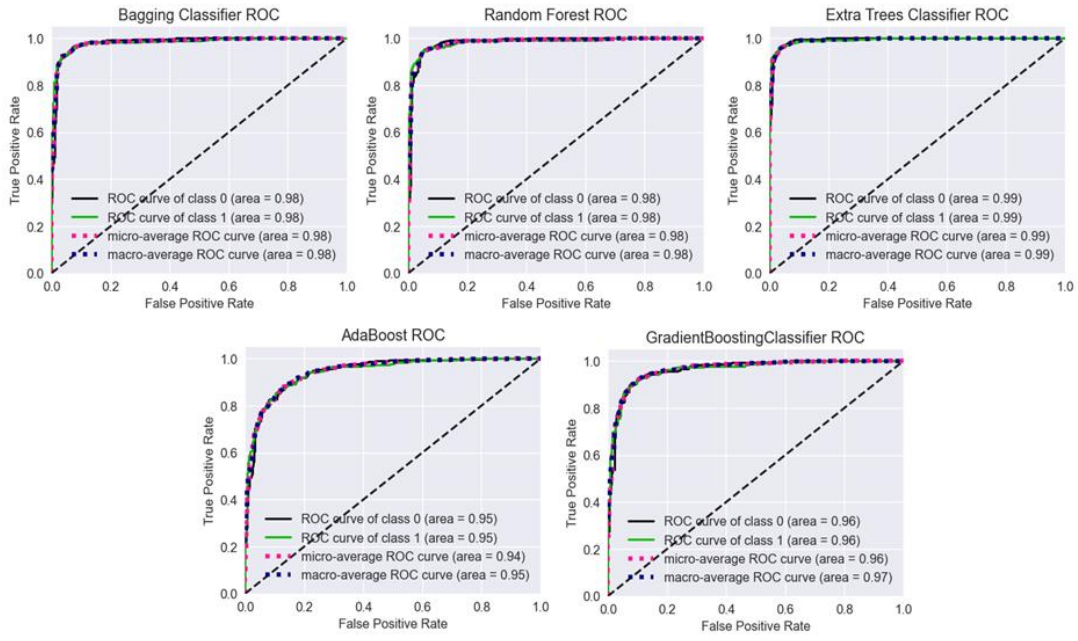
Table 18. Performance of the Ensemble classifiers for Gastric Cancer Big Data

Feature Subset	Models	Accuracy %	Precision %	Recall %	F1_score %	ROC-AUC %	MCC	Brier Score
Selected 15 features from undersampled (**Balanced Dataset**)	Gradient boost (Training)	88.43	88.43	88.56	88.49	96	0.77	0.12
	Gradient boost (**Testing**)	90.76	90.30	91.13	90.72	96	0.82	0.09
	AdaBoost (Training)	86.34	87.58	84.84	86.19	94	0.73	0.14
	AdaBoost (**Testing**)	87.27	88.33	85.63	86.96	95	0.75	0.13
	Random Forest (Training)	93.21	92.17	94.50	93.32	98	0.86	0.07
	Random Forest (**Testing**)	94.39	93.67	95.11	94.39	98	0.89	0.06
	Bagging (Training)	92.91	92.01	94.06	93.02	98	0.86	0.07
	Bagging (**Testing**)	93.94	93.62	94.19	93.90	98	0.88	0.06
	Extra Trees (Training)	95.52	95.01	96.14	95.57	99	0.91	0.04
	Extra Trees (Testing)	**96.06**	**95.74**	**96.33**	**96.04**	**99**	**0.92**	**0.04**

Feature Subset	Models	Accuracy %	Precision %	Recall %	F1_score %	PR-AUC of positive class %	MCC	Brier Score
Selected 15 featuresfrom **Imbalanced Dataset**	Gradient boost (Training)	91.82	82.52	66.47	73.63	83.9	0.69	0.08
	Gradient boost (**Testing**)	93.07	86.12	71.18	77.94	89.3	0.74	0.07
	AdaBoost (Training)	90.82	81.41	60.38	69.33	80.0	0.65	0.09
	AdaBoost (**Testing**)	92.00	83.46	66.76	74.18	86.1	0.70	0.08
	Random Forest (Training)	94.99	91.50	78.08	84.26	91.9	0.82	0.05
	Random Forest (**Testing**)	96.31	93.20	84.71	88.75	95.5	0.87	0.04
	Bagging (Training)	95.36	91.43	80.55	85.65	92.6	0.83	0.05
	Bagging (**Testing**)	97.06	93.52	89.12	91.27	96.1	0.90	0.03
	Extra Trees (Training)	96.63	93.97	85.92	89.76	96.0	0.88	0.03
	Extra Trees (**Testing**)	97.57	95.62	90.00	92.73	98.0	0.91	0.02

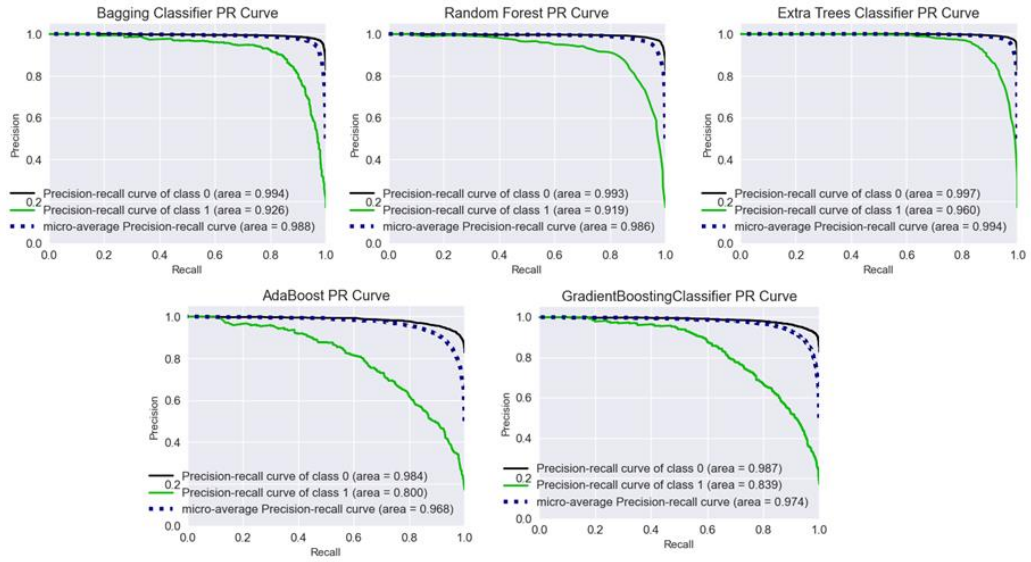


Training set - Balanced Dataset

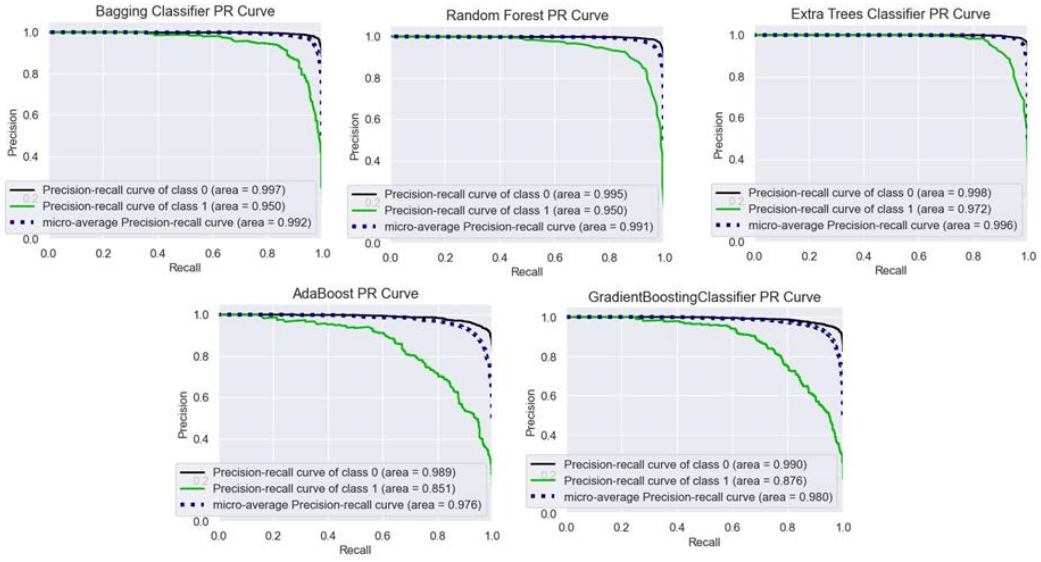


Test set - Balanced Dataset

Fig 37 Receiver Operating Characteristic curves for five ensemble models on balanced dataset



Training set - Imbalanced Dataset



Test set - Imbalanced Dataset

Fig 38 Receiver Operating Characteristic curves for five ensemble models on imbalanced dataset

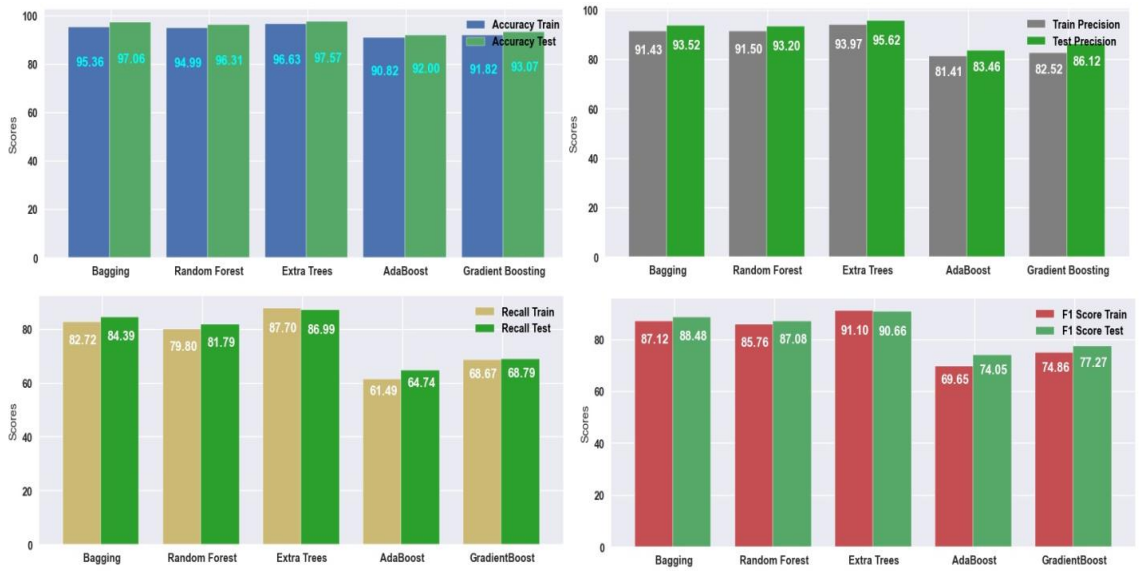


Fig 39. Performances of five ensemble models on imbalanced dataset

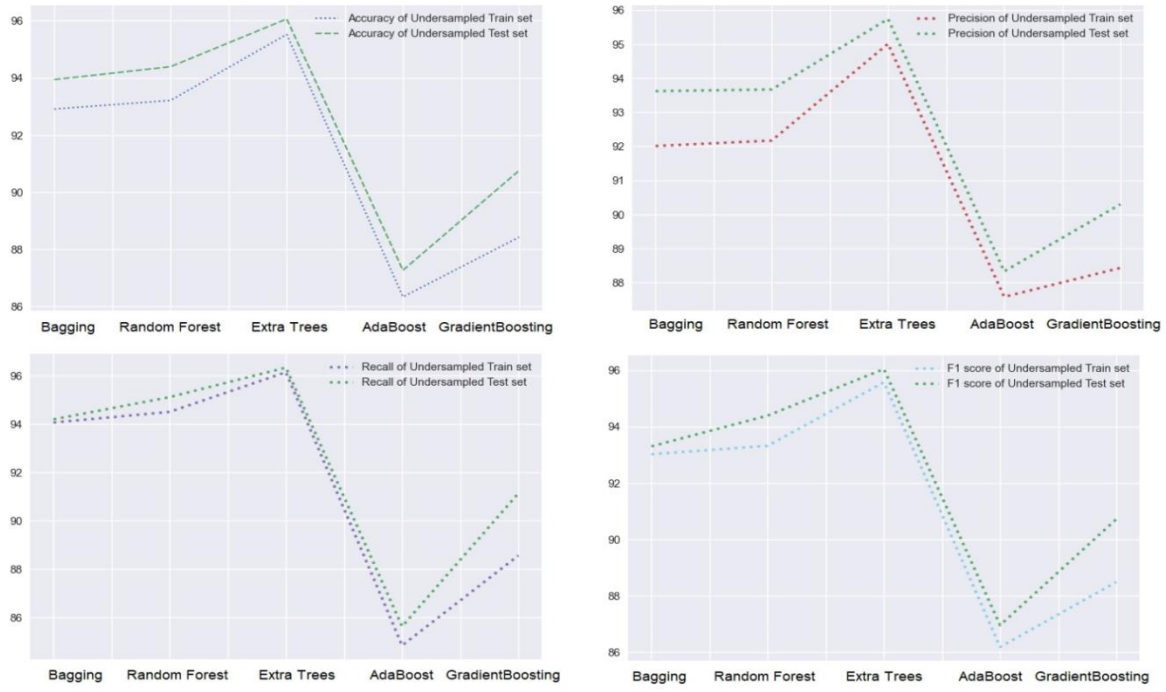


Fig 40. Performances of five ensemble models on balanced dataset

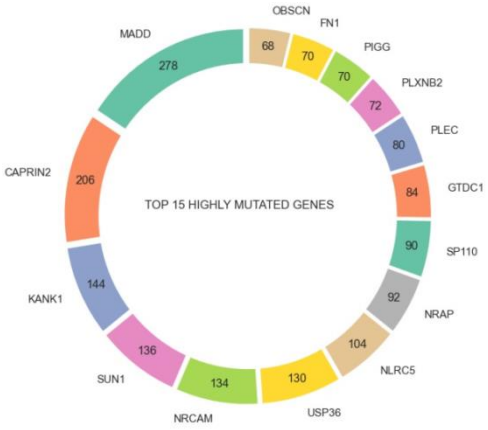


Fig 41. Top 15 Highly Mutated Genes in Gastric Cancer Big Data

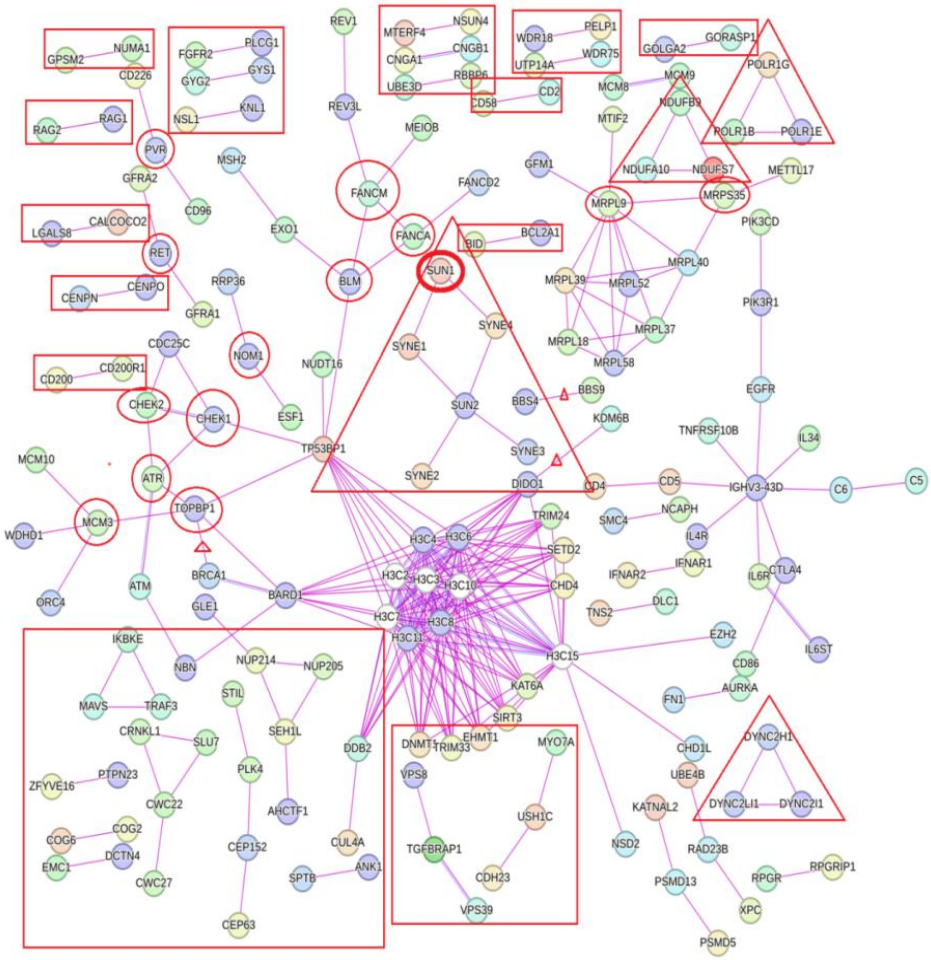


Fig 42.Genes and their functional protein Association in Gastric Cancer Big Data by Experimental Evidence

4.3 MSI Gastric Cancer Classification using Genomic and Epidemiological Features

4.3.1 Learning Curves

All eleven features were incorporated to develop learning curves for the machine learning models: support vector machine, random forest, Naïve Bayes, logistic regression,and Multilayer Perceptron. The dataset comprised only 95 data points and 11 attributes for training. To address potential underfitting, learning curves were generated for the machine learning models (Fig 6). The learning curves showed exclusion of SVM due to huge error gap between the training and validation curves.

4.3.2Feature Selection

To identify the significant features contributing to the MSI-GC, three methods were employed: Ridge regression, extra trees, and recursive elimination methods.The features selected by ridge regression were extra salt, khuva, alcohol, smokeless tobacco product (Khaini / Sadha), and smoked food (Fig 7).The extra trees method selected features including Saum, alcohol, khuva, smoked food, and extra salt (Fig 8). Extra salt, smokeless tobacco products (Khaini / Sadha and Tuibur), alcohol, smoking, and khuva were key features extracted by the recursive elimination method (Table 6).

4.3.3 Machine Learning models for MSI-GC

Random forest and multilayer perceptron classifier demonstrated efficacy in classifying positive and negative MSI-GC. These two models were developed on the three subsets identified by ridge regression, recursive elimination method, and extra trees classifier. The models underwent hyperparameter tuning to optimize performance (Fig 10 and 11). The feature subset extracted by extra trees classifier yielded moderate performance with random forest and multilayer perceptron models, achieving an accuracy of 85%, precision of 81%, ROC curve of 85%, F1-score of 86%, Brier score of 0.14, and Matthews correlation coefficient (MCC) of 0.70.

Superior performance was observed in the aforementioned two models using the feature subset extracted via recursive elimination method, with an accuracy of 94%, precision of 96%, recall of 92%, F1-score of 94%, Brier score of 0.06, and MCC of 0.87. Random forest and multilayer perceptron models exhibited efficient and consistent results with the subset identified by ridge regression, achieving 96% in accuracy, recall, precision, and F1-score, as well as the highest MCC of 0.91 (Table 19). The ROC curves were 96%, 94%, and 85% for subsets of ridge regression, recursive elimination method, and extra trees, respectively (Fig 43).

4.3.4 Identification of Confounding and Driving Factors

Multiple-linear regression was employed on the covariates identified by ridge regression to determine the driving and confounding factors for MSI-GC. Among all the features, khuva exhibited the lowest error and a statistically significant p-value of 0.13 and 0.0007, respectively (Table 20). The regression coefficient of -0.20 indicated that MSI-GC (dependent variable) was dependent on khuva (independent variable), which was evidently a driving factor in the etiology of the disease. Excessive salt consumption, smoked food intake, Khaini / Sadha use, and alcohol consumption were also identified as driving factors. The distribution of khuva consumption data demonstrated a notable distinction between the positive and negative classes of MSI-GC (Fig 44).

Table 19. Performance of MSI-GCClassifiers on three Feature Subsets

Feature Sets	Classifiers	Accuracy %	Precision %	Recall %	ROC curve %	F1 score %	Brier score	MCC
Ridge Regression Extra salt Smoked food Khaini /Sadha Alcohol Khuva	Random Forest	96	96	96	96	96	0.04	+0.91
	Multilayer Perceptron	96	96	96	96	96	0.04	+0.91
Recursive elimination method	Random Forest	94	96	92	94	94	0.06	+0.87
	Multilayer	94	96	92	74	94	0.06	+0.87

Extra salt Khaini /Sadha Tuibur Alcohol Smoking	Perceptron							
<u>Extra Trees Classifier</u> Extra salt Saum Smoked food Alcohol Khuva	Random Forest	85	81	92	85	86	0.14	+0.70
	Multilayer Perceptron	85	81	91	85	86	0.14	+0.70

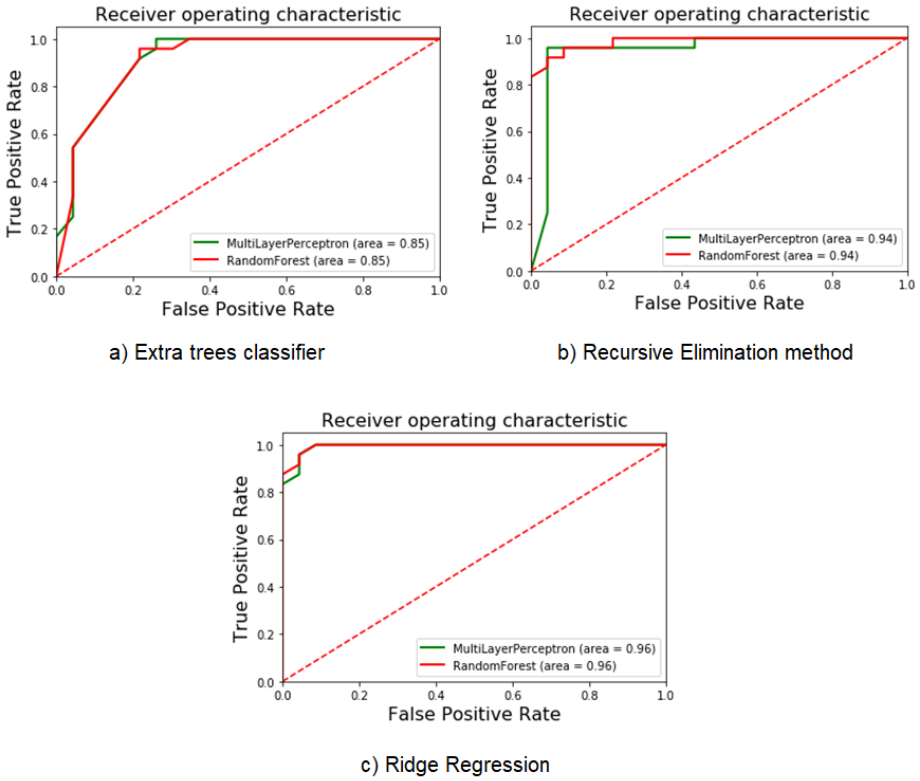


Fig 43. ROC of Random Forest and Multilayer Perceptron models based on theirFeature Sets

Table 20. Identification of Driving Factor and Confounding Factor for MSI-GC

Estimation of coefficients of linear model fitted to data in Ridge Regression Method

Independent Variable	Slope	p-value	Regression Coefficient	Mean Square Error
Extra Salt	-0.0349	0.573	-0.032	0.15
Smoked Food	0.1014	0.127	0.096	0.16
Khaini / Sadha	0.0308	0.608	-0.007	0.15
Alcohol	-0.0633	0.303	-0.066	0.15
Khuva	-0.2142	0.0007	-0.200	0.13

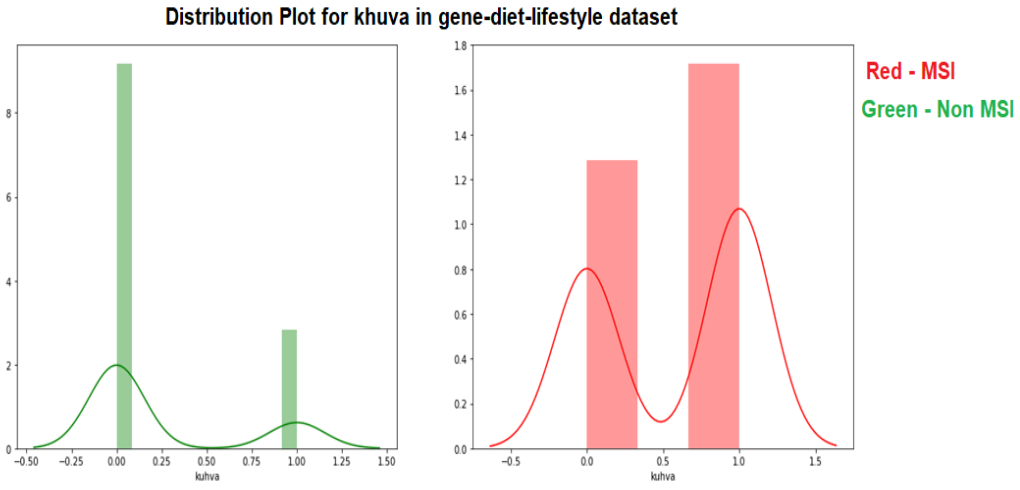


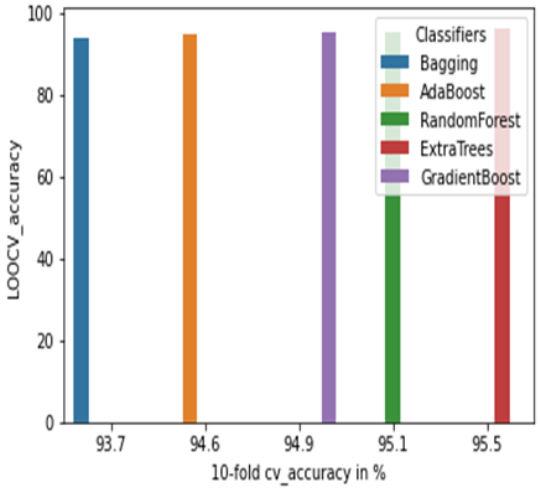
Fig 44. Distribution of MSI and non-MSI by the driving factor

4.4 Breast Cancer Classification from Epidemiological Data

4.4.1 Ensemble Models

Five ensemble models were developed, and their performances were evaluated using leave-one-out cross-validation (LOOCV) and 10-fold cross-validation. The classification accuracies demonstrated consistent results across both validation methods. The Extra Trees classifier achieved an accuracy of 96% with both validation techniques. Gradient boosting, random forest, and AdaBoost classifiers

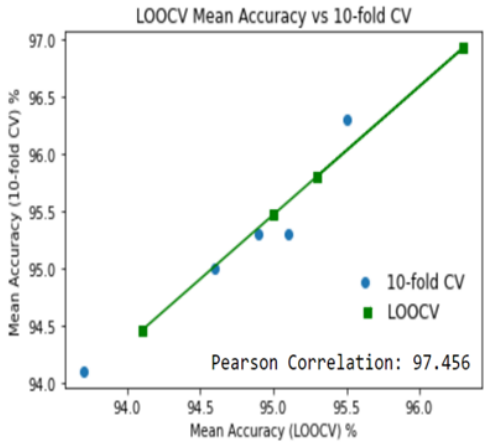
yielded an accuracy of 95% using LOOCV and 10-fold CV. The Bagging classifier attained an accuracy of 94% (Fig 45a). The Pearson correlation coefficient between the accuracies obtained by LOOCV and 10-fold CV was 0.97 (Fig 45a, b, c).



a) Accuracies of 10-fold CV

```
>BaggingClassifier: LOOCV=0.941, cv=0.937
>AdaBoostClassifier: LOOCV=0.950, cv=0.946
>RandomForestClassifier: LOOCV=0.953, cv=0.951
>ExtraTreesClassifier: LOOCV=0.963, cv=0.955
>GradientBoostingClassifier: LOOCV=0.953, cv=0.949
```

b) Accuracies of LOOCV



c) Pearson Correlation between 10-fold CV and LOOCV

Fig 45 Accuracies of 10-fold CV (a), Accuracies of LOOCV (b), Pearson Correlation between 10-fold CV and LOOCV (c)

4.5 Breast Cancer Classification from Germline Data

4.5.1 Convolution Neural Network Models

The performance of CNN models on the four datasets demonstrated superior results in terms of accuracy, precision, recall, f1-score, and ROC curves. The highest accuracy achieved was 98.98% (original dataset) and 98.68% (SMOTEENN dataset), as illustrated in Fig 46. The highest precision was 98.79 (original dataset) and 99.32 (SMOTEENN dataset). The recall and f1-score values were highly satisfactory using both the original and SMOTEENN datasets. Although accuracy, precision, recall, and f1-scores were notably promising using the raw dataset, the ROC curve performance exhibited a slight decrease when compared to the balanced dataset – SMOTEENN (Fig 47).Based on the model evaluation parameters, it was noteworthy that the feature subset identified by SMOTEENN could be the critical features for identifying pathogenic variants from the breast cancer germline dataset (Fig 20).

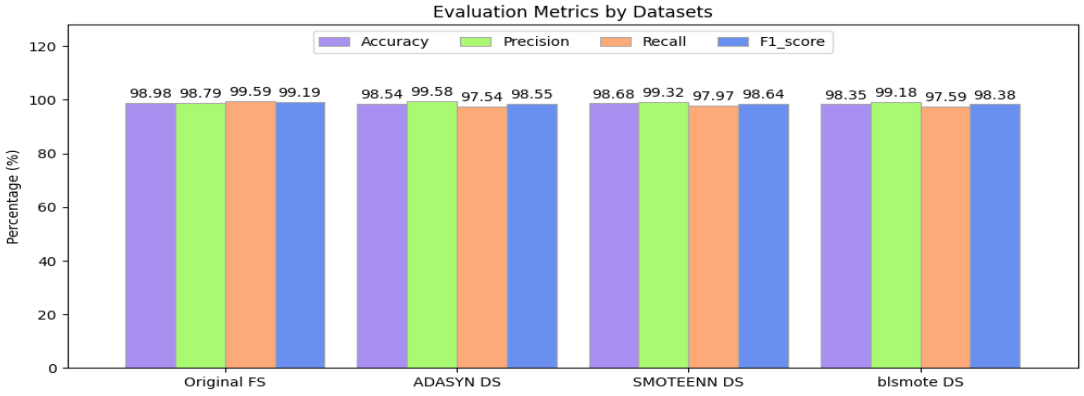


Fig 46. Comparative Performance on Evaluation Metrics based on Imbalanced and Balanced breast cancer datasets

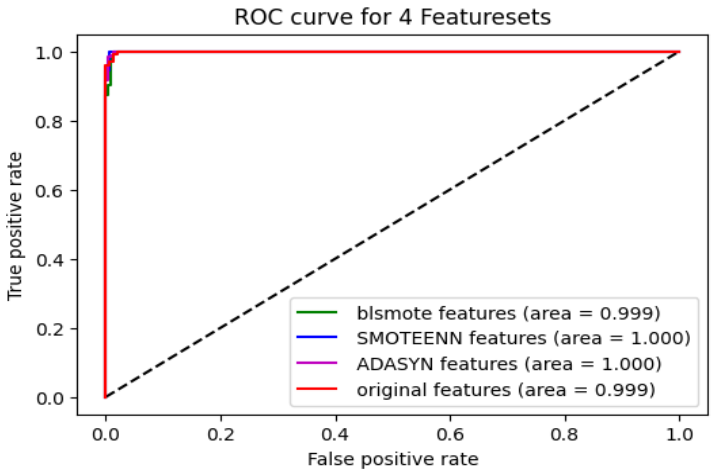


Fig 47.Comparison of Receiver Operating Characteristic curves by imbalanced and balanced breast cancer datasets

4.5.2 Highly Mutated Genes and Genes Functional Connectivity in Breast Cancer

The top 10 mutated genes in the breast cancer genetic dataset were MADD (18%), PLEC (11%), NEB (8%), GTDC1 (8%), TRMP3 (6%), PCM1 (6%), RIMS2 (6%), SETX (5%), DPP10 (5%), and POMT1 (5%)(Fig 48). The MADD gene is associated with cell survival. PLEC is correlated with poor survival, exhibits overexpression in cancer, and interacts with the BRCA2 gene. RIMS2 is implicated in cancer metastasis from breast to brain. GTDC1 is a key contributor to tumor development in breast cancer. DPP10 demonstrates elevated expression in cancer tissues.The graphical representation had shown the interaction of all genes collected from the present breast cancer germline data (Fig 49). There was evidence of hotspots of breast cancer genes which were exhibiting independent interaction between nodes (genes). These were experimentally proven functional connectives explored by string database.

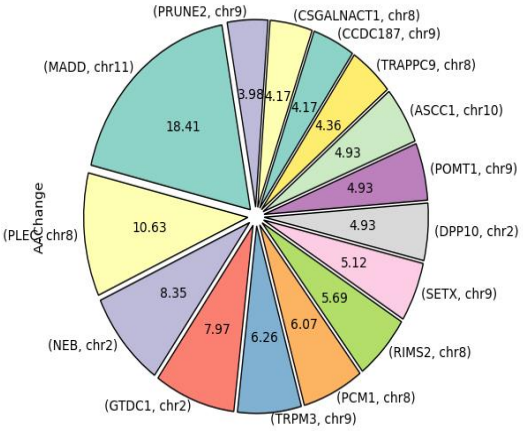


Fig 48.Mutations by top 15 genes and Chromosome wise from Breast Cancer Dataset

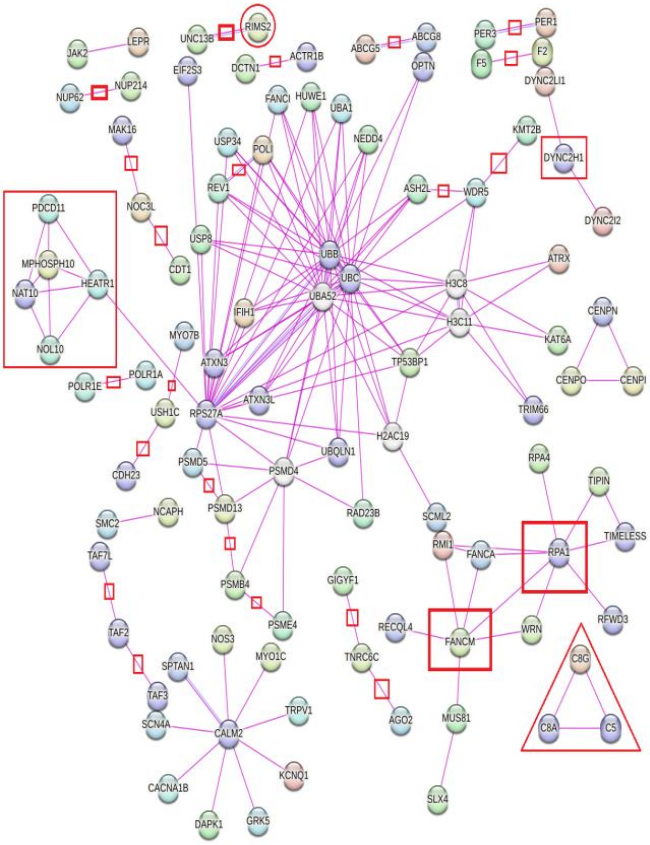


Fig 49.Genes and their functional protein Association in Breast Cancer Big Data by Experimental Evidence

4.6 Head and Neck Cancer Classification from Epidemiological Data

4.6.1 Data Models

Five machine learning classifiers were developed, and their performances were evaluated using accuracy, precision, recall and f1_score. The classification accuracies and f1_scores showed a consistent classifier. The Extra Trees classifier achieved an accuracy of 85%, precision 93%, recall 67%, and f1_score 78%, this was followed by random forest with f1_score of 73% and accuracy of 82%.Linear SVM, decision tree, and Naïve Bayes had metrics of accuracy, and precision above 70% (Fig 50).

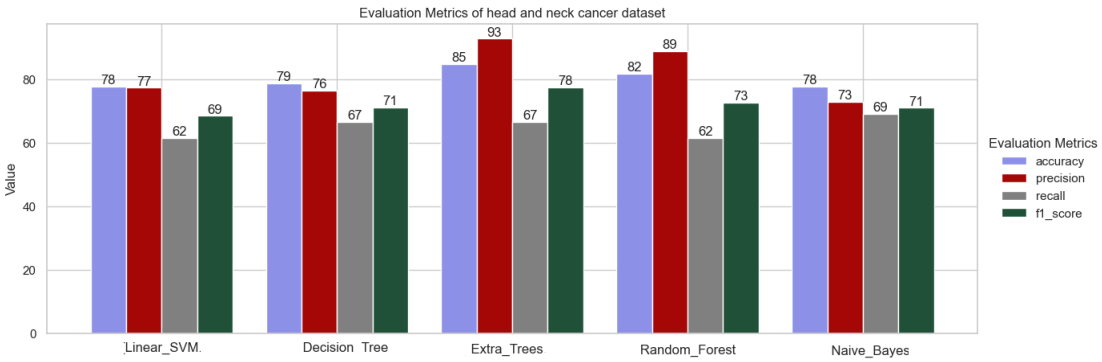


Fig 50.Comparative Performances of five classifiers on Head & Neck cancer Epidemiological dataset

4.7 Head and Neck Cancer Classification from Germline Data

4.7.1 Data models

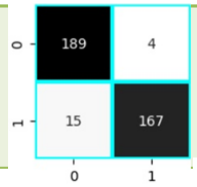
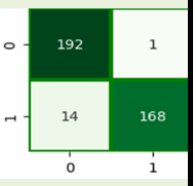
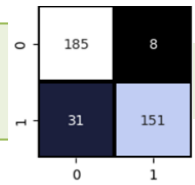
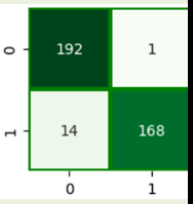
Heat map had eliminated all the allele frequencies from the different population as they had multicollinearity between them (Fig 30). The results obtained by random forest, decision tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and logistic regression were precisely optimized by setting appropriate hyperparameters (Table 12-15 and Fig 31). The accuracy of decision tree and SVM exceeded that of random forest, KNN, and logistic regression, reaching 96% (Table 20). These two algorithms demonstrated very low false positive rates of

1. The results of decision tree and SVM exhibited strong support for the classification results, with each metric of these algorithms being identical. False negatives were considerably higher in logistic regression at 20, although it maintained a low false positive rate of 1. Random forest achieved an accuracy of 95% but exhibited a slightly increased false positive rate of 4 and false negative rate of 15 compared to decision tree and SVM. The accuracy of KNN was 90%, whereas it demonstrated high false positive and false negative rates at 8 and 31, respectively (Table 21). Diagnostic performance of binary classification was measured by Receiver Operating Characteristic (ROC) curves. Among the five classifiers, random forest exhibited the highest Area Under the Curve (AUC) of 99%; however, the model displayed marginally high false positive instances (Table 21 and Fig 51). Consequently, the ROC-AUC of decision tree and SVM were 98%, which demonstrated excellent overall performance in classifying the healthy and head and neck cancer variants from the annotated genomic dataset (Fig 51).

4.7.2 Highly Mutated Genes and Genes Functional Connectivity in Head and Neck Cancer

The top 10 mutated genes in the head and neck cancer genetic dataset were, NEB (15%), CENPF (11%), SYNE2 (10%), FAT1 (10%), OBSCN (9%), TG (9%), MKI670 (9%), SYNE1 (8%), PRUNE2 (8%) AND SETX (8%) (Fig 52).Altered levels in NEB of cancer tissues were seen as compared to normal tissue. Overexpression of CENPF has been observed in oral squamous cell carcinoma. Mutations in the SYNE2 gene had exhibited a significant association with poor prognosis and enhanced metastatic potential in oral squamous cell carcinoma. Mutations in the FAT1 gene had enhanced tumor growth and metastasis in head and neck squamous cell carcinoma. The graphical representation had shown the interaction of all genes collected from the present head and neck germline data (Fig 53). There was evidence of hotspots of head and neck cancer genes which were exhibiting independent interaction between nodes (genes). These were experimentally proven functional connectives explored by string database.

Table 21.Comparative Performances of five classifiers on Head & Neck cancer

Model	Accuracy (%)	Confusion Matrix
Random Forest	95	
Decision Tree	96	
KNN	90	
SVM	96	
Logistic Regression	94	

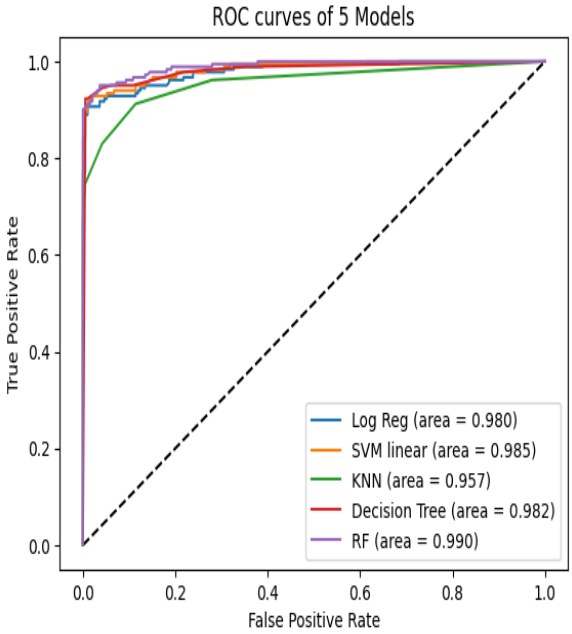


Fig 51. Comparison of Receiver Operating Characteristic curves by head and neck cancer datasets

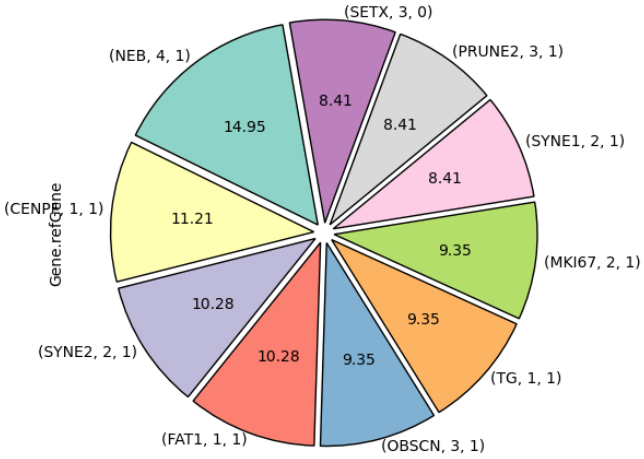


Fig 52.Mutations by top 10 genes and Chromosome wise from Head and Neck Cancer Dataset

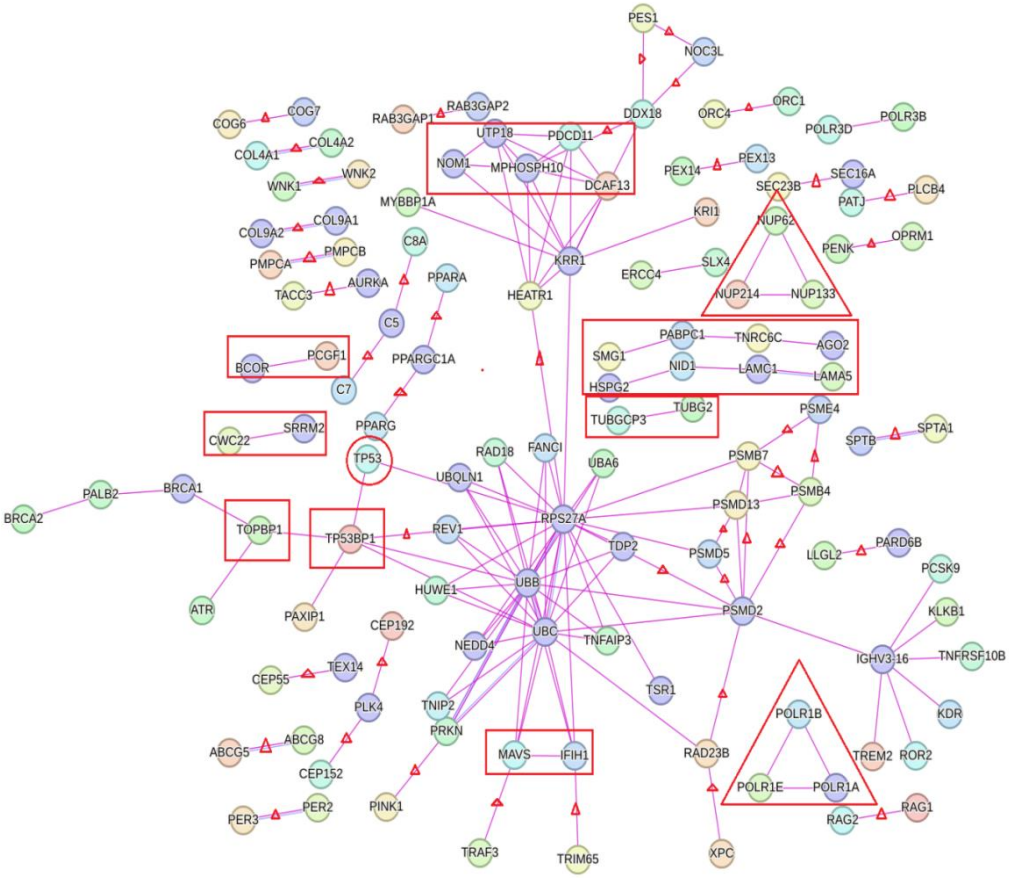


Fig 53.Genes and their functional protein Association in Head and Neck Cancer Big Data by Experimental Evidence

4.8 Integration of Healthy, Gastric, Breast and Head & Neck Cancer Classification

4.8.1 Proposed Multilayer Perceptron Model

The accuracy of the multilayer perceptron model attained 75% in this multi-class classification. The model was constructed with two hidden layers comprising 16 and 8 neurons each, which achieved satisfactory accuracy while utilizing reduced computational resources and time. Other architectures of multilayer perceptron models demonstrated comparable performance; however, in terms of time and computational efficiency, the architecture with 16 and 8 neuron hidden layers yielded promising results (Table 22).

Table 22.MLP model performance under different Architectures

S.No	MLP Hidden Layer Architectures	Train Accuracy %	Test Accuracy %
1.	[24,16,8]	71	75
2.	[16,8,6]	74	75
3.	[16,8,4]	75	75
4.	**[16,8]**	**73**	**75**

```
MULTILAYER PERCEPTRON with hidden layers : 16,8
Test Accuracy: 0.7537878787878788
Train Accuracy: 0.7388059701492538
Confusion Matrix:
 [[46  7  6  7]
 [ 6 61  1  2]
 [10  1 51  7]
 [13  4  1 41]]
```

Fig 54. Confusion Matrix of Proposed Multilayer Perceptron Model

Furthermore, this architecture exhibited the lowest false positive and false negative rates when compared to alternative architectures (Fig 54). Training and testing

accuracies of all four architectures were shown in table 22 and there was no evidence of overfitting in the dataset. It is noteworthy that a multilayer perceptron model of low complexity was employed.

4.8.2 Performance of Ensemble Data Models on Big Data

There were 35.7% of unique variations in the big dataset of gastric, breast, and head and neck cancers. Approximately, 56.9% of variations were found to be common among all three tumors, while only 7.4% were found to be prevalent in any of the two cancer types (Fig 55). Genes USH2A, MK167, NEB, and OBSCN were found common and highly mutated in the above-mentioned cancers (Fig 56). DHAH5, PARP4 and PRUNE2 were in top 10 mutated genes speficically in gastric, breast and head & neck cancer, respectively. FAT1, SYNE1 and SYNE2 were found common between gastric and head & neck cancers; while, none found common between gastric and breast; breast and head & neck cancers (Fig 56). An interesting result was obtained on integerting gastric, breast and head & neck germline characteristic which had revealed more unique genes among these three cancers. MDN1 in GC; XIRP2 and FBN3 in BC; were predominantly mutated in this dataset (Fig 57).

AdaBoost algorithms achieved the maximum accuracy of 84.18%, precision of 82.15%, recall of 86.83%, f1_score of 84.43%, and ROC_AUC of 84.21% when classifying pathogenicity within the three cancer kinds (gastric, breast, and head and neck) (Table 23). When compared to other classifiers, the metrics obtained by AdaBoost revealed a well-balanced classifier that identified benign and pathogenic variations while maintaining a balance of precision and recall.

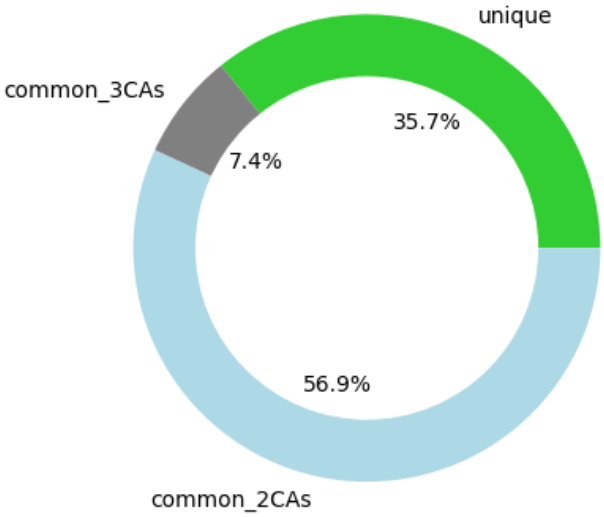


Fig 55. Variants in Big Dataset of Gastric, Breast and Head and Neck Cancers.

Fig 56. Top 10 Mutated genes in each Gastric, Breast and Head & Neck Cancers.

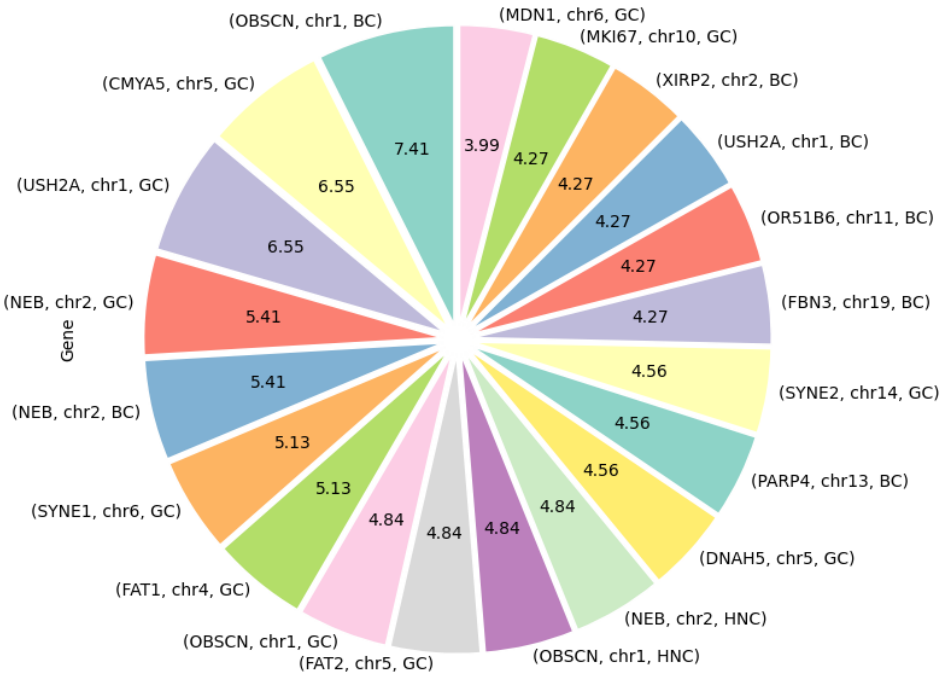


Fig 57. Top 20 Mutated genes by integrating (Gastric, Breast and Head & Neck Cancers) germline data

Table 23. Performance of Ensemble Classifiers on Big Dataset of Gastric, Breast and Head and Neck Cancers.

Model	Accuracy %	Precision %	Recall %	F1_Score %	ROC_AUC %
AdaBoost	84.18	82.15	86.83	84.43	84.21
Extra Trees	72.13	64.53	96.75	77.43	72.42
Random Forest	71.73	63.94	98.05	77.41	72.04
Bagging classifier	70.76	63.25	97.40	76.70	71.08

Chapter 5. DISCUSSION

5.1 Classification of Gastric Cancer using Epidemiological Features

Logistic regression is a fundamental and widely used statistical method in machine learning on medical data (Schober and Vetter, 2021). One of the key strengths of logistic regression is its interpretability and its simplicity also makes it less prone to overfitting compared to more complex models, especially when dealing with limited data. Using the present dataset a custom-made logistic regression was developed for gastric cancer classification from epidemiological features which had optimal performance. This model effectively correlated the achieved results and theoretical knowledge through smooth convergence.

This section evaluated the logistic regression model's effectiveness, considering all risk factors as independent variables. With a 0.5 threshold, the model achieved 98.51% accuracy. Figure 35 illustrates the rapid error rate reduction by the cost function and gradient descent. Global minima were reached after 150,000 iterations, with a 0.000915 learning rate and 0.22 line intercept. When iterations were reduced to 50,000, maintaining the same learning rate but with a -0.04 intercept, accuracy decreased to 97.01%, as shown in Figure 35. This figure also demonstrates that while the error rate declined rapidly after 50,000 iterations due to the gradient descent optimizer, accuracy was compromised. Scaling features between 0 and 1 enhanced the model's accuracy and error performance.

Previous studies had demonstrated logistic regression's effectiveness in various medical applications. It achieved 95.7% accuracy in classifying breast cancer tumors (Khairunnahar et al. 2019), and 97.18% accuracy in a breast cancer dataset comparison study (Sultana et al. 2018). Zhu et al.'s integrated model attained 97.04% accuracy for diabetes prediction (Zhu et al. 2019), while other study reported 90% accuracy in predicting comorbidities in children with cerebral palsy (Bertoncelli et al. 2020). These results indicate that logistic regression models perform comparably to ensemble machine learning techniques in healthcare contexts.

The current study's logistic regression model revealed strong associations between all features and gastric cancer incidence in Mizo populations. Many of these features align with globally recognized risk factors for gastric cancer. Excessive salt intake, processed meats, alcohol consumption, and limited physical activity are among the risk factors for gastric cancer in Asian populations (Karagianni et al. 2010). However, conflicting findings exist, such as a Chinese study reporting lower gastric cancer risk among those preferring non-salty food and non-drinkers (Zhong et al. 2012).

In the Mizo population, several factors were identified as risks for gastric cancer, including the consumption of smoked foods, use of tobacco products, early-age smoking, limited fruit intake, secondhand smoke exposure, and a family history of cancer (Hui et al. 2023; Zhang et al. 2021). The present study revealed a strong negative correlation between Khuva (a combination of betel leaf and betel nut) and MSI-GC. Additionally, electromagnetic radiation was identified as a potential risk factor for gastric cancer in this population (Siddoo-Atwal. 2018). The current research achieved 98% accuracy utilizing Logistic Regression, demonstrating its efficacy in identifying risk factors and facilitating decision-making processes.

5.2 Classification of Gastric Cancer using Genomic Features

This investigation focuses on the development of ensemble classifiers to differentiate between benign and pathogenic variants in gastric cancer whole exome big data. The initial variants were annotated utilizing functional prediction tools, conservation scores, and allele frequency databases, including gnomAD and 1000 genome projects (Auton et al. 2015). These annotated features were employed to construct ensemble models. The models exhibited enhanced performance with high data fidelity, achieved by labeling target variants only when all six prediction tools (MutationTaster, FATHMM_pred, LR_pred, RadialSVM, SIFT_pred, and Polyphen2_HDIV) concurred on their benign/pathogenic status. The dataset was refined from 192781 to 72471 variants after eliminating redundancies. The ensemble models were trained on 70% of the non-redundant variants and tested on the remaining 30%.

The models were trained using optimal hyperparameters obtained through 5-fold gridsearchCV and validated using PR-AUC for imbalanced datasets and ROC-AUC for balanced datasets. For the imbalanced dataset, three key validation metrics - PR-AUC, recall, and MCC - were considered as gold standards for performance evaluation (Hancock et al. 2023). The optimal classifiers were selected based on minimum brier scores and maximum recall, precision, and PR-AUC, as false positives could lead to erroneous drug target identification (Kim, et al. 2020). To address multicollinearity and prevent model overfitting, non-collinear feature subsets were extracted for both training and testing sets.

The study also examined the impact of LASSOCV-selected features on the ensemble models. LASSOCV regression's unique ability to tune λ during cross-validation and handle multicollinearity without data loss was emphasized. Despite the extensive feature set, not all allele frequencies, conservation scores, and functional prediction tool scores significantly contributed to model performance. The most influential features for determining variant pathogenicity were identified as CADD_phred, VEST3_score, phyloP46way_placental, AF_asj, AF_eas, Alt, Start, ExAC_SAS, ExAC_AMR, Ref, Genes, ExAC_OTH, CADD_raw, non_neuro_AF_popmax, and AAChanges. For the balanced dataset, four out of five models demonstrated the highest ROC-AUC, while for the imbalanced dataset, three out of five models exhibited the highest PR-AUC. Both scenarios yielded high MCC scores exceeding 0.80. To validate the consensus classification using the top 15 significant features, the study employed both imbalanced and balanced (undersampled) datasets. These findings suggest that the identified features play a crucial role in determining pathogenicity in gastric cancer big data (Noroozi et al. 2023).

CADD tools have consistently demonstrated efficacy in identifying potential breast cancer driver genes with p-values ≤ 0.01, followed by VEST3 prediction scores (Nono et al. 2019). This investigation also revealed that two whole-exome functional prediction scores (CADD raw and CADD phred) exhibited high and significant scores compared to 13 other selected features (Fig 5 and Table 4). A

random forest classifier developed by Jain for pre-mRNA splicing prediction achieved 100% specificity, 93.3% accuracy, and only 3 false positives with a 0.6 cutoff. The models' performance in that study improved significantly when phyloP46way_placental, phyloP46way_primate, CADD raw, and CADD phred scores were incorporated into the dataset (Jain 2014). Despite being conducted a decade ago, phyloP and CADD scores remain crucial factors in medical big data mining. In accordance with these features, the extra trees classifier presented here demonstrated relatively higher performance over time, achieving 96% accuracy, 95.74% precision, and 96.33% recall with a low error rate of 0.04 (brier score) on a balanced dataset.

This study also exhibited notable performance using extra trees classifiers on an imbalanced dataset, with a PR-AUC of 98% and the lowest error rate of 0.02 due to low false positives. A recent report indicates that in silico tools for predicting pathogenicity continue to evolve through regular bug fixes and improvements, suggesting that CADD and VEST3 scores can be considered key factors in determining pathogenicity and identifying new candidate genes/variants (Gracia et al. 2022). Allele frequencies from various populations (African/African American, East Asian, South Asian, and Exome Aggregation Consortium database) were among the top 10 selected features in this study. Previous research has demonstrated that allele frequencies are useful in detecting genetic misdiagnosis and identifying benign variants (Kobayashi at al. 2017).

An interesting pattern emerged, with genes, chromosomal positions, reference bases, and altered bases identified as significant features, their importance validated by machine learning model performances. Studies have shown that reference base alterations at specific positions cause gastric cancer in somatic mutations (Zhang et al. 2018). Ongoing research aims to identify potential driving factors from genes and their positional aspects concerning altered bases. This study can serve as a valuable tool for narrowing down searches to identify driver genes, chromosomal positions, reference bases, and altered bases. The present study found unified predictions visualized from various features, including CADD_phred, VEST3_score,

phyloP46way_placental, AF_asj, AF_eas, Alt, Start, ExAC_SAS, ExAC_AMR, Ref, Genes, ExAC_OTH, CADD_raw, non_neuro_AF_popmax, and AAChanges, by reproducing valid prediction results. The classifiers developed in this work primarily demonstrated excellent performance in distinguishing between benign and pathogenic variants from gastric cancer big data.

The 15 key features underwent training and testing on both imbalanced and balanced datasets, with the latter created through random undersampling. Additionally, these identified features significantly reduce computational time during variant annotation, potentially expediting diagnosis and treatment outcomes. These genetic characteristics can be combined with epidemiological factors to efficiently predict disease outcomes and prognosis rates. Extra trees classifiers demonstrated superior performance in terms of accuracy, precision, recall, and f1_score across both balanced and imbalanced datasets. Several crucial steps were implemented to prevent erroneous results and misinterpretation: (i) determining optimal tuning hyperparameters using 5-fold GridSearchCV; (ii) systematically reducing horizontal dimensionality through a valid feature selection method by fine-tuning LASSOCV; and (iii) labeling targets only when all 6 functional prediction tools agreed, ensuring a balance between precision and recall (minimizing false positives and negatives).

Through examination of all features, the top 15 best-fit features were identified, with their significance demonstrated through ensemble learning performances (Table 18, and Fig 37-40). As genes were among the 15 signatures obtained in this study, a pie chart was created to illustrate the 15 most frequently mutated genes, as depicted in Fig 41. The chart clearly indicated that MADD and CAPRIN2 genes were mutated in nearly 200 locations. Notably, CAPRIN2 has been identified as a poor molecular prognosis marker for nasopharyngeal cancer (Qiu, et al. 2022). MADD has been studied as an essential factor for cancer cell survival (Jayarama et al. 2014). These two genes, identified through machine learning, correspond well with existing known factors. An independent tri-nodular genes interaction was observed in this study, MYO7A, USH1C and CDH23. These were study in other African and Asian populations (Nakanishi et al. 2010).

5.3 Classification of MSI Gastric Cancer using Genomic and Epidemiological Features

To assess model fit, learning curves were generated for five supervised ML algorithms (support vector machine, random forest, logistic regression, multilayer perceptron, and Naive Bayes) utilizing all eleven attributes. The curves indicated that all algorithms, with the exception of support vector machine, were well-suited for classifying the dataset (Fig 6). The support vector machine exhibited a wider disparity between training and validation scores, while the other four algorithms demonstrated smooth convergence with reduced error gaps between training and validation score curves (Fig 6). To eliminate irrelevant attributes, identify strongly correlated features, and mitigate overfitting, feature selection was conducted using 3 robust methods: recursive elimination, ridge regression, and extra trees classifier. Ridge regression identified Khaini/sadha, khuva,extra salt, smoked food, alcohol as the top 5 potential factors contributing to MSI-GC (Fig 7). The recursive elimination technique identified tuibur,extra salt, alcohol,Khaini/sadha, smoking as the 5 most significant features (Table 6). The extra trees model determined extra salt, saum, smoked food, alcohol, and khuva as the top five correlated features (Fig 8). Data models were constructed using the four selected algorithms (random forest, multilayer perceptron, logistic regression, and Naive Bayes) on the three groups of feature sets identified by the extraction methods. The feature selection process and learning curve analysis confirmed the absence of underfitting and overfitting in the dataset. These crucial data preprocessing steps were implemented to enhance accuracy and minimize misclassification of datapoints prior to model development.

Features selected through ridge regression, including khuva,extra salt, smoked food, alcohol, Khaini/sadha were utilized to construct Logistic Regression, Random Forest, Multilayer Perceptron, and Naive Bayes classifiers. The Random Forest and Multilayer Perceptron models were fine-tuned using hyperparameters (Fig 10 and 11). Hyperparameter optimization was a crucial step in model training, significantly enhancing performance during the testing phase (DeCastro-García et al. 2019). For the Random Forest model, 100 decision trees were selected (Fig 10),

based on research by Oshiro et al. which demonstrated that doubling the number of trees beyond 64-128 did not yield additional information gain, with optimal performance (100% AUC) achieved within this range (Oshiro et al. 2012).

The multilayer perceptron model underwent similar fine-tuning using critical hyperparameters. The neural network comprised three layers with 25 hidden neurons each. To address the vanishing gradient issue, the relu activation function was employed. A momentum of 0.9 was established to achieve global minima. Given the small dataset size, the 'lbfgs' (Limited memory Broyden Fletcher Goldfarb Shanno) optimizer was selected (Najafabadi, et al. 2017). The model was successfully optimized with 100 iterations (Fig 11), and deviating from these parameters resulted in unsatisfactory outcomes. Both Random Forest and multilayer perceptron models demonstrated exceptional performance, achieving 96% for accuracy, precision, recall, F1 measure, and ROC. These results clearly indicated that both algorithms were well-balanced classifiers, with +0.91 MCC and 0.04 brier score (Table 19 and Fig 43). The +0.91MCCsignified satisfactory classifier performance. The brier score of 0.04, falling below 0.175, indicated a very precise model in terms of both performance and accuracy (Xiong et al. 2023).

The feature set identified through recursive elimination encompassed tuibur, extra salt, alcohol, Khaini/sadha, and smoking. These features were utilized to construct data models employing random forest, Naive Bayes, logistic regression, and multilayer perceptron algorithms. During the training phase, random forest and multilayer perceptron classifiers underwent optimization using hyperparameter settings as illustrated in Figs. 10 and 11, respectively. Both random forest and multilayer perceptron models yielded identical results: 94% accuracy, 96% precision, 92% recall, 94% ROC, 94% F1 score, 0.06 brier score, and +0.87 MCC (Table 19 and Fig 43). The random forest classifier had been adjusted for gini index and entropy to enhance information gain on both methods producing equivalent evaluation metrics. Although, MCC and brier score fell below the accepted ranges (0.175, +1), but other metrics were not satisfactory, these results were insufficient to

conclusively demonstrate that the recursive elimination method feature set causes MSI-GC (Table 19).

The model development utilized a feature subset derived from the extra trees classifier, which included khuva, extra salt, smoked food,saum, alcohol. The resulting models achieved performance metrics of 85% accuracy, 81% precision, 91% recall, 85% ROC, 86% F1 score, 0.14 brier score, and +0.70 MCC (Table 18 and Fig 43). Notably, the features of importance from the extra trees classifier did not exhibit significant correlation with MSI-GC, as evidenced by below-average evaluation metric scores when compared to the other two feature sets obtained through ridge regression and recursive elimination methods.

A multiple linear regression analysis incorporating various covariates was employed to determine confounding factors (McNamee, 2005). The analysis revealed that excessive salt consumption, intake of smoked meat and vegs, habit ofchewing khaini/sadha, and drinking alcohol (independent variables) were associated with MSI-GC (dependent variable), indicating as possible confounding factors. Khuva emerged as the primary contributor to MSI-GC, demonstrating a statistically significant p-value of 0.0007 and a slope of -0.2142, which indicates a strong negative correlation between Khuva and MSI-GC compared to other factors (Table 20). The regression coefficient of -0.20 indicated that MSI-GC was highly dependent on Khuva, and the mean square error of 0.13 was the lowest among all attributes, suggesting minimal discrepancy between actual and predicted values (Table 20 and Fig 44). Previous studies have demonstrated that betel quid (Khuva) chewers are susceptible to MSI in head and neck cancer patients (Lin et al. 2016), and it is also a major factor in causing multiple MSI regions in oral cancer (Su et al. 2019). This study provides the first evidence that smoking tobacco, excessive salt intake, smoked food consumption, and alcohol use are confounding factors, while Khuva is the primary driver of MSI in gastric cancer. Backtracking of Khuva data distribution showed distinct demarcation between MSI positive and negative cases (Fig 44).

5.4 Classification of Breast Cancer using Epidemiological Features

This research employed 5 ensemble classifiers to forecast breast cancer in women from the Mizo population using epidemiological characteristics. Individuals with detrimental habits, such as addiction to various forms of tobacco, smoked foods, and alcohol, are more susceptible to developing different types of cancer. These models were constructed using non-invasive attributes likely to be risk factors for breast cancer. Model performance was assessed using leave-one-out cross-validation (LOOCV) and 10-fold cross-validation. Among the obtained accuracies, the extra trees algorithm surpassed AdaBoost, random forest, and gradient boosting models, achieving meanLOOCV accuracies of 96.3% and 10-fold CV of 95.5% (Fig 45a, b). Due to the imperfect predictive accuracy of many new machine learning classifiers, this study utilized both LOOCV and 10-fold CV to verify prediction accuracy (Fatima et al. 2020). Furthermore, a correlation score of 97.45% between the mean accuracies generated by these two cross-validation methods reinforced the models' prediction accuracy and result replicability (Fig 45 c) (Hosni et al. 2017).

Ensemble models have proven highly effective to enhance classification accuracy, with boosting techniques (AdaBoost algorithm) which had been widely used in various cancer classifications. However, for this dataset, bagging techniques (Random Forest and Extra Trees classifiers) performed exceptionally well (Vaka et al. 2020). Notably, the models generated in this study achieved accuracies ≥ 94% using non-invasive breast cancer features through bagging and boosting, surpassing the accuracy score of the XGboost classifier applied to Chinese women Hou et al. (2020). Research indicates that prolonged adherence to detrimental habits, such as using smokeless tobacco products (tuibur, sadha, gutkha), smoking, consuming alcohol, and chewing pan, can lead to genetic mutations associated with cancer, particularly tobacco-induced cancers, as reported by the Center for Disease Control and Prevention (US), 2010.

A potential strong association between tuibur, sadha, and pan chewing and breast cancer in Mizo women has been suggested, drawing parallels to a study by (Spangler et al. 2001) on Native American women (Eastern Band Cherokee). The similarities in lifestyle and dietary habits between Eastern Band Cherokee and Mizo

women, both from tribal populations, increase the likelihood of shared risk factors. Additionally, research has shown that women with a family history of breast cancer who engage in detrimental lifestyles are at higher risk, aligning with findings from (Zodinpuii et al. 2020). While this study did not find a direct association between smoking and breast cancer, exposure to secondhand smoke remains a significant risk factor for women, as demonstrated by (Li et al. 2015).

Although numerous machine learning models have been developed to address various aspects of breast cancer, including diagnosis, prognosis, screening, treatment outcomes, and genetic mutations, this study concentrates on identifying significant detrimental lifestyle factors that can be avoided to reduce the incidence of the disease (Assiri et al. 2020; Wu et al. 2021).

5.5 Classification of Breast Cancer using Genomic Features

Random Forest analysis of breast cancer NGS data identified key attributes from both balanced (SMOTEENN) and imbalanced (original) datasets. These attributes include Chr, Start, End, CADD raw score, GERP++_RS, LRT score, Polyphen2_HVAR score, MutationTaster score, SIFT score, and SiPhy_29way_logOdds (Rentzsch et al. 2019; Dong et al. 2015; Schwarz et al. 2014; Davydov et al. 2010; and Adzhubei et al. 2013). SMOTEENN enhances training with class balance by eliminating noise in the data and generating new instances based on nearest neighbors (Husainet al. 2025). To address multicollinearity among features, a heat map was employed (James et al. 2013), resulting in the removal of Polyphen2_HDIV and MutationAssessor scores. Feature selection, reduction of overfitting, and extraction of non-linear relationships between features are accomplished using Random Forest classifier (Pudjihartono et al. 2022). The selection of genetic attributes was further enhanced by developing Convolution Neural Networks (CNNs) on four datasets, including an imbalanced dataset and three oversampled datasets: ADASYN, SMOTEENN, and blsmote. Kalman-SMOTE technique had been applied to eliminate noise in the dataset that encompassed both the original data and the synthetically generated samples, which had yielded extensive testing on a diverse range of datasets (Thejas et al. 2022). ADASYN and

SMOTE oversampling methods had significantly improved classification performance on the small datasets with less positive rates (Zhu et al. 2024).A study was designed incorporating SVM for imputation and Adaptive Synthetic (ADASYN) sampling for feature utilization to address missing values, while leveraging CNN-generated features to enhance the model's capabilities. This model combination achieved exceptional performance, with 99.99% in accuracy, precision, recall, and F1 score (Munish 2024).Based on the model evaluation parameters, it was noteworthy that the feature subset identified by SMOTEENN could be the critical features for identifying pathogenic variants from the breast cancer germline dataset (Fig 20).CNNs had served as an effective tool for constructing classification models for breast cancer germline data, utilizing the SMOTEENN data balancing technique to address imbalance issues. Additionally, Random Forest Classifier identified significant patterns contributing to the disease. These findings can be applied to improve diagnosis accuracy, provide genetic counseling, identify suitable drug targets, and conduct personalized breast cancer risk assessments. There is a need to increase the number of healthy cases in the genomic data so the real data can be utilized to support the findings.

5.6 Classification of Head and Neck Cancer using Epidemiological Features

The classification results showed very interesting insights to the head and neck cancer risk factors. Cases who were age above 45, high frequency smoking, using snuff and alcohol intake, and prolonged alcohol intake had shown 85% accuracy by Extra trees classifier in compared to other classifiers. There were considerable imbalanced between false positive and false negatives, all the five models were able to minimize the false positives but had shoot up in false negatives. These might have been an impact as a result of removing the missing values and outliers. The above-identified risk factors had been shown to have greater interactions in Nepal population with odd’s ratio of 12.83 (Chang et al. 2020).

5.7 Classification of Head and Neck Cancer using Genomic Features

The top significant 10 features had been identified through Extra trees model: Chr, Start, End, Gene, MutationTaster score, FATHMM score, GERP++_RS, AF, non_neuro_AF_popmax, and ExAC_AMR. ExtraTrees classifier has not only been applied on medical data, it has been utilized for selecting relevant features for each type of security (intrusion detection). The extreme learning ensemble model, combined with softmax layer, has achieved an accuracy of 98%, which has been instrumental in identifying intrusion detection in real time (Sharma et al. 2019).The leukocyte classification by ResNet50 and feature selection by extra trees classifier yielded the highest accuracy of 90.76% (Baby et al. 2021).In the current dataset, the positive class was down-sampled due to significant data imbalance to mitigate bias towards the negative class. Similarly, downsampling was performed single cell transcriptome data based on unbiased datapoints which had effectively clustered the samples(Ren et al. 2022).

To assess the generalizability of the obtained results, a suite of five machine learning models (Logistic Regression, Random Forest, Decision Tree, Support Vector Machine and KNN) were developed (Carvalho et al. 2024; Kurian et al. 2022). Strength of this research is the integration of forward selection with grid search cross-validation hyperparameter tuning to identify the most relevant characteristics of genetic variants in head and neck cancer. However, a limitation is the potential loss of valuable information during the down-sampling process of the positive class.

5.8 Classification of Breast Cancer, Gastric Cancer, Head & Neck cancer

The present work had identified five common risk factors associated with gastric, breast, and head & neck cancers: khuva (paan), tuibur (aqueous tobacco extract), smoking, alcohol, snuff (smokeless tobacco products), and smoked food. These risk factors were reported in other population by evaluating them using machine learning classifiers and statistical tools (Roy et al. 2024; Zami et al. 2024 and Zodinpuii et al. 2022). The present study had developed a Multi-layer Perceptron

model that had effectively addressed this multi-class classification challenge (distinguishing between gastric, breast, head & neck cancers, and healthy cases) without overfitting.

In cancer classification, the MLP model can be trained on diverse features extracted from medical imaging, genetic data, or clinical information (Yang et al. 2022). The input layer receives these features, while the hidden layers perform intermediate computations to identify relevant patterns. The output layer subsequently provides the final classification, typically indicating the presence or absence of cancer, or differentiating between various cancer types.When applied to multiclass classification, the Multilayer Perceptron (MLP) output layer typically comprises one neuron per class. MLP had identified 38 genes signature patterns from breast cancer subtypes with accuracy of 93.56% (Yang et al. 2022). Regarding the existing methods and results obtained via MLPs on multiclass classification problems, significant advancements have been made in recent years. These improvements have led to enhanced accuracy and efficiency in categorizing complex datasets across various domains. However, challenges persist in optimizing neural network architectures and addressing issues such as overfitting and computational complexity for large-scale multiclass problems. Strengths of the proposed MLP model: To prevent overfitting, the model's hidden layer, neuron count, and 'lbfgs' optimizer were carefully selected. SMOTE oversampling was utilized to mitigate underfitting. Limitations of the proposed MLP model: Future research should aim to increase both the sample size and the number of features to elucidate additional factors concealed within epidemiological and genomic characteristics.

The results obtained using AdaBoost classifier (AI tools) can be used to design diagnostic gene-panel that are specific to gastric, breast and head & neck cancers in Mizo population (cancer high-risk). DNAH5 mutant had been involved in gastric cancer progression which was evident in the present study (Song et al. 2023). This gene was identified by machine learning tool, found to be one of the top most mutated genes in the gastric cancer big dataset (Fig 56). This study has also identified PARP4 in top 10 mutated genes of breast cancer and it had been found to

be potential candidate for breast cancer in other population (Prawira et al. 2019). In head and neck cancer, PRUNE2 was among the top 10 and it had been playing a role in oral and thyroid cancers (Storvall et al. 2023). First 20 mutated genes can be further screened to develop multi-cancer panel for Mizo population (Fig 57). This Multi-Cancer Panel examines genes linked to various organ systems that are largely related with adult-onset, non-syndrome cancer predisposition disorders. These gene panels can detect genetic alterations for gastric, breast, and head and neck cancers that could serve us better in risk assessment, diagnosis, and treatment choices.

Chapter 6. SUMMARY

The significant findings of this study are:

1. An algorithmic decision-making model was developed utilizing logistic regression to classify gastric cancer and healthy individuals based on epidemiological features, achieving an accuracy of 98.51%.

2. Saum (fermented pork fat), utilization of aluminum cookware, consumption of Zozial (infiltrated local cigar), initiation of smoking before 18 years of age, tuibur (aqueous tobacco extract), sadha (smokeless tobacco product), chewing of areca nuts (Khuva, zardhapan, supari), residence or occupation in proximity to cellular phone towers, and food storage in plastic containers were positively correlated with the incidence of gastric cancer in Mizoram.

3. Excessive salt intake, consumption of smoked foods, use of smokeless tobacco products (Khuva, Khaini/Sadha), and alcohol consumption were identified as risk factors associated with Microsatellite Instability in Gastric Cancer (MSI-GC).

4. This is the first study that had identified Khuva as a driving factor for MSI-GC, while excessive salt consumption, smoked food intake, Khaini/Sadha usage, and alcohol consumption were determined to be confounding factors for MSI-GC.

5. No biasness was observed during feature selection using 5-fold LASSOCV on balanced and imbalanced datasets.

6. Functional annotation (CADD and VEST3 scores, allele frequencies (East Asian, South Asian,Latino, and Ashkenazi Jewish), evolutionary conserved score (phyloP), gene, start position of the mutation, number of amino acid changes in gene (AAChange – use defined feature) reference allele (Ref) and mutated allele (Alt) were identified 96% of times as key attributes that causes pathogenicity in gastric cancer.

7. Ensemble classifiers have demonstrated superior overall performance in distinguishing between breast cancer cases and healthy controls based on epidemiological characteristics, achieving accuracies of 96.3% and 95.5% through LOOCV and 10-fold cross validation, respectively.

8. Heterozygous mutations were found to be higher in breast cancer dataset compared to healthy controls.

9. Features identified by Random forest were chromosome, position (start and end), function prediction scores (CADD, GREP, Polyphen, MutationTaster, LRT and SIFT) which showed well-balanced classification results on breast cancer big data.

10. On oversampling healthy datapoints by three distinct techniques (ADASYN, SMOTEENN and blsmote) from breast cancer big dataset had generated same feature subsets.

11. CNN model designed on (SMOTEENN) balanced dataset had good trade-off between precision and recall of 99.32% and 97.97% respectively with ROC-AUC of 100%

12. Individuals above the age of 45, with frequent smoking, snuff use, and prolonged alcohol consumption, were classified with 85% accuracy using the Extra Trees classifier on head and neck epidemiological characteristics.

13. Decision tree and SVM classifies have produced 96% accuracy with low false positive rate of 1 using head and neck germline variants with healthy controls.

14. Non-invasive characteristics were examined to elucidate potential cancer risk factors for gastric, breast, and head and neck malignancies.

15. Specific dietary habits and food preferences were found to distinctively contribute various cancer disparities.

16. A high-performing MLP model was selected based on its reduced computational time and enhanced efficiency to classify breast, gastric, head & neck and healthy controls.

17. Genes (nodes) and their functional protein association in gastric cancer big dataset were identified based on their experimental evidences: edges (two nodes interaction): GPSM2-NUMA1; RAG1-RAG2; MTERF4-NSUN4; WDR18-PELP1; CD58-CD2; and significant hubs (more than two nodes interactions): FANCM-MEIOB-FANCA-BLM-EXO1-TP53BP1; MCM3-TOPBP1-TP53BP1-CHEK1-CDC25C; TGFBRAP1-VPS39-VPS8; CHD23-USH1C-MYO7A; CWC22-CRNKL1-SLU7; PVR-CD96-CD226.

18. Genes (nodes) and their functional protein association in breast cancer big dataset were identified based on their experimental evidences: edges (two nodes interaction): UNC13B-RIMS2; ABCG5-ABCG8; PER3-PER1; F5-F2; DCTN1-ACTR1B; POLR1E-POLA1; and significant hubs (more than two nodes interactions): MPHOSPH10-PDCD11-NAT10-HEATR1-NOL10; CDH23-USH1C-MYO7B; RPA1-RPA4-TIPIN-TIMELESS-WRN-RFWD3-FANCA; DYNC2H1-DYNC2I2-DYNC2LI1; TNRC6C-AGO2-GIGYF1.

19. Genes (nodes) and their functional protein association in head and neck cancer big dataset were identified based on their experimental evidences: edges (two nodes interaction):BCOR-PCGF1; CWC22-SRRM2; CEP55-TEX14; ABCG5-ABCG8; PER3-PER2; LLGL2-PARD6B; TUBGCP3-TUBG2; COG6-COG7; COL4A1-COL4A2; WNK1-WNK2; ORC4-ORC1; TACC3-AURKA; BCOR-PCGF1; ORC4-ORC1; PENK-OPRM1; and significant hubs (more than two nodes interactions): TP53BP1-TOPBP1-BRAC1-ATR-TP53;TRAF3-MAVS-IFIH1-TRIM65; RAD23B-XPC-PSMD2-TDP2; C5-C7-C8A; DDX18-NOC3L-PES1-PDCD11.

20. Genes: MDN1, MKI67, DNAH5, FAT2, FAT1, SYNE1, SYNE2, USH2A, and CMYA5 in GC; XIRP2, USH2A, OR51B6, FBN3, PARP4 in BC; NEB AND OBSCN were common in all the three cancer types found to hold high prospective to

develop gastric, breast and head & neck cancer in Mizo population which were evident from machine learning and deep learning models.

Machine learning algorithms were able to evidently identify the gastric, breast and head & neck cancer risk factors from diet-lifestyle with accuracies greater than 95%. Deep learning algorithms have shown great efficiency in detecting the foremost genetic traits of germline variants from gastric, breast and head & neck cancers with precision and recall greater than 95%. The epidemiological risk pattern of the above-mentioned three cancers had been achieved with accuracy of 75% using computationally less intensive architecture designed using MLP. The genomic risk factors of these three cancers had given an accuracy of 84.18% and f1_score of 84.43% by AdaBoost-hyperparameter tuned algorithm. The algorithm had identified twenty genes by integrating germline cancer dataset which will be signature genes to develop multi-cancer genetic panel for this cancer-burden population. There were many node (gene) interactions between each cancer used in this study, they are showing very interesting and unique patterns which gives different insight to understand the cause of cancer and its propagation within the cells.

Chapter 7. BIBILIOGRAPHY

Abdulsadig, R. S., & Rodriguez-Villegas, E. (2024).A comparative study in class imbalance mitigation when working with physiological signals.*Frontiers in Digital Health, 6*, 1377165.

Adam, R., Dell'Aquila, K., & Hodges, L. (2023). Deep learning applications to breast cancer detection by magnetic resonance imaging: A literature review. *Breast Cancer Research, 25*, 87.

Adzhubei, I., Jordan, D. M., &Sunyaev, S. R. (2013).Predicting functional effect of human missense mutations using PolyPhen-2.*Current Protocols in Human Genetics, 2013*(Chapter 7), Unit7.20.

Akselrod-Ballin, A., Chorev, M., &Shoshan, Y. (2019). Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, *292*(2), 331–342.

Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A., &Mäkitie, A. A. (2024). Artificial intelligence-driven radiomics in head and neck cancer: Current status and future prospects. *International Journal of Medical Informatics, 188*, 105464.

Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., Bahiyah, N., Benes, F., & Rhee, J. (2022).Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. *Computers*, *11*(9), 136

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: *http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed 22 September 2022*.

Anitha, M., Gayathri, S., Nickolas, S., & Bhanu, M. S. (2020). Feature engineering-based automatic breast cancer prediction. *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 247–256.

Arlot, S., & Celisse, A. (2010).A survey of cross-validation procedures for model selection.*Statistical Surveys, 4*, 40–79. Arya, M., Sastry, G. H., Motwani, A., Kumar, S., &Zaguia, A. (2022, February 15). A novel extra tree ensemble optimized DL framework (ETEODL) for early detection of diabetes. *Frontiers in Public Health, 9*, 797877.

Arya, M., Sastry, G. H., Motwani, A., Kumar, S., & Zaguia, A. (2022). A Novel Extra Tree Ensemble Optimized DL Framework (ETEODL) for Early Detection of Diabetes. Front. Public Health, 9:797877.

Assiri, A.S., Nazir, S., & Velastin, S. (2020). Breast tumor classification using Ensemble machine learning Method.*Journal of Imaging, 6,* 39.

Auton, A., Brooks L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., & Abecasis, G.R. (2015) A global reference for human genetic variation.*Nature. 1;526(7571):68-74*. 1000 Genomes Project Consortium;.

Baby, D., Devaraj, S. J., Hemanth, J. &ANISHIN, R. M. (2021) Leukocyte classification based on feature selection using extra trees classifier: atransfer learning approach. *Turkish Journal of Electrical Engineering and Computer Sciences,:11;(29):*8

Bang, C., Bernard, G., Le, W. T., Lalonde, A., Kadoury, S., &Bahig, H. (2023).Artificial intelligence to predict outcomes of head and neck radiotherapy.*Clinical and Translational Radiation Oncology, 39*, 100590.

Bertoncelli, C. M., Altamura, P., Vieira, E. R., Iyengar, S. S., Solla, F., &Bertoncelli, D. (2020). PredictMed: A logistic regression–based model to predict health conditions in cerebral palsy. *Health Informatics Journal, 26*(4), 2105–2118.

Binzagr, F. (2024).Explainable AI-driven model for gastrointestinal cancer classification.*Frontiers in Medicine (Lausanne), 11*, 1349373.

Bolger, A.M., Lohse, M., &Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics 30 (15), 2114–2120.*

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Brindha, S.K., Senthil Kumar, N., Chenkual, S., Lalruatfela, S., & Zomuana, T., Ralte, Z., Maitra, A., & Basu, A., & Nath, Prem. (2020).Data Mining for Early Gastric Cancer Etiological Factors from Diet-Lifestyle Characteristics.*2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 65–73.

Broggi, G., Maniaci, A., Lentini, M., Palicelli, A., Zanelli, M., Zizzo, M., Koufopoulos, N., Salzano, S., Mazzucchelli, M., &Caltabiano, R. (2024). Artificial intelligence in head and neck cancer diagnosis: A comprehensive review with emphasis on radiomics, histopathological, and molecular applications. *Cancers, 16*(21), 3623.

Carriero, A., Groenhoff, L., Vologina, E., Basile, P., &Albera, M. (2024). Deep learning in breast cancer imaging: State of the art and recent advancements in early 2024. *Diagnostics*, *14*(8), 848.

Carvalho, D., Capriles, P., & Goliatt, L. (2024). Comparative analysis of machine learning models for breast cancer patients' survival prediction. *Advances in Artificial Intelligence and Data Engineering* (pp. 305-315).

Çetin, A., &Büyüklü, A. (2024).A new approach to K-nearest neighbors distance metrics on sovereign country credit rating. *Kuwait Journal of Science, 52*(1).

Chakraborty, P., Kurkalang, S., Ghatak, S., Das, S., Palodhi, A., Sarkar, S., Dhar, R., Chenkual, S., Pachuau, L., Zohmingthanga, J., Pautu, J. L., Zomuana, T., Lalruatfela, S. T., Zothanzama, J., Kumar, N. S., &Maitra, A. (2023). Deep sequencing reveals recurrent somatic mutations and distinct molecular subgroups in gastric cancer in Mizo population, North East India. *Genomics, 115*(6), 110741.

Chang, C.P., Siwakoti B, Sapkota, A., Gautam, D.K., Lee, Y.A., Monroe, M., Hashibe, M. (2020) Tobacco smoking, chewing habits, alcohol drinking and the risk of head and neck cancer in Nepal.*Int J Cancer*, 1;147(3):866-875.

Chen, Q., Meng, Z., Liu, X., Jin, Q., & Su, Ran.(2018). Decision Variants for the Automatic Determination of Optimal Feature Subset in RF-RFE.*Genes. 9*. 301.

Chen, R.C., Dewi, C., Huang, S.W. et al. (2020). Selecting critical features for data classification based on machine learning methods. *J Big Data 7*, 52.

Chicco, D., and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak, 20*, 16.

Chollet, F., et al., et al. (2015).Keras.Retreived from https://github.com/keras-team/keras.

Cobaleda, C., Godley, L. A., Nichols, K. E., Wlodarski, M. W., & Sanchez-Garcia, I. (2024). Insights into the molecular mechanisms of genetic predisposition to hematopoietic malignancies: The importance of gene-environment interactions. *Cancer Discovery, 14*(3), 396–405.

Das, D., Nayak, M., Pani, S. K. (2019).Missing Value Imputation-A Review.*International Journal of Computer Sciences and Engineering, 7*, 548-558.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++.*PLoS Comput Biol, 2*;6(12):e1001025.

DeCastro-García, N., Muñoz Castañeda, Á., García, D., & Carriegos, M.V. (2019). Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm.*Complexity, 1-16.*

Dias, R., & Torkamani, A. (2019) Artificial intelligence in clinical and genomic diagnostics. *Genome Med,* 11, 70

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies.*Hum Mol Genet, 15*;24(8):2125-37.

Dong, Q., Liu, M., Chen, B., Zhao, Z., Chen, T., Wang, C., Zhuang, S., Li, Y., Wang, Y., Ai, L., Liu, Y., Liang, H., Qi, L., &Gu, Y. (2021). Revealing biomarkers associated with PARP inhibitors based on genetic interactions in cancer genome. *Computational and Structural Biotechnology Journal, 19*, 4435-4446.

Douville, C., Masica, D. L., Stenson, P. D., Cooper, D. N., Gygax, D. M., Kim, R., Ryan, M., & Karchin, R. (2016).Assessing the pathogenicity of insertion and deletion variants with the Variant Effect Scoring Tool (VEST-Indel). *Human Mutation, 37*(1), 28–35.

El-Hassani, F. Z., Amri, M., Joudar, N.-E., & Haddouch, K. (2024). A new optimization model for MLP hyperparameter tuning: Modeling and resolution by real-coded genetic algorithm. *Neural Processing Letters, 56*.

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*.

Emmanuel, T., Maupong, T., & Mpoeleng, D. (2021).A survey on missing data in machine learning.*Journal of Big Data, 8*, 140.

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020).Prediction of breast cancer, comparative review of machine learning techniques, and their analysis.*IEEE Access, 8*, 150360–150376.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010).Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

Garcia, F. A. O., de Andrade, E. S., & Palmero, E. I. (2022).Insights on variant analysis in silico tools for pathogenicity prediction. *Frontiers in Genetics, 13*, 1010327.

Gupta, P., & Malhi, A. (2018).Using deep learning to enhance head and neck cancer diagnosis and classification.*ICSCAN*, 1-6.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines.*Machine Learning, 46*(1–3), 389–422.

Hajian-Tilaki K. (2013).Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.*Caspian J Intern Med*, 4(2):627-35.

Hancock, J. T., Khoshgoftaar, T. M., & Johnson, J. M. (2023).Evaluating classifier performance with highly imbalanced big data.*Journal of Big Data, 10*, 42.

Hasanin, T., Khoshgoftaar, T. M., &Leevy, J. L. (2019). Severely imbalanced big data challenges: Investigating data sampling approaches. *Journal of Big Data, 6*, 107.

Hemphill, S.A,, Tollit, M., Kotevski, A., and Heerde, J.A. (2015). Predictors of Traditional and Cyber-Bullying Victimization: A Longitudinal Study of Australian Secondary School Students. *J Interpers Violence*, 30(15):2567-90.

Hoque, N., Bhattacharyya, D. K., & Kalita, J. (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications, 41*, 6371–6385.

Hosni, M., Idri, A., Abran, A., & Nassif, A. B. (2017).On the value of parameter tuning in heterogeneous ensembles effort estimation.*Soft Computing*, 1–34.

Hui, Y., Tu, C., Liu, D., Zhang, H., & Gong, X. (2023). Risk factors for gastric cancer: A comprehensive analysis of observational studies. *Frontiers in Public Health, 10*, 892468.

Hunter, J. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*.90-95.

Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T., & Toma, M. (2025). SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models. *Algorithms*, *18*(1), 37.

Hussain, S., Ali, M., Naseem, U., Nezhadmoghadam, F., Jatoi, M. A., Gulliver, T. A., &Tamez-Peña, J. G. (2024). Breast cancer risk prediction using machine learning: A systematic review. *Frontiers in Oncology, 14*, 1343627.

Imaizumi, H., Sato, K., Nishihara, A., Minami, K., Koizumi, M., Matsuura, N., &Hieda, M. (2018). X-ray-enhanced cancer cell migration requires the

linker of nucleoskeleton and cytoskeleton complex. *Cancer Science, 109*(4), 1158–1165.

James, G., Daniela, Witten., Trevor, Hastie., & Robert, T. (2013). An Introduction to Statistical Learning with Application in r. *Springer*, eISBN: 978-1-4614-7137-7

Jayarama, S., Li, L. C., Ganesh, L., Mardi, D., Kanteti, P., Hay, N., Li, P., &Prabhakar, B. S. (2014). MADD is a downstream target of PTEN in triggering apoptosis. *Journal of Cellular Biochemistry, 115*(2), 261–270.

Jordaan, E.M. & Smits, G. (2002).Estimation of the regularization parameter for support vector regression.*IJCNN.2002, 3,* 2192 - 2197.

Karagianni, V., &Triantafillidis, J. K. (2010). Prevention of gastric cancer: Diet modifications. *Annals of Gastroenterology, 23*(4), 237–242.

Khairunnahar, L., Hasib, M. A., Rezanur, R. H. B., Islam, M. R., &Hosain, M. K. (2019).Classification of malignant and benign tissue with logistic regression.*Informatics in Medicine Unlocked, 16*, 100189

Khosravi, M., Jasemi, S. K., Hayati, P., Javar, H. A., Izadi, S., &Izadi, Z. (2024). Transformative artificial intelligence in gastric cancer: Advancements in diagnostic techniques. *Computers in Biology and Medicine, 183*, 109261.

Kim, H. E., Kim, H. H., Han, B. K., Kim, K. H., Han, K., Nam, H., Lee, E. H., & Kim, E. K. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: A retrospective, multireader study. *Lancet Digital Health, 2*(3), e138–e148.

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S.E., & Topper, S.E. (2017). Pathogenic variant burden in the ExAC database: An empirical approach to evaluating population data for clinical variant interpretation. *Genome Medicine, 9*, 13.

Kumar, Y., Shrivastav, S., &Garg, K. (2024). Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification.*Scientific Reports, 14*, 25006.

Kurian, B., & Jyothi, V. L. (2022).Comparative analysis of machine learning methods for breast cancer classification in genetic sequences.*Journal of Environmental and Public Health, 2022*, 7199290.

Lee, W., Lee, H., Lee, H., Park, E. K., Nam, H., & Kooi, T. (2023).Transformer-based deep neural network for breast cancer classification on digital breast tomosynthesis images.*Radiology: Artificial Intelligence, 5*(3), e220159.

Lee, W., & Seo, K. (2022).Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning.*Big Data Research, 28*, 100314.

Leitheiser, M., Capper, D., Seegerer, P., Lehmann, A., Schüller, U., Müller, K.R., Klauschen, F., Jurmeister, P., & Bockmayr, M. (2022). Machine learning models predict the primary sites of head and neck squamous cell carcinoma metastases based on DNA methylation. J Pathol, 2022 256(4):378-387

Li, B., Wang, L., Lu, M.S., Mo, X.F., Lin, F.Y., Ho, S.C., et al. (2015) Passive Smoking and Breast Cancer Risk among Non-Smoking Women: A Case-Control Study in China. *PLoS ONE,* 10(4): e0125894

Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs With BWA-MEM. https://doi.org/10.48550/arXiv.1303.3997.

Li, L., Chen, Y., Shen, Z., Zhang, X., Sang, J., Ding, Y., Yang, X., Li, J., Chen, M., Jin, C., Chen, C., & Yu, C. (2020). Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow-band imaging.*Gastric Cancer, 23*(1), 126–132.

Lin, J. C., Wang, C. C., Jiang, R. S., Wang, W. Y., & Liu, S. A. (2016). Microsatellite alteration in head and neck squamous cell carcinoma patients from a betel quid-prevalent region. *Sci Rep 6*, 22614.

Liu, C., Duan, Y., Zhou, Q., Wang, Y., Gao, Y., Kan, H., & Hu, J. (2023). A classification method of gastric cancer subtype based on residual graph convolution network.*Frontiers in Genetics, 13*, 1090394.

Liu, L., Wu, X., & Li, S. (2022). Solving the class imbalance problem using ensemble algorithm: Application of screening for aortic dissection.*BMC Medical Informatics and Decision Making, 22*, 82.

Lysdahlgaard, S. (2023).Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs.*Radiography (London), 29*(6), 1132–1138.

Madani, M., Behzadi, M. M., &Nabavi, S. (2022). The role of deep learning in advancing breast cancer detection using different imaging modalities: A systematic review. *Cancers, 14*(21), 5334.

McNamee R. (2005). Regression modelling and other methods to control confounding.Occup Environ Med.;62(7):500-6, 472.

Munshi, R.M. (2024) Correction: Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. *PLOS ONE, 19*(2): e0298980

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 18-20

Najafabadi, M.M., Khoshgoftaar, T.M., Villanustre, F., & Holt, J. (2017). Large-scale distributed L-BFGS . J Big Data 4, 22.

Nakanishi, H., Ohtsubo, M., Iwasaki, S., Hotta, Y., Takizawa, Y., Hosono, K., Mizuta, K., Mineta, H., &Minoshima, S. (2010). Mutation analysis of the MYO7A and CDH23 genes in Japanese patients with Usher syndrome type 1.*Journal of Human Genetics*, *55*(12), 796–800.

National Centre for Disease Informatics and Research.(n.d.).Consolidated report of population-based cancer registries, 2006–2008, 2009–2011, 2012–2014.*National Cancer Registry Programme (NCRP-ICMR)*. Retrieved from https://ncdirindia.org/Reports.aspx

Nono, A. D., Chen, K., & Liu, X. (2019).Comparison of different functional prediction scores using a gene-based permutation model for identifying cancer driver genes.*BMC Medical Genomics*, *12* (Suppl 1), 22.

Noroozi, Z., Orooji, A., &Erfannia, L. (2023).Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction.*Scientific Reports*, *13*, 22588.

Oshiro, T. M., Perez, P. S., &Baranauskas, J. A. (2012). How many Trees in a Random Forest? *Lecture Notes in Computer Science*, *7376*, 154.

Pachuau, L., Lalremmawia, H., Ralte, L., Vanlalpeka, J., Pautu, J. L., Chenkual, S., Zomuana, T., Lalruatfela, S. T., Zohmingthanga, J., Chhakchhuak, L., Varma, A. K., & Kumar, N. S. (2025). Uncovering novel pathogenic variants and pathway mutations in triple-negative breast cancer among the endogamous Mizo tribe.*Breast Cancer Research and Treatment*, *209*(2), 375–387.

Pachuau, L., Zami, Z., Nunga, T., Zodingliana R., Zoramthari, R., Lalnuntluanga R., Sangi Z., Rinmawii, L., Senthil Kumar, N., & Lalhruaitluanga, H.(2022). First-degree family history of cancer can be a potential risk factor among head and neck cancer patients in an isolated Mizo tribal population, Northeast India. *Clinical Epidemiology and Global Health*, *13*, 100954.

Payel, C., Kurkalang, S., Ghatak, S., Das, S., Palodhi, A., Sarkar, S., Dhar, R., Chenkual, S., Pachuau, L., Zohmingthanga, J., Pautu, J. L., Zomuana, T., Lalruatfela, S. T., Zothanzama, J., Kumar, N. S., &Maitra, A. (2023). Deep sequencing reveals recurrent somatic mutations and distinct molecular subgroups in gastric cancer in Mizo population, North East India. *Genomics*, *115*(6), 110741.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., *et al.*, *J Machine Learning* Research, 12, pp. 2825-2830, 2011

Prawira, A., Munusamy, P., Yuan, J., Chan, C.H.T., Koh, G.L., Shuen, T.W.H., Hu, J., Yap, Y.S., Tan, M.H., Ang, P., & Lee, A.S.G. (2019). Assessment of PARP4 as a candidate breast cancer susceptibility gene. Breast Cancer Res Treat, 177(1):145-153.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W. & O'Sullivan, J.M. (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform.* 2:927312.

Qiu, L., Zhou, R., Zhou, L., Yang, S., & Wu, J. (2022). CAPRIN2 upregulation by LINC00941 promotes nasopharyngeal carcinoma ferroptosis resistance and metastatic colonization through HMGCR. *Frontiers in Oncology, 12*, 931749.

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., &Atashi, A. (2022). Prediction of breast cancer using machine learning approaches. *Journal of Biomedical Physics & Engineering*, *12*(3), 297–308.

Ren, J., Zhang, Q., Zhou, Y., Hu, Y., Lyu, X., Fang, H., Yang, J., Yu, R., Shi, X., & Li, Q. (2022). A downsampling method enables robust clustering and integration of single-cell transcriptome data. *Journal of Biomedical Informatics, 130*, 104093.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., &Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research, 47*(D1), D886–D894.

Reva, B., Antipin, Y., & Sander, C. (2011).Predicting the functional impact of protein mutations: Application to cancer genomics.*Nucleic Acids Research, 39*(17), e118.

Rohatgi, O., Agrawal, N., Goswami, S., Vivek, V., Chaudhuri, M., &Aparasu, R. (2023). MSR75 Machine learning models in prediction of head and neck squamous cell carcinoma survivability. *Value in Health, 26*, S291.

Roy, A. D., Mukherjee, M., Kumar, S., & Kumar, D. (2024).The burden of gastrointestinal cancers in north east India with special reference to Helicobacter pylori infection. *Indian J Forensic Community Med, 11*(4):147-151

Rufibach, K. (2010). Use of Brier score to assess binary predictions.*Journal of clinical epidemiology,63*, 938-9.

Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons, and challenges. *Journal of Infection and Public Health, 13*(9), 1274–1289.

Safakish, A., Sannachi, L., DiCenzo, D., Kolios, D., PejovicMilic, A., & Czarnota G. J. (2023). Predicting head and neck cancer treatment outcomes with pre-treatment quantitative ultrasound texture features and optimizing machine learning classifiers with texture-of-texture features. *Frontiers in Oncology, 13*, 1258970.

Salehin, I., & Kang, D.-K. (2023). A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics, 12*(14), 3106.

Schober, P., & Vetter, T. R. (2021).Logistic regression in medical research.*Anesthesia& Analgesia, 132*(2), 365–366.

Schwarz, J.M., Cooper, D.N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods, 11*(4):361-2.

Sergey, I., &Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv Preprint arXiv:1502.03167.*

Shadeed, I., Alwan, J., &Abd, D. (2020).The effect of gamma value on support vector machine performance with different kernels.*International Journal of Electrical and Computer Engineering (IJECE), 10*, 5497-5506.

Shaon, Md. S. H., Karim, T., Shakil, Md.S.,& Hasan, Md. Z. (2024). A Comparative Study of Machine Learning Models with LASSO and SHAP Feature Selection for Breast Cancer Prediction.*Healthcare Analytics, 6*. 100353.

Sharma, J., Giri, C., Granmo, OC. et al. (2019). Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. *EURASIP J. on Info.Security, 15.*

Shi, J., Cheng, C., Ma, J., Liew, C. C., &Geng, X. (2018).Gene expression signature for detection of gastric cancer in peripheral blood.*Oncology Letters, 15*(6), 9802–9810.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., & Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models.*Human Mutation, 34*(1), 57–65.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., & Gaunt, T. R. (2013). Predicting the functional, molecular,

and phenotypic consequences of amino acid substitutions using hidden Markov models.*Human Mutation, 34*(1), 57–65.

Siddoo-Atwal, C. (2018). Electromagnetic radiation from cellphone towers: A potential health hazard for birds, bees, and humans. *Environmental Health Perspectives, 126*(9).

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research, 40*, W452–W457.

Song, Z., Song, X., Li, H., Cheng, Z., Mo, Z., & Yang, X. (2023). Identification and validation of a prognostic-related mutant gene DNAH5 for hepatocellular carcinoma. Front Immunol, 14:1236995.

Spangler, J. G., Michielutte, R., Bell, R. A., &Dignan, M. B. (2002).Association between smokeless tobacco use and breast cancer among Native-American women in North Carolina.*Ethnicity & Disease, 11*(1), 36–43.

Stark, L., Kasajima, A., &Stögbauer, F. (2024). Head and neck cancer of unknown primary: Unveiling primary tumor sites through machine learning on DNA methylation profiles. *Clinical Epigenetics, 16*, 47.

Storvall, S., Ryhänen, E., Karhu, A., & Schalin-Jäntti, C. (2023). Novel PRUNE2 Germline Mutations in Aggressive and Benign Parathyroid Neoplasms. Cancers, 15(5), 1405.

Su, S. C., Chang, L. C., Lin, C. W., Chen, M. K., Yu, C. P., Chung, W. H., & Yang, S. F. (2019). *Human Genetics*, *138*, 1379.

Sultana, J., &Jilani, A. K. (2018).Predicting breast cancer using logistic regression and multi-class classifier.*International Journal of Engineering and Technology, 7*(4.20), 22–26.

Sun, Y., Zhu, S., & Ma, K. (2019).Identification of 12 cancer types through genome deep learning.*Scientific Reports, 9*, 17256.

Sylvain, A., & Alain, C. (2010).A survey of cross-validation procedures for model selection.*Statistical Surveys, 4*, 40–79.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research, 51*(D1), D638-D646.

Taksler, G. B., Keating, N. L., & Rothberg, M. B. (2018).Implications of false-positive results for future cancer screenings.*Cancer, 124*(11), 2390–2398.

Taninaga, J., Nishiyama, Y., Fujibayashi, K.,Gunji, T., Sasabe, N., Iijima, K., and Naito, T. (2019). Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. *Scientific Reports, 9*, 12384.

Teh, K., Armitage, P., Tesfaye, S., Selvarajah, D., & Wilkinson, I. D. (2020). Imbalanced learning: Improving classification of diabetic neuropathy from magnetic resonance imaging. *PLoS ONE, 15*(12), e0243907.

Thapa, S., Lalrohlui, F., Ghatak, S., Zohmingthanga, J., Lallawmzuali, D., Pautu, J. L., &Senthil Kumar, N. (2016). Mitochondrial complex I and V gene polymorphisms associated with breast cancer in Mizo-Mongoloid population. *Breast Cancer, 23*(4), 607-616.

Thejas, G. S., Hariprasad, Y., Iyengar, S.S., & Sunitha, N. R. (2022). An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets. *Machine Learning with Applications,8,* 100267. 10.

Tsai, C.F., Chen, K. C., & Lin, W.C. (2024). Feature Selection and Its Combination with Data Over-sampling for Multi-Class Imbalanced Datasets. *Applied Soft Computing,153*, 111267.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A.J. (2019).Machine learning algorithm validation with a limited sample size. *PLoS ONE*, 14(11): e0224365.

Vaka, A. R., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging machine learning.*ICT Express, 6*(4), 320–324.

Van der Auwera, G.A., &O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, 1st ed. *O'Reilly Media, California*.

van Rijn, E., Eichenlaub, J.B., Lewis, P.A., Walker, M.P., Gaskell, M.G., Malinowski, J.E., & Blagrove, M. (2015). The dream-lag effect: Selective processing of personally significant events during Rapid Eye Movement sleep, but not during Slow Wave Sleep. *Neurobiol Learn Mem, 122*:98-109.

Viering, T., & Loog, M. (2021). The shape of learning curves: A review. *arXiv:2103.10948*.

Wang, H., Shi, J., Yang, Y., Ma, K., & Xue, Y. (2024). Machine learning methods predict recurrence of pN3b gastric cancer after radical resection. *Translational Cancer Research, 13*(3), 1519–1532.

Wang, L. (2024). Mammography with deep learning for breast cancer detection.*Frontiers in Oncology, 14*, 1281922.

Wu, J., & Hicks, C. (2021).Breast cancer type classification using machine learning.*Journal of Personalized Medicine, 11*(2), 61.

Xiong, M., Deng, A., Koh, P., Wu, J., Li, S., Xu, J., & Hooi, B. (2023).Proximity-Informed Calibration for Deep Neural Networks.*10.48550/arXiv.2306.04590*.

Yadav, R. P., Ghatak, S., Chakraborty, P., Lalrohlui, F., Kannan, R., Kumar, R., Pautu, J. L., Zomingthanga, J., Chenkual, S., Muthukumaran, R., &Senthil Kumar, N. (2018). Lifestyle chemical carcinogens associated with mutations in cell cycle regulatory genes increase susceptibility to gastric cancer risk. *Environmental Science and Pollution Research International, 25*(31), 31691–31704.

Yang, X., Zheng, Y., Xing, X., Sui, X., Jia, W., & Pan, H. (2022).Immune subtype identification and multi-layer perceptron classifier construction for breast cancer.*Frontiers in Oncology, 12*, 943874.

Yang, Y., Khorshidi, H. A., &Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Frontiers in Digital Health, 6*, 1430245.

Yu, S. H., Cai, J. H., Chen, D. L., Liao, S. H., Lin, Y. Z., Chung, Y. T., & Wang, C. C. N. (2021). LASSO and bioinformatics analysis in the identification of key genes for gynecologic cancer.*Journal of Personalized Medicine, 11*(11), 1177.

Zami, Z., Pachuau, L., Bawihtlung, Z., Khenglawt, L., Hlupuii, L., Lalthanpuii, C., Hruaii, V., Lalhruaitluanga, H., & Kumar, N.S. Treatment regimens and survival among patients with head and neck squamous cell carcinoma from Mizo tribal population in northeast India - a single centre, retrospective cohort study. *Lancet Reg Health Southeast Asia*, *24*:100377.

Zeng, P., Zhao, Y., Liu, J., Liu, L., Zhang, L., Wang, T., Huang, S., & Chen, F. (2014). Likelihood ratio tests in rare variant detection for continuous phenotypes. *Ann Hum Genet*, 78(5):320-32.

Zhang, J., Qiu, W., Liu, H., et al. (2018). Genomic alterations in gastric cancers discovered via whole-exome sequencing. *BMC Cancer, 18*, 1270.

Zhang, K., Wang, H., & Cheng, Y. (2024). Early gastric cancer detection and lesion segmentation based on deep learning and gastroscopic images. *Scientific Reports, 14*, 7847.

Zhang, R., Li, H., Li, N., Shi, J.-F., Li, J., Chen, H.-D., Yu, Y.-W., Qin, C., Ren, J.-S., Chen, W.-Q., & He, J. (2021). Risk factors for gastric cancer: A large-scale, population-based case-control study. *Chinese Medical Journal, 134*(16), 1952–1958.

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine, 4*(11), 218.

Zhong, C., Li, K. N., Bi, J. W.,& Wang, B.C. (2012).Sodium intake, salt taste and gastric cancer risk according to Helicobacter pylori infection, smoking, histological type and tumor site in China.*Asian Pacific Journal of Cancer Prevention, 13*, 2481–2484.

Zhu, C., Idemudia, C. U., & Feng, W. (2019).Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques.*Informatics in Medicine Unlocked, 17*, 100179.

Zhu, J., Pu, S., He, J., Su, D., Cai, W., Xu, X., & Liu, H. (2024). Processing imbalanced medical data at the data level with assisted-reproduction data as an example.*BioData Mining, 17*.

Zodinpuii, D., Pautu, J. L., Zothankima, B., Khenglawt, L., Lallawmzuali, D., Lalmuanpuii, R., Zuali, L., Ralte, L., Muthukumaran, R. B., Varma, A. K., Zothanzama, J., & Kumar, N. S. (2022). Breast cancer is significantly associated with cancers in the first- and second-degree relatives in ethnic Mizo-Mongoloid population, Northeast India. *National Journal of Community Medicine, 13*(9), 606-611.

Zomawia, E., Zami, Z., Vanlallawma, A., Senthil Kumar, N., Zothanzama, J., Tlau, L., Chhakchhuak, L., Pachuau, L., Pautu, J. L., &Hmangaihzuali, E. V. L. (2023). Cancer awareness, diagnosis and treatment needs in Mizoram, India: Evidence from 18 years trends (2003–2020). *The Lancet Regional Health – Southeast Asia, 17*, 100281.

BIO DATA OF THE CANDIDATE

Brindha Senthil Kumar

Education qualification: Ph.D. (pursuing) in Computer Science and Engineering

Address : Dept of Computer Engineering, Mizoram University, Mizoram - 796004, India

Email: brindhamzu@gmail.com

Career Objective

I would like to contribute myself in sharing my knowledge and develop Intelligent Decision-Making models to combat against life-threatening diseases. I like to learn and carry out research in mining biological information using Artificial Intelligence Tools**.** The huge accumulation of genetic and epidemiological characteristics is great insight to seek the signature mutations, patterns of diseases and to find the pathways of the biological reactions.

Educational Qualifications

1. **Doctorate**: **Pursuing PhD**, Department of Computer Engineering, Mizoram University
   Year: 2020 – till date

2. **Postgraduation**: M.Tech in Computer Science and Engineering, Department of Information Technology, MizoramUniversity
   Year: 2018 – 2020**CGPA**:9.66 (**Gold Medal**)

3. **Undergraduation**: B.E in Computer Science and Engineering, BharathiarUniversity, Coimbatore, Tamilnadu
   Year: 1995 – 1999**CGPA**:6.55

3. **Higher Secondary**: Our Lady of Lourdes Matriculation Higher Secondary School, Pollachi, Tamilnadu
   Year Passed: 1995
   Percentage: 76.60%

Certification Courses

Certifications	***Grade***	***Year***	***/Institutions***
Certification Course on Modern Machine Learning	B	2022	IIIT, Hyderabad, **India**
Certification in C++, Core Java and J2EE	First class	2004	SSI Computer Institute (**Pollachi, India**).
Certification in Oracle with Developer 2000 and Visual Basic	First Class	1999	TULEC Computer Education (**Coimbatore, India**)
Medical Transcriptionist	-	2000	Baston Solution (**Coimbatore, India**)
IELTS	**6/9**	2006	-

Technical Skills

LANGUAGES : C, C++, Java, JSP/Servlets, J2EE.

PACKAGES : Python, Scikit learn, Numpy, Pandas, Seaborn, Tensorflow, Keras, PHP, VISUAL BASIC, .NET & Developer.

DATABASES : SQL Server Oracle and MS Access

OPERATING SYSTEMS : Windows, DOS, UNIX

AREAS OF INTEREST

Artificial Intelligence, Biomedical Informatics ,Programming.

Work Experience and Projects Completed

Position	***Duration***	***Nature of Duties***	***Company/Institution***
Project Assistant (SERB, New Delhi)	Nov. 2022 – till date	To identify the cancer-causing factors by integrating epidemiological and genomic data using Artificial Intelligence Approaches	Department of Computer Engineering, Mizoram University, Mizoram, **India**
M.Tech Project	May 2019 - August 2020	To predict Early Gastric Cancer (EGC) factors from diet and lifestyle characteristics using supervised machine learning	Department of Computer Engineering, Mizoram University, Mizoram, **India**& Bioinformatics Infrastructural

		algorithms. To detect the cofounding risk factors causing MSI-GC by integrating genetic data and diet-lifestyle factors in Mizo population.	Facility, Mizoram University, **India**
Project Assistant	2 August 2009 - 18 May 2012	Developed Database for Mizoram butterflies and Snakes using client-server technology and JTool – PROT-PROP to predict subcellular location of a protein using JAVA. Predicted structure of Bt-cry class 1 proteins using homology modelling and validated and studies their toxicity using bioinformatics tools. Aiding students and scholars with the sequences analysis in validating, editing and submission of the sequences in NCBI database. Ground work had been done to design a model for protein classification using machine language algorithms (WEKA).	Bioinformatics Infrastructure Facility, Mizoram University, Aizawl, **India**
Visiting Faculty	May – June, 2009. March 2010. Feb-June, 2011.	**Taught** Basic of Computer programming, networking, operating systems, Databases and datamining concepts	Department of Biotechnology, Mizoram University, Mizoram, **India**
Visiting Faculty	Feb – July, 2009. Aug - Dec., 2009. April – June, 2011.	**Taught** Artificial Intelligence course	Department of Mathematics and Computer Science, Mizoram University, Mizoram, **India**
Visiting Faculty	July 2008 - 31 July 2009	**Taught** JAVA, Database Management Systems, Data Structures & Algorithms, Computer Fundamentals, Packages and Databases	ICFAI University Mizoram, Aizawl, **India**
Scholar(full time) Tutor(part ime)	July 2005 - Jan. 2007	Programming in JAVA, VB Assisting/Proof Reading computer science articles. **Taught** Java Programming	Institute of Computer Applications, Sun-yat sen University, **China** Sun-yat sen University, **China**

Publications

1. **Brindha Senthil Kumar**, Doris Z, Lalawampuii P, Saia C, John Z, Senthil kumar N, Lal Hmingliana (2022). Ensemble modelling for early breast cancer predication from diet and lifestyle. IFAC PapersOnLine 55-1, 429–435.
2. **Brindha Senthil Kumar**, Payel C, Saia C, John Z, Jeremy L P, Prem Nath, Arindam M, Senthil Kumar Nachimuthu, Lal Hmingliana (2022). Lifestyle and Dietary factors causing microsatellite instability gastric cancer detected using ensemble modelling. Journal of Scientific Research 14(3), 943-958.
3. **Brindha Senthil Kumar**, Vanlalwmpuia R, Freda R, John Z, Lalruatpuii H, Senthil Kumar N, Lal Hmingliana (2022). A Multilayer perceptron model to predict risk factors of type II diabetes mellitus. International Journal of Food and Nutritional Sciences. DOI: 10.4103/ijfans_110-22
4. **Brindha Senthil Kumar**, Chhuani L, Jahau L, Sarmah M, Senthil Kumar N, Vanlalpeka H, Lal Hmingliana (2022). Precise stratification of gastritis associated risk factors by handling outliers with feature selection in multilayer perceptron model. Lecture Notes in Electrical Engineering, vol 998. Springer.
5. **Brindha Senthil Kumar**, Senthil Kumar N, Ganesh, Rai, and Lal Hmingliana (2022). Machine learning to determine COVID-19 awareness and knowledge among the housemaids in Mizo population. J Evid Based Med Healthc 2022:9(8):2.
6. **Brindha Senthil Kumar**, Harvey V, John Z, Senthil Kumar N, Lal Hmingliana (2021). Logistic regression for gastric cancer classification using epidemiological risk factors in cases and controls. Science and Technology Journal, 9(2): 149-155.
7. **Brindha Senthil Kumar**, Senthilkumar N, Chenkual S, Lalruatfela S T, Zomuana T, Ralte Z, Maitra A, Basu A, Prem Nath (2020). Data Mining for Early Gastric Cancer Etiological Factors from Diet-Lifestyle Characteristics. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 65-73. Published in IEEE Xplorer. doi: 10.1109/ICICCS48265.2020.9121142.

8. **Brindha Senthil Kumar**, Z Ralte, AK Passari, VK Mishra, BM Chutia, BP Singh, Gurusubramanian G, Senthil Kumar N(2013) Characterization of Bacillus thruingiensis Cry1 class proteins in relation to their insecticidal action. Interdisciplinary Sciences: Computational Life Sciences 5(2): 127-135.
9. R Zomuanpuii, L Ringngheti, **Brindha Senthil Kumar**, Gurusubramanian G, Senthil Kumar N (2013) ITS2 characterization and Anopheles species identification of the subgenus Cellia**.** Acta tropica 125(3): 309-319.
10. **Brindha Senthil Kumar**, Sangzula S, Gurusubramanian G, Senthil Kumar N (2012) Prot-Prop: J-tool to predict the subcellular location of proteins based on physiochemical characterization**.** Interdisciplinary Sciences: Computational Life Sciences 4(4): 296-301.
11. Pranjal K, **Brindha Senthil Kumar,** Soundarajan K, Senthil kumar N(2012) Analysis of aminoacids pattern in receptor tyrosine kinase using Boolean Association Rule**.** Bioinformation 8(8): 344.
12. **Brindha Senthil Kumar** (2009) Mizoram Butterflies Data Storage-Retrieval Using Client-Server Technology: ICFAI University Journal of Computer Sciences. 2(1): 56-63.
13. H Tejmala, **Brindha Senthil Kumar**, G Gurusubramanian, N Senthilkumar (2011) In-silico comparison of distal-less protein variation in insects. Science vision 11(4):189-197.
14. **Brindha Senthil Kumar**, Sangzuala Sailo, Liansangmawii chhakchhuak, Pranjal Kalita, G Gurusubramanian, NS Kumar (2011) Protein 3D structure determination using homology modeling and structure analysis. Science Vision 11(3):125-133.
15. **Brindha Senthil Kumar** (2008) Era of Humo-machines: Revolution in Artificial Intelligence & Robotics. *Science Vision* 8(4):146 - 151.
16. Lalrotluanga, Issac L, Catherine V, Zothansanga, **Brindha Senthil Kumar**, Senthilkumar N, and Gurusubramanian G (2008) Butterfly Faunal Diversity in Aizawl, Mizoram, India. *Science Vision 8(3):* 65 – 75.

Invited Talks

1. Empowering Farmers with precise Machine Learning Models in 2-days national workshop training for AI-Driven Pest and Disease Monitoring for sustainable Agriculture in Mizoram, February 18, 2025.
2. The potential for Artificial Intelligence in Biomedical Sciences, September 19, 2024.
3. Online guest lecture on Machine Learning in Bioinformatics, December, 13, 2023.
4. Hands-on Artificial Intelligence in Cancer Genomics in Workshop and Hands-on on AI in Medicine, July, 12, 2023.
5. Trailblazing Artificial Intelligence to Classify Cancer Risk Factors from Diet and Lifestyle in National Seminar on applications of NGS and AI in healthcare, December, 20, 2022.
6. Machine Learning Models for Cancer Genomics in FDP on Reconnoitering Conventional & Avant-garde Methodologies to Excavate Safety Signals from Real World Data – A road map for integrated pharmacovigilance modus operandi in National Seminar on applications of NGS and AI in healthcare, September 9, 2022.

Conferences Attended

- One Day National Conference on Innovation in Biotechnology and Bioengineering (BIOKRITI 2024) on 27 April 2024, presented a paper entitled "Decision-Making Support System to predict the prognosis of Squamous cell carcinoma treatment", (**First prize**).

- International Conference on Millets: Genomic and Biotechnological Approaches: Impact on Health and Nutraceuticals (BIORUJIVITH 2024) on 27 March 2024, presented a paper entitled "Interpretable machine learning approach to predict the millets production in India", (**Third prize**).

- International Conference on Frontier Areas of Science & Technology (ICFAST 2022) organized by University of Hyderabad during September 9-10, 2022,

presented a paper entitled “Multi-class cancers classification using neural networks from non-invasive factors”.

- 4th International Conference on Machine Learning and Signal Processing (MISP 2022) organized by NIT, Raipur during March 12-14, 2022, presented a paper entitled “Precise stratification of gastritis associated risk factors by handling outliers with feature selection in multilayer perceptron model”.

- 7th International Conference on Advances in control & optimization of dynamical systems (ACDOS 2022) organized by NIT, Silchar during Feb. 22-25, 2022, presented a paper entitled “Ensemble Modelling for Early Breast Cancer Prediction from Diet and Lifestyle”.

- 4th International Conference on Recent Trends in Bioengineering (ICRTB-2021) during 12 – 13 February, 2021, abstract entitled “Lifestyle and Dietary factors causing Microsatellite Instability Gastric Cancer detected using Ensemble modeling”, (**First prize**).

- 2nd Annual Convention of the NEAST & International Seminar on Recent Advances in Science and Technology organized by Mizoram University during 16-18 November 2020, presented an abstract entitled “A Multilayer Perceptron based model to identify Type II Diabetes Mellitus from Epidemiological Data”.

- International Conference on Intelligent Computing and Control Systems (ICICCS 2020) organized by Vaigai College of Engineering during 13-15 May 2020, presented a paper entitled “Data mining for Early Gastric Cancer Etiological Factors from Diet-Lifestyle characteristic”.

- International Seminar on Computational Intelligence at Bharathiar University, Coimbatore on 30 December 2008.

Workshops/Training/Faculty Development Programs Attended

- Understanding AI Coding Assistants for Programming 15-16 Feburary 2025, by NPTEL, IIT Madras.
- Emerging Point-of-Care Diagnostic Technologies for Affordable Healthcare from Nov 25-30, 2024 at IIT Kharagpur.
- Building an Innovation Mindset for Student 14-16 November, 2020 by NPTEL, IIT Madras
- AI for signal processing on 6 July 2024 by NPTEL, IIT Madras.
- WEKA: Exploring Machine Learning for Healthcare during 6-7 April, 2024 by NPTEL, IIT Madras.
- Single Nucleotide Polymorphism (SNP) detection and analysis using NGS during 2-4, November, 2023 by Nano science and Technology Consortium.
- Micoarray based gene expression analysis using R programming during 6-8, October 2023 by Nano science and Technology Consortium.
- 3-day Offline Hands-on Workshop on Introduction to Machine Learning in Genomics during 21-23, Sep. 2023 by IBAB, Bangalore
- Applications of machine learning techniques in biology using Weka during 9-10 Sep. 2023 by NPTEL, IIT Madras
- 7 Days Training Programme on Machine Learning Using Python during 1-8 Feb, 2023
- High-End Workshop (DST-Karyashala) on Next Generation Sequence Data Analysis in Human Diseases during 17-22 October, 2022
- Hands-on training and workshop on computational biology and biomedical informatics during 25-31, July 2022.
- Online Elementary FDP on Artificial Intelligence and Machine Learning, during 04-08 October, 2021.
- One Day Faculty development program on Artificial Intelligence and Machine Learning in Biological Data Analysis, 29 March 2021.
- A 10 days workshop on Application of Artificial Intelligence in Bioinformatics and Biomedical Science, during 27 January to 06 February 2021

- Short term course on Research Methodology: Research & Publication Ethics, during 24-30 November 2020.
- One week AICTE Short Term Training Programme on Artificial Intelligence & Machine Learning, during 28 September to 03 October 2020

- A 10-days Live Online FDP / Short-Training on AI, ML & Deep Learning conducted by Eduxlabs, during 12 to 23 August 2020
- 7 Days Workshop on Hands-on Machine Learning using Python Programming, conducted by Department of Information Technology, Gauhati University, during 16 - 22 August 2020
- A 10-day workshop on Python for Everybody, conducted by Surana College, Peenya, Bangalore, during 9 to 19 July 2020
- A one week online short term course on LaTeX, conducted by Department of Information Technology, Mizoram University in associated with Spoken Tutorial Project, IIT Bombay, during 29 June to 3 July 2020.
- RBCDSAI'S International Summit on Data Science and AI, during 18-19 June 2020, organized by IIT Madras.
- Advanced workshop on Next Generation Sequencing and Data Analysis, organized by NIBMG, Kalyani, between 15-23 October 2019.

Webinars Attended

- One day national webinar on Emerging capabilities of Artificial Intelligence organized by Department of Computer Engineering and Computer Society of India (CSI), Kolkata, on 23 July 2021.
- Webinar on The Science of Learning, conducted by Dept of ECE and Dept of Physics, Mizoram University, on 25 May 2021.
- Webinar on Using Mathematical models to understand Epidemics, conducted by Department of Zoology, Mizoram University, on 19 May 2021.
- National webinar on Impact of Wearables and Internet of Things on Healthcare, conducted by Department of Information Technology, Mizoram University, on 15 July 2020

- National webinar on Software Development – Waterfall vs Agile Methodology, conducted by Department of Information Technology, Mizoram University, on 13 July 2020
- RBCDSAI's International Summit on Data Science and AI, organized by Robert Bosch Centre for Data Science and Artificial Intelligence, Indian Institute of Technology Madras, between 18 to 20 June 2020
- National webinar on Artificial Intelligence and Cloud Technology, conducted by Department of Information Technology, Mizoram University, on 28 May 2020

PARTICULARS OF THE CANDIDATE

NAME OF THE CANDIDATE	:	**BRINDHA SENTHIL KUMAR**
DEGREE	:	Ph.D.
DEPARTMENT	:	Computer Engineering
TITLE OF THE THESIS	:	Integration of Epidemiological and Cancer Genome Data using Artificial Intelligence Approach in Mizo Population
DATE OF ADMISSION	:	06.11.2020

APPROVAL OF RESEARCH PROPOSAL

1. DRC	:	12.04.2021
2. BOS	:	27.04.2021
3. SCHOOL BOARD	:	20.05.2021
MZU REGISTRATION NO.	:	1807416
Ph.D. REGISTRATION NO. & DATE	:	MZU/Ph. D./1617 of 06.11.2020
EXTENSION (IF ANY)	:	N/A

Head

Department of Computer Engineering

ABSTRACT

INTEGRATION OF EPIDEMIOLOGICAL AND CANCER GENOME DATA USING ARTIFICIAL INTELLIGENCE APPROACH IN MIZO POPULATION

AN ABSTRACT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

BRINDHA SENTHIL KUMAR

MZU REGISTRATION NO.: 1807416

Ph. D. REGISTRATION NO.: MZU/Ph. D./1617 of 06.11.2020

DEPARTMENT OF COMPUTER ENGINEERING

SCHOOL OF ENGINEERING AND TECHNOLOGY

MARCH,2025

INTEGRATION OF EPIDEMIOLOGICAL AND CANCER GENOME DATA USING ARTIFICIAL INTELLIGENCE APPROACH IN MIZO POPULATION

BY

BRINDHA SENTHIL KUMAR

Department of Computer Engineering

SUPERVISOR

Dr. LALHMINGLIANA

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Computer Engineering of Mizoram University, Aizawl.

INTRODUCTION AND REVIEW OF LITERATURE

Machine Learning (ML) models have demonstrated considerable potential in advancing cancer research and clinical practice across various types of cancer, including gastric, breast, and head and neck cancers. These models utilize extensive datasets of patient information, imaging studies, and molecular profiles to enhance diagnosis, prognosis, and treatment planning. In gastric cancer, ML algorithms have been employed to analyze endoscopic images, facilitating the detection of early-stage tumors and prediction of cancer progression (Khosraviet al. 2024). It also had provided a promising approach in medical diagnostics and treatment planning. These advanced computational techniques leverage large datasets of medical images, patient records, and genetic information to accurately categorize different types and stages of gastric cancer (Kumar et al. 2024). Machine learning algorithms, such as support vector machines, random forests, and gradient boosting, have demonstrated significant efficacy in identifying key features and patterns associated with various stages of gastric cancer (Wang et al. 2024). These algorithms can analyze complex combinations of clinical, histopathological, and molecular data to provide more precise and personalized classifications. Gastric cancer classification utilizing machine learning and deep learning (DL) algorithms has expanded beyond traditional methods, incorporating multi-modal data analysis and advanced neural network architectures.

The application of explainable AI techniques has improved the interpretability of these complex models, allowing clinicians to understand the reasoning behind the classifications and fostering trust in AI-assisted decision-making processes (Binzagr 2024). The data-driven approach not only enhances the precision of gastric cancer management but also facilitates more efficient clinical trials and accelerated drug discovery processes. As research in this field continues to progress, it is anticipated that deep learning algorithms will play an increasingly crucial role in clinical decision-making processes, potentially transforming the landscape of head and neck cancer management. In head and neck cancer, ML techniques have been applied to analyze CT and MRI scans, aiding in tumor

delineation for radiation therapy planning. Furthermore, across all three cancer types, ML models are being investigated to predict treatment outcomes, identify biomarkers, and personalize therapy regimens based on individual patient characteristics and genetic profiles. Machine learning models have been applied to various types of cancer, including gastric, breast, and head and neck cancers.

Understanding the causes, trends, and distribution of cancer in populations is crucial, and epidemiological risk factors are crucial for cancer prevention and management because they assist researchers, medical professionals, and public health officials in identifying high-risk individuals, putting in place efficient screening programs, and creating public health initiatives. These include inherited traits, exposures to the environment (like carcinogens), dietary and lifestyle choices (like diet and exercise), and infectious agents (Zodinpuii et al. 2022).However, epidemiological data alone is insufficient to identify the core cause of gastric cancer at the gene level. Genetic-level changes are primarily attributable to diet and lifestyle habits in various cancer types. Mutations can lead to cancer by disrupting normal cell functions and promoting uncontrollable cell growth. Certain genetic mutations in oncogenes have the probability to cause cancer. Mutations can also inactivate tumor suppressor genes, which prevent uncontrolled cell division and help repair DNA damage. Some genetic mutations are inherited from parents and can significantly increase the risk of certain cancers, this can help with early detection and prevention strategies. Most cancer-related mutations also occur in somatic cells and are not passed down to offspring (Chakraborty et al. 2023). These somatic mutations accumulate over time due to environmental factors (like smoking, UV radiation, or exposure to chemicals etc.), aging, or random errors during cell division. Understanding genetic mutations allows for personalized medicine, where treatments are tailored to an individual's specific genetic profile.

A support vector machine (SVM) model utilizing quantitative ultrasound scan features demonstrated 81% sensitivity and 76% specificity in distinguishing between responders and non-responding patients. This work used higher-order texture-of-texture features to improve predictive model performance (Safakish et al. 2023).

Random forest, neural network, elastic net penalized logistic regression, and support vector machine algorithms were employed to predict the primary site of head and neck squamous cell tumors based on their DNA methylation profiles. The classifiers demonstrated high classification accuracy, with the support vector machine achieving 89% accuracy (Leitheiser et al. 2021).Machine learning models were developed utilizing the Surveillance, Epidemiology, and End Results (SEER) registries database of head and neck squamous cell carcinoma. The Support Vector Machine (SVM) demonstrated an accuracy of 70%, while the decision tree exhibited an Area Under the Receiver Operating Characteristic (AUROC) of 72%, in comparison to logistic regression and random forest models (Rohatgi et al. 2023).

Several machine learning models have been developed for cancer classification using medical, laboratory, diet-lifestyle and medical image datasets.ML and DL models have been gaining significance in biological classification and regression to elucidate the underlying risk factors in complex diseases. Signs and symptoms of the same cancer type have been utilized to develop biomarkers (Saba 2020). Similar approaches could be designed at the population level to identify key mutations and novel mutations in a particular community. Furthermore, integrating WES and epidemiological data could facilitate more precise cancer therapy and cancer risk identification. These approaches may not be sufficient to determine the features responsible for causing breast cancer from genetic and diet-lifestyle perspectives.

In Mizoram, since 2003 the Incidence Rate of cancers in both male and female has been following an increasing trend. Stomach cancer is the leading cause of cancer death among male, followed by oesophagus, head & neck and lung. Among female, lung cancer is the leading cause of death followed by cervix, breast, oesophagus, head &neck. It is apparent that Mizoram is experiencing an alarming rise in cancer incidence and mortality rates, leading to the state being dubbed the cancer capital of India (NCDIR, 2014; Eric Zomawia et al. 2023). Earlier studies in Mizo population had shown both well-known and novel mutations in cancer-related genes such as *TP53*, *MUC6*, and *ARID1A*, APC, AKT1 and CDH1 in Gastric Cancer

(PayelChakraborty et al. 2023) and *TP53, CACNA1E, IGSF3,RYR1,CDH1*and *FAM155A* in somatic samples and *ATM, STK11*, *CDKN2A* in germline Triple Negative Breast Cancer samples (Pachuau et al. 2025).

Consumption of smoked and salted meat, “sa-um” (fermented pork fat), fish and soda as a food additive have been shown to increase the risk of Stomach Cancer in Mizoram (Yadav et al. 2018) and high intake of animal fats, mainly from red meat was correlated with an increased risk of breast cancer in Mizo population (Thapa et al. 2016). There is a need of cancer research in integrated aspects including epidemiology, etiology, lifestyle factors and food habits, genetic mutations screening and ML methods for proper diagnosis and prognosis.

The study focuses on the high-risk Mizo population in Northeast India, addressing a critical knowledge gap in cancer research for this specific community. By analyzing point mutations in relation to etiological factors, the research aims to develop more accurate and efficient methods for early cancer detection. This approach not only promises to improve cancer diagnosis but also offers a non-invasive, cost-effective, and rapid alternative to traditional biopsy procedures. The integration of community-specific data will contribute to a better understanding of the cancer burden in the Mizo population and facilitate the extraction of clinical knowledge tailored to their unique genetic and environmental context. Ultimately, this research has the potential to significantly impact cancer prevention and treatment strategies, particularly in resource-limited settings. These models aim to enhance diagnostic accuracy, risk stratification, and treatment planning for cancers by utilizing the complementary information provided by epidemiological and genomic data.

The present study is a seminal work that integrates epidemiological and genomic data to elucidate the etiological factors for human cancer types. Epidemiological and genomic data of various cancer types are available, enabling the integration of these two feature sets. This research aims to address the knowledge gap between epidemiological and genomic data by enabling early detection of cancer through the identification of significant point mutations in relation to the etiological

factors that cause cancer in the high-risk Mizo population of Northeast India. This study will also contribute to the aggregation of data from multiple sources specific to the Mizo population, which will greatly aid in quantifying the burden of this disease and extracting clinical knowledge specific to the community. Although biopsy is considered the gold standard for cancer diagnosis, it is an invasive, costly, and time-consuming procedure, particularly in remote areas. The present methods will provide non-invasive, time-efficient, and rapid approaches for accurately predicting the disease and its causes at both genetic and diet-lifestyle levels.

The present study also represents a significant advancement in cancer research by integrating epidemiological and genomic data to identify etiological factors for various human cancer types. This integration of diverse datasets allows for a more comprehensive understanding of cancer development and progression. By combining population-level epidemiological data with genetic information, researchers can uncover complex relationships between environmental factors, lifestyle choices, and genetic predispositions that contribute to cancer risk. This approach has the potential to revolutionize cancer prevention and early detection strategies by providing a more nuanced understanding of cancer etiology. This information can be used to implement targeted screening programs, develop personalized prevention strategies, and guide treatment decisions for patients with hereditary cancer syndromes.

OBJECTIVES

1) Identification of etiological risk factors from cancer epidemiological data by constructing effective **Machine Learning models**.

2) Detection of disease causing **mutations from genomic variants** between cancer types and healthy controls using Machine Learning models.

3) To find the association between **epidemiological factors and genomic traits** by integrating epidemiological and genomic data.

METHODS

Epidemiological Data of Gastric Cancer patients and Study Population

The epidemiological data for gastric, breast and head & neck were obtained from Department of Biotechnology, Mizoram University which was collected from Civil Hospital Aizawl and Mizoram State Cancer Institute, Mizoram, Northeast India. The data utilized in this investigation were collected after obtaining informed consent from cancer patients and healthy individuals.

Genomic Variant Data

The germline variant dataset was obtained from Mizoram cancer patients in Variant Call Format (VCF) for gastric, breast, and head and neck cancers from Department of Biotechnology, Mizoram University. This study had utilized germline variants of afore-mentioned cancer types, as germline variants play a crucial role in understanding cancer predisposition and inheritance patterns(Cobaleda et al. 2024). The DNA isolated from the patient's blood was subjected to Whole Exome Sequencing (100× coverage) in IlluminaNovaSeq 6000 and the quality was checked using FastQC v0.12.1, adapter sequences removed using Trimmomatic v0.39 (Andrews, 2010; Bolger et al., 2014), aligned to GRCh38.p13 reference genome using the BWA-MEM algorithm followed by sorting the aligned reads, marking PCR duplicates, base quality score recalibration (dbSNP138.vcf), variant calling (GATK v4.2.0.0 HaplotypeCaller) and variant quality score recalibration (VQSR) (Li, 2013; Van der Auwera and O'Connor, 2020). The variants were annotated in ANNOVAR using RefGen, ClinVar, 1000 genome, avsnp15, esp3500, gnomad211_exome databases.

The annotated VCF files contained minor allele frequencies of all geographical populations, functional predictions for each single nucleotide polymorphism (SNP): Combined Annotation Dependent Depletion (CADD) (Rentzsch et al. 2019), Polymorphism Phenotyping v2 (PolyPhen-2) (Adzhubei et al. 2013), Functional Analysis through Hidden Markova Models (FATHMM) (Shihab et al. 2012), Variant Effect Scoring Tool (VEST) (Douville, et al. 2016), Sorting

Intolerant From Tolerant (SIFT) (Sim et al. 2012), Likelihood Ratio Test (LRT) (Zeng et. al. 2014), MutationAssessor (Reva et al. 2011), MutationTaster, RadialSVM, and several hand-crafted features (Heterogeneous SNV, homogenous SNV, Amino acid alteration_SNV, exon_splicing SNV, startloss_SNV, stoploss_SN), chromosome number, gene names, reference base, altered base, and positions.

Classification of Gastric Cancer using Epidemiological Features

The dataset comprises 117 gastric cancer patients and 148 healthy controls from the Mizo population. Twenty-nine features were included, encompassing information about dietary habits, lifestyle factors, environmental exposures, and health history. The dataset exhibited missing values, which were subsequently imputed utilizing their respective class median values, as the features were discrete (Emmanuel et al. 2021).There was substantial bias in smoking, alcohol, and other smokeless tobacco product intake between cases and controls, as the control samples included a higher number of female participants, with the majority falling between 23 to 35 years old.All features in the dataset were normalized using MinMax scaling.

A logistic regression model was developed to classify healthy individuals and gastric cancer patients based on epidemiological features. The model predicted the outcome of the target variable based on a threshold value. If the outcome exceeded the threshold, the observation was classified as cancer; and if the outcome was below the threshold, the observation was classified as healthy. In this study, the threshold was set at 0.5 for the logistic regression to differentiate between gastric cancer patients and healthy individuals.

Classification of Gastric Cancer using Genomic Features

Gastric Cancer Big Dataset

A total of 192,781 variants were identified. During the quality control process, variants with read depth below the mean value were excluded, and a mean coverage of 93% was maintained. Following quality control measures, 72,471 benign

variants and 8,788 pathogenic variants remained. Pathogenic variants were assigned a label of 1, while benign variants were assigned a label of 0, serving as the class label (dependent variable). The class labels were assigned only if the variant was predicted to be benign or pathogenic by all six prediction tools: MutationTaster, FATHMM_pred, LR_pred, RadialSVM, SIFT_pred, and Polyphen2_HDIV. To minimize noise, avoid biasness in model's performanceand improve the model's generalization to distinguish between cancerous (pathogenic) and non-cancerous (benign) SNPs, columns containing missing values were removed, resulting in 36 features being subjected to data visualization and modeling.

The features exhibited varying ranges; consequently, the training dataset was scaled using standard scalar. The present study utilized LASSO regression with 5-fold cross-validation to identify the key features contributing to variant pathogenicity. Significant 16 features that contribute to the target value, while the coefficients of other irrelevant features were reduced to zero. Given that the present dataset was large and highly imbalanced, in order to address noise, decision uncertainty, and reduce computational complexity, this study adopted L1-penalty-based feature selection (LASSO) to extract optimal features (Yu et al., 2021).

Imbalanced Dataset

On utilizing the top 15 genetic traits selected by 5-fold LASSOCV, an imbalanced dataset was constructed with the size and features.

Random undersampling(Balanced)

The dataset presented exhibited significant imbalance, which may result in inadequate learning of the positive minority class. To mitigate poor performance on the minority class and prevent overfitting, a balanced dataset was randomly subsampled (undersampling) and utilized to assess the viability of the features selected by the 5-fold LASSOCV method.

Ensemble Models

The dataset was split into 70% training and 30% for testing. The train dataset was utilized to identify optimal hyperparameters using a five-fold GridSearchCV procedure. A set of five ensemble classifiers were selected: Gradient boost, AdaBoost, Random forest, Bagging, and extra trees algorithms. To predict pathogenic and benign variants from gastric cancer big data, five ensemble models were selected: Gradient boost, AdaBoost, Random forest, Bagging, and extra trees classifiers. The models were optimized with hyperparameters obtained through 5-fold cross-validation grid search method on the training datasets using 2 feature subsets.

Classification of MSI Gastric Cancer using Genomic and Epidemiological Features

The study included 79 gastric cancer patients who had been diagnosed with the disease through biopsy and endoscopy. The dataset had eleven major risk factors associated with gastric cancer. The dataset comprised eleven attributes, encompassing both diet and lifestyle data. Data labeling was performed using attributes including: smoked food, alcohol,Khaini/sadha, khuva,tuibur, extra salt, saum, smoking, age, sex, and target (MSI - 1 or non MSI - 0). Learning curves were generated to address underfitting, and feature selection was performed to mitigate overfitting (Viering and Loog, 2021). Learning curves were generated for Logistic Regression, Naive Bayes, Multilayer Perceptron, Support Vector Machine, and Random Forest. Feature selection and redundant data elimination were conducted using three established methods: ridge regression (Friedman et al. 2010), extra trees classifier (Hemphill et al. 2014), and recursive feature elimination methods (Guyonet al. 2002). In thisstudy, lifestyle and food habits were evaluated for their risk for MSI-GC. The lifestyle-diet dataset was randomly divided in a 70:30 ratio; 95 instances allocated as training and 47 instances as testing set, with random state=0. Accuracy, precision, recall, F1 score, brier score, MCC, and Receiver operating characteristics (ROC) (Brier, 1950; Tilaki, 2013; Chicco and Jurman, 2020) were significant metrics to evaluate the performance of the models.

Classification of Breast Cancer using Epidemiological Features

The dataset consisted of 15 independent and 1 dependent features, with a total of 321 data points. The dataset had 7% missing value which may be attributed to patients' compromised physical condition during therapy. Missing values were replaced with their respective group median values after applying one-hot encoding (Santhoshkumar et al. 2020). Five ensemble learning classifiers were selected and the performance was validated using leave one out cross validation (LOOCV) and 10-fold cross validation (10-fold CV).

Classification of Breast Cancer using Genomic Features

Breast Cancer Big Dataset

Initially, 203,380 variants were identified in breast cancer samples. After quality control and removal of missing values, 1,213 variants were filtered out. In healthy individuals, 4,407,672 variants obtained from the source, in which 755 variants retained following the removal of missing values and quality control in terms of depth and coverage. All variants from breast cancer samples were labeled as 1, while variants from healthy individuals were labeled as 0.

Feature Selection by Random Forest

Random Forest with hyperparameter tuning was employed to identify the key features contributing to variant pathogenicity. This method combines multiple decision trees to reduce variance. The inherent randomness of the model facilitates the evaluation of feature importance and is extensively utilized in datasets with high-dimensional features (Chen et al., 2020).

Feature Subsets

Feature selection in combination with oversampling techniques offers several advantages when generating feature subsets for machine learning tasks (Tsai et al. 2024). This approach helps to reduce dimensionality, mitigate the curse of dimensionality, and improve model performance by focusing on the most

discriminative attributes. Ten significant features from the raw dataset (imbalanced), ADASYN dataset, SMOTEENN dataset, and blsmote, were selected to form four independent datasets.

Convolution Neural Networks Model for Breast Cancer Big Data Classification

The hyperparameters of CNN models werebatch size of 32, 300 epochs, ReLU activation function for input and hidden layers, sigmoid activation function for the output layer, binary cross-entropy loss function, and Adam optimizer. These hyperparameters were optimized across all four CNN models developed using the raw dataset (imbalanced), ADASYN, SMOTEENN, and BLSMOTE datasets. The dataset was partitioned as train (60%), valid (20%) and test (20%) to evaluate the performance of the CNN models. The CNN models successfully achieved global minima, as evidenced by the smooth and consistent training accuracy, validation accuracy, training loss, and validation loss graphs, which were plotted for the raw dataset (imbalanced), ADASYN, SMOTEENN, and BLSMOTE datasets.

Classification of Head & Neck Cancer using Epidemiological Features

The dataset consisted of 14 independent and 1 dependent features, with a total of 300 data points. There were 200 instances for healthy individuals (controls) and 100 instances for head and neck cancer (cases). The categorical variables were converted to numerical values using label encoder. The dataset was divided into train and test set with ratio of 3:1, respectively. Further, train set was scaled using StandardScaler and transformed on test set. The missing values and outliers were removed to enhance data integrity and reduce the negative implication on patient care.

Classification of Head and Neck Cancer using Genomic Features

Head and Neck Cancer Big Dataset

In total, there were 5,141,996 variants identified in head and neck cancer samples. Of these, only 60,824 variants were located in the region of interest (exonic). Following quality control measures and removal of missing values, 5,784 variants remained. In

healthy individuals, 4,407,672 variants were initially identified, with 755 variants retained after quality control procedures and removal of missing values, considering depth and coverage. All variants from head and neck cancer samples were labeled as 1, while variants from healthy individuals were labeled as 0.

Random Downsampling

Random downsampling in machine learning is a technique that can address class imbalance issues by reducing the number of samples in the majority class.

Data Preprocessing

The dataset was divided into training and test sets. The training dataset was scaled using MinMaxscaler. Missing values in attributes and instances were removed to mitigate potential bias in data modeling.

Feature selection using Extra Trees

The Extra Trees classifier, optimized through hyperparameter tuning, was employed to identify the key features contributing to variant pathogenicity.The algorithm's hyperparameters were fine-tuned to optimize its performance, ensuring the most effective identification of key features within the dataset.

Data Models

Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and logistic regression classifiers were employed to develop models utilizing a large head and neck cancer dataset. The hyperparameters for these models were optimized using the training set, and their efficacy was subsequently evaluated using the reserved test dataset.

Multi-class classification by integrating Gastric, Breast and Head & Neck Cancer features

Data oversampling

The dataset comprised five independent features and one dependent feature, with a total of 590 observations across all cancer types (gastric, breast, and head and neck cancer) and healthy individuals. The dataset contained 200, 173, 117, and 100 observations from healthy individuals, breast cancer patients, gastric cancer patients, and head and neck cancer patients, respectively. The dataset exhibits uneven class distributions; consequently, Synthetic Minority Over-sampling Technique (SMOTE) was employed to oversample gastric, breast, and head & neck cancer data.

Feature Selection using Mutual Info Classifier

The integrated dataset was subjected to feature extraction using mutual info classifier. Feature extraction by mutual information (MI) classifier has a robust methodology as it computes the mutual information between each feature and the target variable (Hoque et al. 2014). This method effectively ranks and selects features that are most informative for classification tasks.

Multilayer Perceptron Model

Multilayer perceptrons (MLPs) offer several advantages for multiclass classification tasks. Their capacity to learn complex, non-linear decision boundaries renders them well-suited for handling datasets with intricate relationships between features and multiple classes (El-Hassani et al. 2024). This part of the work had presented an optimized multilayer perceptron model that was specific to this integrated dataset.

RESULTS

Gastric Cancer Classification from Epidemiological Data

The proposed logistic regression model demonstrated a promising accuracy of 98.5% by achieving the global minimum with an intercept of 0.22 and a learning rate of 0.000915. Saum (fermented pork fat), the use of aluminum cookware, plastic

utensils for food storage, smoking of Zozial (infiltrated local cigar), smoking before the age of 18,passive smoking, sadha (smokeless tobacco product), tuibur (aqueous tobacco extract), chewing betel nuts (kuhva, zardhapan, supari), residing or working in proximity to cell phone towers, and family history of cancer were identified as epidemiological risk factors for gastric cancer in the Mizoram population.

Gastric Cancer Classification from Big Data

LASSO model was experimentally tuned based on 5-fold CV with hyperparameters: max iteration was 100000 and tolerance was 0.000001, which had found fifteen significant features for causing pathogenicity in gastric cancer. AAChanges, CADD_raw, non_neuro_AF_popmax, ExAC_AMR, Ref, Genes, ExAC_OTH, ExAC_SAS, and Start were found to be negatively correlated; whereas CADD_phred, VEST3_score, phyloP46way_placental, AF_asj, AF_eas, and Alt were positively correlated.

Models

The Extra Trees classifier applied to the balanced dataset demonstrated the highest and most effective performance across all metrics, with 96.06% accuracy, 95.74% precision, 96.33% recall, 96.0F1-score 4, ROC_AUC of 99%, MCC of 0.92, and Brier score of 0.04. Its performance on the imbalanced dataset exhibited the highest accuracy at 97.57% and lowest Brier score at 0.02. It was also observed that there were marginally uneven values across all metrics: precision 95.62%, recall 90%, F1-score 92.73%, PR-AUC 98%, and MCC of 0.91. These results indicated a more robust generalization capability for both positive and negative classes on the balanced dataset.Random forest had become the second-best performing model on a balanced dataset showed accuracy 94.39%, precision 93.67%, recall 95.11%, F1-score 94.39%, ROC-AUC 98%, MCC 0.89, and brier score 0.06. This model could better identify most pathogenic variants. When comparing these values with the imbalanced dataset metrics, there were significant decreases in recall (84.71%) and F1-score (88.75%). Although the accuracy of the imbalanced dataset was 96.31%,

which was found to be superior to that of the balanced dataset, this parameter alone cannot be considered significant in terms of assessing the model's performance.

On the balanced dataset, it achieved an accuracy of 93.94%, precision of 93.62%, recall of 94.19%, F1-score of 93.90%, ROC-AUC of 98%, MCC of 0.88, and Brier score of 0.06. The classifier on the imbalanced dataset demonstrated slightly higher margins in evaluation metrics, with an accuracy of 97.06%, which was comparable to the extra trees accuracy on the imbalanced dataset. Precision, recall, F1-score, and ROC exhibited a minor decrease of 93.52%, 89.12%, 91.27%, and 96.1%, respectively. AdaBoost and Gradient boosting algorithms demonstrated moderate performance in determining the pathogenicity of the variants. Precision and recall scores on the balanced dataset were notably more favorable compared to values obtained from imbalanced datasets. Gradient boost on the imbalanced dataset yielded a precision of 86.12% and recall of 71.18%, whereas on the balanced dataset, it achieved a precision of 90.30% and recall of 91.13%. Conversely, AdaBoost with the imbalanced dataset produced a precision of 83.46% and recall of 66.76%, while performing more effectively on the balanced dataset with a precision of 88.33% and recall of 85.63%.In the aforementioned results of bagging classifiers and boosting classifiers, all models demonstrated satisfactory performance on the balanced dataset; consequently, accuracy, MCC, and Brier score potentially can led to misinterpretation on the imbalanced dataset.

MSI Gastric Cancer Classification using Genomic and Epidemiological Features

All eleven features were incorporated to develop learning curves for the machine learning models: support vector machine, random forest, Naïve Bayes, logistic regression, and Multilayer Perceptron. The learning curves showed exclusion of SVM due to huge error gap between the training and validation curves.To identify the significant features contributing to the MSI-GC, three methods were employed: Ridge regression, extra trees, and recursive elimination methods.The features selected by ridge regression were extra salt, khuva, alcohol, smokeless tobacco product (Khaini / Sadha), and smoked food.The extra trees method selected features

including Saum, alcohol, khuva, smoked food, and extra salt. Extra salt, smokeless tobacco products (Khaini / Sadha and Tuibur), alcohol, smoking, and khuva were key features extracted by the recursive elimination method.

Machine Learning models for MSI-GC

Random forest and multilayer perceptron classifier demonstrated efficacy in classifying positive and negative MSI-GC. These two models were developed on the three subsets identified by ridge regression, recursive elimination method, and extra trees classifier. Random forest and multilayer perceptron models exhibited efficient and consistent results with the subset identified by ridge regression, achieving 96% in accuracy, recall, precision, and F1-score, as well as the highest MCC of 0.91.

Identification of Confounding and Driving Factors

Multiple-linear regression was employed on the covariates identified by ridge regression to determine the driving and confounding factors for MSI-GC. Among all the features, khuva exhibited the lowest error and a statistically significant p-value of 0.13 and 0.0007, respectively. The regression coefficient of -0.20 indicated that MSI-GC (dependent variable) was dependent on khuva (independent variable), which was evidently a driving factor in the etiology of the disease. Excessive salt consumption, smoked food intake, Khaini / Sadha use, and alcohol consumption were also identified as driving factors. The distribution of khuva consumption data demonstrated a notable distinction between the positive and negative classes of MSI-GC.

Breast Cancer Classification from Epidemiological Data

Ensemble Models

Five ensemble models were developed, and their performances were evaluated using leave-one-out cross-validation (LOOCV) and 10-fold cross-validation. The Extra Trees classifier achieved an accuracy of 96% with both validation techniques. Gradient boosting, random forest, and AdaBoost classifiers yielded an accuracy of 95% using LOOCV and 10-fold CV. The Bagging classifier attained an accuracy of

94%. The Pearson correlation coefficient between the accuracies obtained by LOOCV and 10-fold CV was 0.97

Breast Cancer Classification from Germline Data

Convolution Neural Network Models

The performance of CNN models on the four datasets demonstrated superior results in terms of accuracy, precision, recall, f1-score, and ROC curves. The highest accuracy achieved was 98.98% (original dataset) and 98.68% (SMOTEENN dataset). The highest precision was 98.79 (original dataset) and 99.32 (SMOTEENN dataset). The recall and f1-score values were highly satisfactory using both the original and SMOTEENN datasets. Although accuracy, precision, recall, and f1-scores were notably promising using the raw dataset, the ROC curve performance exhibited a slight decrease when compared to the balanced dataset – SMOTEENN.Based on the model evaluation parameters, it was noteworthy that the feature subset identified by SMOTEENN could be the critical features for identifying pathogenic variants from the breast cancer germline dataset.

Head and Neck Cancer Classification from Epidemiological Data

Data Models

Five machine learning classifiers were developed, and their performances were evaluated using accuracy, precision, recall and f1_score. The classification accuracies and f1_scores showed a consistent classifier. The Extra Trees classifier achieved an accuracy of 85%, precision 93%, recall 67%, and f1_score 78%, this was followed by random forest with f1_score of 73% and accuracy of 82%. Linear SVM, decision tree, and Naïve Bayes had metrics of accuracy, and precision above 70%.

Head & Neck Cancer Classification from Germline Data

Data models

Heat map had eliminated all the allele frequencies from the different population as they had multicollinearity between them .The results obtained by random forest, decision tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and logistic regression were precisely optimized by setting appropriate hyperparameters. The accuracy of decision tree and SVM exceeded that of random forest, KNN, and logistic regression, reaching 96%. These two algorithms demonstrated very low false positive rates of 1. The results of decision tree and SVM exhibited strong support for the classification results, with each metric of these algorithms being identical. False negatives were considerably higher in logistic regression at 20, although it maintained a low false positive rate of 1.

Random forest achieved an accuracy of 95% but exhibited a slightly increased false positive rate of 4 and false negative rate of 15 compared to decision tree and SVM. The accuracy of KNN was 90%, whereas it demonstrated high false positive and false negative rates at 8 and 31, respectively. Diagnostic performance of binary classification was measured by Receiver Operating Characteristic (ROC) curves. Among the five classifiers, random forest exhibited the highest Area Under the Curve (AUC) of 99%; however, the model displayed marginally high false positive instances. Consequently, the ROC-AUC of decision tree and SVM were 98%, which demonstrated excellent overall performance in classifying the healthy and head and neck cancer variants from the annotated genomic dataset.

Integration of Healthy, Gastric, Breast and Head & Neck Cancer Classification

Proposed Multilayer Perceptron Model

The accuracy of the multilayer perceptron model attained 75% in this multi-class classification. The model was constructed with two hidden layers comprising 16 and 8 neurons each, which achieved satisfactory accuracy while utilizing reduced computational resources and time. Other architectures of multilayer perceptron

models demonstrated comparable performance; however, in terms of time and computational efficiency, the architecture with 16 and 8 neuron hidden layers yielded promising results.

Performance of Ensemble Data Models on Big Data

There were 35.7% of unique variations in the big dataset of gastric, breast, and head and neck cancers. Approximately, 56.9% of variations were found to be common among all three tumors, while only 7.4% were found to be prevalent in any of the two cancer types. Genes USH2A, MK167, NEB, and OBSCN were found common and highly mutated in the above-mentioned cancers. DHAH5, PARP4 and PRUNE2 were in top 10 mutated genes speficically in gastric, breast and head & neck cancer, respectively. FAT1, SYNE1 and SYNE2 were found common between gastric and head & neck cancers; while, none found common between gastric and breast; breast and head & neck cancers. An interesting result was obtained on integerting gastric, breast and head & neck germline characteristic which had revealed more unique genes among these three cancers. MDN1 in GC; XIRP2 and FBN3 in BC; were predominantly mutated in this dataset. AdaBoost algorithms achieved the maximum accuracy of 84.18%, precision of 82.15%, recall of 86.83%, f1_score of 84.43%, and ROC_AUC of 84.21% when classifying pathogenicity within the three cancer kinds (gastric, breast, and head and neck). When compared to other classifiers, the metrics obtained by AdaBoost revealed a well-balanced classifier that identified benign and pathogenic variations while maintaining a balance of precision and recall.

DISCUSSION

Classification of Gastric Cancer using Epidemiological Features

This section evaluated the logistic regression model's effectiveness, considering all risk factors as independent variables. With a 0.5 threshold, the model achieved 98.51% accuracy.. Global minima were reached after 150,000 iterations, with a 0.000915 learning rate and 0.22 line intercept. When iterations were reduced to 50,000, maintaining the same learning rate but with a -0.04 intercept, accuracy decreased to 97.01%.

The current study's logistic regression model revealed strong associations between all features and gastric cancer incidence in Mizo populations. Many of these features align with globally recognized risk factors for gastric cancer. Excessive salt intake, processed meats, alcohol consumption, and limited physical activity are among the risk factors for gastric cancer in Asian populations (Karagianni et al. 2010).

Classification of Gastric Cancer using Genomic Features

Annotated features were employed to construct ensemble models. The models exhibited enhanced performance with high data fidelity, achieved by labeling target variants only when all six prediction tools (MutationTaster, FATHMM_pred, LR_pred, RadialSVM, SIFT_pred, and Polyphen2_HDIV) concurred on their benign/pathogenic status. The dataset was refined from 192781 to 72471 variants after eliminating redundancies. The ensemble models were trained on 70% of the non-redundant variants and tested on the remaining 30%.

The models were trained using optimal hyperparameters obtained through 5-fold gridsearchCV and validated using PR-AUC for imbalanced datasets and ROC-AUC for balanced datasets. For the imbalanced dataset, three key validation metrics - PR-AUC, recall, and MCC - were considered as gold standards for performance evaluation (Hancock et al. 2023). The optimal classifiers were selected based on minimum brier scores and maximum recall, precision, and PR-AUC, as false positives could lead to erroneous drug target identification (Kim, et al. 2020). To address multicollinearity and prevent model overfitting, non-collinear feature subsets were extracted for both training and testing sets.

The study also examined the impact of LASSOCV-selected features on the ensemble models. LASSOCV regression's unique ability to tune λ during cross-validation and handle multicollinearity without data loss was emphasized. The most influential features for determining variant pathogenicity were identified as CADD_phred, VEST3_score, phyloP46way_placental, AF_asj, AF_eas, Alt, Start, ExAC_SAS, ExAC_AMR, Ref, Genes, ExAC_OTH, CADD_raw,

non_neuro_AF_popmax, and AAChanges. For the balanced dataset, four out of five models demonstrated the highest ROC-AUC, while for the imbalanced dataset, three out of five models exhibited the highest PR-AUC. Both scenarios yielded high MCC scores exceeding 0.80. To validate the consensus classification using the top 15 significant features, the study employed both imbalanced and balanced (undersampled) datasets. These findings suggest that the identified features play a crucial role in determining pathogenicity in gastric cancer big data (Noroozi et al. 2023).

In accordance with these features, the extra trees classifier presented here demonstrated relatively higher performance over time, achieving 96% accuracy, 95.74% precision, and 96.33% recall with a low error rate of 0.04 (brier score) on a balanced dataset. This study also exhibited notable performance using extra trees classifiers on an imbalanced dataset, with a PR-AUC of 98% and the lowest error rate of 0.02 due to low false positives. A recent report indicates that in silico tools for predicting pathogenicity continue to evolve through regular bug fixes and improvements, suggesting that CADD and VEST3 scores can be considered key factors in determining pathogenicity and identifying new candidate genes/variants (Gracia et al. 2022). Allele frequencies from various populations (African/African American, East Asian, South Asian, and Exome Aggregation Consortium database) were among the top 10 selected features in this study. Previous research has demonstrated that allele frequencies are useful in detecting genetic misdiagnosis and identifying benign variants (Kobayashi at al. 2017).

The 15 key features underwent training and testing on both imbalanced and balanced datasets, with the latter created through random undersampling. Additionally, these identified features significantly reduce computational time during variant annotation, potentially expediting diagnosis and treatment outcomes. These genetic characteristics can be combined with epidemiological factors to efficiently predict disease outcomes and prognosis rates. Extra trees classifiers demonstrated superior performance in terms of accuracy, precision, recall, and f1_score across both balanced and imbalanced datasets. Several crucial steps were implemented to

prevent erroneous results and misinterpretation: (i) determining optimal tuning hyperparameters using 5-fold GridSearchCV; (ii) systematically reducing horizontal dimensionality through a valid feature selection method by fine-tuning LASSOCV; and (iii) labeling targets only when all 6 functional prediction tools agreed, ensuring a balance between precision and recall (minimizing false positives and negatives).

Classification of MSI Gastric Cancer using Genomic and Epidemiological Features

To assess model fit, learning curves were generated for five supervised ML algorithms (support vector machine, random forest, logistic regression, multilayer perceptron, and Naive Bayes) utilizing all eleven attributes. The curves indicated that all algorithms, with the exception of support vector machine, were well-suited for classifying the dataset. The support vector machine exhibited a wider disparity between training and validation scores, while the other four algorithms demonstrated smooth convergence with reduced error gaps between training and validation score curves. To eliminate irrelevant attributes, identify strongly correlated features, and mitigate overfitting, feature selection was conducted using 3 robust methods: recursive elimination, ridge regression, and extra trees classifier. Ridge regression identified Khaini/sadha, khuva, extra salt, smoked food, alcohol as the top 5 potential factors contributing to MSI-GC. The recursive elimination technique identified tuibur, extra salt, alcohol, Khaini/sadha, smoking as the 5 most significant features. The extra trees model determined extra salt, saum, smoked food, alcohol, and khuva as the top five correlated features. Data models were constructed using the four selected algorithms (random forest, multilayer perceptron, logistic regression, and Naive Bayes) on the three groups of feature sets identified by the extraction methods. The feature selection process and learning curve analysis confirmed the absence of underfitting and overfitting in the dataset. These crucial data preprocessing steps were implemented to enhance accuracy and minimize misclassification of datapoints prior to model development.

Features selected through ridge regression, including khuva,extra salt, smoked food, alcohol, Khaini/sadha were utilized to construct Logistic Regression,

Random Forest, Multilayer Perceptron, and Naive Bayes classifiers. The Random Forest and Multilayer Perceptron models were fine-tuned using hyperparameters. Hyperparameter optimization was a crucial step in model training, significantly enhancing performance during the testing phase (DeCastro-García et al. 2019). For the Random Forest model, 100 decision trees were selected, based on research by Oshiro et al., which demonstrated that doubling the number of trees beyond 64-128 did not yield additional information gain, with optimal performance (100% AUC) achieved within this range (Oshiro et al. 2012).

A multiple linear regression analysis incorporating various covariates was employed to determine confounding factors (McNamee, 2005). The analysis revealed that excessive salt consumption, intake of smoked meat and vegs, habit of chewing khaini/sadha, and drinking alcohol (independent variables) were associated with MSI-GC (dependent variable), indicating as possible confounding factors. Khuva emerged as the primary contributor to MSI-GC, demonstrating a statistically significant p-value of 0.0007 and a slope of -0.2142, which indicates a strong negative correlation between Khuva and MSI-GC compared to other factors. The regression coefficient of -0.20 indicated that MSI-GC was highly dependent on Khuva, and the mean square error of 0.13 was the lowest among all attributes, suggesting minimal discrepancy between actual and predicted values. Previous studies have demonstrated that betel quid (Khuva) chewers are susceptible to MSI in head and neck cancer patients (Lin et al. 2017), and it is also a major factor in causing multiple MSI regions in oral cancer (Su et al. 2019). This study provides the first evidence that smoking tobacco, excessive salt intake, smoked food consumption, and alcohol use are confounding factors, while Khuva is the primary driver of MSI in gastric cancer.

Classification of Breast Cancer using Epidemiological Features

This research employed 5 ensemble classifiers to forecast breast cancer in women from the Mizo population using epidemiological characteristics. Individuals with detrimental habits, such as addiction to various forms of tobacco, smoked foods, and alcohol, are more susceptible to developing different types of cancer. Among the

obtained accuracies, the extra trees algorithm surpassed AdaBoost, random forest, and gradient boosting models, achieving mean LOOCV accuracies of 96.3% and 10-fold CV of 95.5%. Due to the imperfect predictive accuracy of many new machine learning classifiers, this study utilized both LOOCV and 10-fold CV to verify prediction accuracy (Fatima et al. 2020). Furthermore, a correlation score of 97.45% between the mean accuracies generated by these two cross-validation methods reinforced the models' prediction accuracy and result replicability (Hosni et al. 2017).

Ensemble models have proven highly effective to enhance classification accuracy, with boosting techniques (AdaBoost algorithm) which had been widely used in various cancer classifications. However, for this dataset, bagging techniques (Random Forest and Extra Trees classifiers) performed exceptionally well (Vaka et al. 2020). Notably, the models generated in this study achieved accuracies $\geq$ 94% using non-invasive breast cancer features through bagging and boosting, surpassing the accuracy score of the XGboost classifier applied to Chinese women Hou et al. (2020). Research indicates that prolonged adherence to detrimental habits, such as using smokeless tobacco products (tuibur, sadha, gutkha), smoking, consuming alcohol, and chewing pan, can lead to genetic mutations associated with cancer, particularly tobacco-induced cancers, as reported by the Center for Disease Control and Prevention (US), 2010.

A potential strong association between tuibur, sadha, and pan chewing and breast cancer in Mizo women has been suggested, drawing parallels to a study by (Spangler et al. 2001) on Native American women (Eastern Band Cherokee). The similarities in lifestyle and dietary habits between Eastern Band Cherokee and Mizo women, both from tribal populations, increase the likelihood of shared risk factors. Additionally, research has shown that women with a family history of breast cancer who engage in detrimental lifestyles are at higher risk, aligning with findings from (Zodinpuii et al. 2020). While this study did not find a direct association between smoking and breast cancer, exposure to secondhand smoke remains a significant risk factor for women, as demonstrated by (Li et al. 2015).

Classification of Breast Cancer using Genomic Features

Random Forest analysis of breast cancer NGS data identified key attributes from both balanced (SMOTEENN) and imbalanced (original) datasets. These attributes include Chr, Start, End, CADD raw score, GERP++_RS, LRT score, Polyphen2_HVAR score, MutationTaster score, SIFT score, and SiPhy_29way_logOdds (Rentzsch et al. 2019; Dong et al. 2015; Schwarz et al. 2014; Davydov et al. 2010; and Adzhubei et al. 2010). SMOTEENN enhances training with class balance by eliminating noise in the data and generating new instances based on nearest neighbors (Husain et al., 2025). To address multicollinearity among features, a heat map was employed (James et al., 2013), resulting in the removal of Polyphen2_HDIV and MutationAssessor scores. Feature selection, reduction of overfitting, and extraction of non-linear relationships between features are accomplished using Random Forest classifier (Pudjihartono et al. 2022).

The selection of genetic attributes was further enhanced by developing Convolution Neural Networks (CNNs) on four datasets, including an imbalanced dataset and three oversampled datasets: ADASYN, SMOTEENN, and blsmote. ADASYN and SMOTE oversampling methods had significantly improved classification performance on the small datasets with less positive rates (Zhu et al. 2024).A study was designed incorporating Support Vector Machine (SVM) for imputation and Adaptive Synthetic (ADASYN) sampling for feature utilization to address missing values, while leveraging Convolution Neural Network (CNN)-generated features to enhance the model's capabilities. This model combination achieved exceptional performance, with 99.99% in accuracy, precision, recall, and F1 score (Munish 2024).Based on the model evaluation parameters, it was noteworthy that the feature subset identified by SMOTEENN could be the critical features for identifying pathogenic variants from the breast cancer germline dataset. CNNs had served as an effective tool for constructing classification models for breast cancer germline data, utilizing the SMOTEENN data balancing technique to address imbalance issues. Additionally, Random Forest Classifier identified significant patterns contributing to the disease. These findings can be applied to improve

diagnosis accuracy, provide genetic counseling, identify suitable drug targets, and conduct personalized breast cancer risk assessments. There is a need to increase the number of healthy cases in the genomic data so the real data can be utilized to support the findings.

Classification of Head and Neck Cancer using Epidemiological Features

The classification results showed very interesting insights to the head and neck cancer risk factors. Cases who were age above 45, high frequency smoking, using snuff and alcohol intake, and prolonged alcohol intake had shown 85% accuracy by Extra trees classifier in compared to other classifiers. There were considerable imbalanced between false positive and false negatives, all the five models were able to minimize the false positives but had shoot up in false negatives. These might have been an impact as a result of removing the missing values and outliers. The above-identified risk factors had been shown to have greater interactions in Nepal population with odd's ratio of 12.83 (Chang et al. 2020).

Classification of Head & Neck Cancer using Genomic Features

The top significant 10 features had been identified through Extra trees model:Chr, Start, End, Gene, MutationTaster score, FATHMM score, GERP++_RS, AF, non_neuro_AF_popmax, and ExAC_AMR. ExtraTrees classifier has not only been applied on medical data, it has been utilized for selecting relevant features for each type of security (intrusion detection). The extreme learning ensemble model, combined with softmax layer, has achieved an accuracy of 98%, which has been instrumental in identifying intrusion detection in real time (Sharma et al. 2019).The leukocyte classification by ResNet50 and feature selection by extra trees classifier yielded the highest accuracy of 90.76% (Baby et al. 2021).In the current dataset, the positive class was down-sampled due to significant data imbalance to mitigate bias towards the negative class. Similarly, downsampling was performed single cell transcriptome data based on unbiased datapoints which had effectively clustered the samples(Ren et al., 2022).

To assess the generalizability of the obtained results, a suite of five machine learning models (Logistic Regression, Random Forest, Decision Tree, Support Vector Machine and KNN) were developed (Carvalho et al., 2024; Kurian et al., 2022). A strength of this research is the integration of forward selection with grid search cross-validation hyperparameter tuning to identify the most relevant characteristics of genetic variants in head and neck cancer. However, a limitation is the potential loss of valuable information during the down-sampling process of the positive class.

Multi-class classification by integrating Gastric, Breast and Head & Neck Cancer features

The present work had identified five common risk factors associated with gastric, breast, and head & neck cancers: khuva (paan), tuibur (aqueous tobacco extract), smoking, alcohol, snuff (smokeless tobacco products), and smoked food. These risk factors were reported in other population by evaluating them using machine learning classifiers and statistical tools (Roy et al., 2024; Zami et al., 2024 and Zodinpuii et al., 2022). The present study had developed a Multi-layer Perceptron model that had effectively addressed this multi-class classification challenge (distinguishing between gastric, breast, head & neck cancers, and healthy cases) without overfitting.

Regarding the existing methods and results obtained via MLPs on multiclass classification problems, significant advancements have been made in recent years. These improvements have led to enhanced accuracy and efficiency in categorizing complex datasets across various domains. However, challenges persist in optimizing neural network architectures and addressing issues such as overfitting and computational complexity for large-scale multiclass problems. Strengths of the proposed MLP model: To prevent overfitting, the model's hidden layer, neuron count, and 'lbfgs' optimizer were carefully selected. SMOTE oversampling was utilized to mitigate underfitting. Limitations of the proposed MLP model: Future research should aim to increase both the sample size and the number of features to

elucidate additional factors concealed within epidemiological and genomic characteristics.

The results obtained using AdaBoost classifier (AI tools) can be used to design diagnostic gene-panel that are specific to gastric, breast and head & neck cancers in Mizo population (cancer high-risk). DNAH5 mutant had been involved in gastric cancer progression which was evident in the present study (Song et al. 2023). This gene was identified by machine learning tool, found to be one of the top most mutated genes in the gastric cancer big dataset. This study has also identified PARP4 in top 10 mutated genes of breast cancer and it had been found to be potential candidate for breast cancer in other population (Prawira et al. 2019). In head and neck cancer, PRUNE2 was among the top 10 and it had been playing a role in oral and thyroid cancers (Storvall et al. 2023). First 20 mutated genes can be further screened to develop multi-cancer panel for Mizo population. This Multi-Cancer Panel examines genes linked to various organ systems that are largely related with adult-onset, non-syndrome cancer predisposition disorders. These gene panels can detect genetic alterations for gastric, breast, and head and neck cancers that could serve us better in risk assessment, diagnosis, and treatment choices.

SUMMARY

The significant findings of this study are:

1. An algorithmic decision-making model was developed utilizing logistic regression to classify gastric cancer and healthy individuals based on epidemiological features, achieving an accuracy of 98.51%.

2. Saum (fermented pork fat), utilization of aluminum cookware, consumption of Zozial (infiltrated local cigar), initiation of smoking before 18 years of age, tuibur (aqueous tobacco extract), sadha (smokeless tobacco product), chewing of areca nuts (Khuva, zardhapan, supari), residence or occupation in proximity to cellular phone towers, and food storage in plastic containers were positively correlated with the incidence of gastric cancer in Mizoram.

3. Excessive salt intake, consumption of smoked foods, use of smokeless tobacco products (Khuva, Khaini/Sadha), and alcohol consumption were identified as risk factors associated with Microsatellite Instability in Gastric Cancer (MSI-GC).

4. This is the first study that had identified Khuva as a driving factor for MSI-GC, while excessive salt consumption, smoked food intake, Khaini/Sadha usage, and alcohol consumption were determined to be confounding factors for MSI-GC.

5. No biasness was observed during feature selection using 5-fold LASSOCV on balanced and imbalanced datasets.

6. Functional annotation (CADD and VEST3 scores, allele frequencies (East Asian, South Asian, Latino, and Ashkenazi Jewish), evolutionary conserved score (phyloP), gene, start position of the mutation, number of amino acid changes in gene (AAChange – use defined feature) reference allele (Ref) and mutated allele (Alt) were identified 96% of times as key attributes that causes pathogenicity in gastric cancer.

7. Ensemble classifiers have demonstrated superior overall performance in distinguishing between breast cancer cases and healthy controls based on epidemiological characteristics, achieving accuracies of 96.3% and 95.5% through LOOCV and 10-fold cross validation, respectively.

8. Heterozygous mutations were found to be higher in breast cancer dataset compared to healthy controls.

9. Features identified by Random forest were chromosome, position (start and end), function prediction scores (CADD, GREP, Polyphen, MutationTaster, LRT and SIFT) which showed well-balanced classification results on breast cancer big data.

10. On oversampling healthy datapoints by three distinct techniques (ADASYN, SMOTEENN and blsmote) from breast cancer big dataset had generated same feature subsets.

11. CNN model designed on (SMOTEENN) balanced dataset had good trade-off between precision and recall of 99.32% and 97.97% respectively with ROC-AUC of 100%

12. Individuals above the age of 45, with frequent smoking, snuff use, and prolonged alcohol consumption, were classified with 85% accuracy using the Extra Trees classifier on head and neck epidemiological characteristics.

13. Decision tree and SVM classifies have produced 96% accuracy with low false positive rate of 1 using head and neck germline variants with healthy controls.

14. Non-invasive characteristics were examined to elucidate potential cancer risk factors for gastric, breast, and head and neck malignancies.

15. Specific dietary habits and food preferences were found to distinctively contribute various cancer disparities.

16. A high-performing MLP model was selected based on its reduced computational time and enhanced efficiency to classify breast, gastric, head & neck and healthy controls.

17. Genes (nodes) and their functional protein association in gastric cancer big dataset were identified based on their experimental evidences: edges (two nodes interaction): GPSM2-NUMA1; RAG1-RAG2; MTERF4-NSUN4; WDR18-PELP1; CD58-CD2; and significant hubs (more than two nodes interactions): FANCM-MEIOB-FANCA-BLM-EXO1-TP53BP1; MCM3-TOPBP1-TP53BP1-CHEK1-CDC25C; TGFBRAP1-VPS39-VPS8; CHD23-USH1C-MYO7A; CWC22-CRNKL1-SLU7; PVR-CD96-CD226.

18. Genes (nodes) and their functional protein association in breast cancer big dataset were identified based on their experimental evidences: edges (two nodes interaction): UNC13B-RIMS2; ABCG5-ABCG8; PER3-PER1; F5-F2; DCTN1-ACTR1B; POLR1E-POLA1; and significant hubs (more than two nodes interactions): MPHOSPH10-PDCD11-NAT10-HEATR1-NOL10; CDH23-USH1C-MYO7B;

RPA1-RPA4-TIPIN-TIMELESS-WRN-RFWD3-FANCA; DYNC2H1-DYNC2I2-DYNC2LI1; TNRC6C-AGO2-GIGYF1.

19. Genes (nodes) and their functional protein association in head and neck cancer big dataset were identified based on their experimental evidences: edges (two nodes interaction):BCOR-PCGF1; CWC22-SRRM2; CEP55-TEX14; ABCG5-ABCG8; PER3-PER2; LLGL2-PARD6B; TUBGCP3-TUBG2; COG6-COG7; COL4A1-COL4A2; WNK1-WNK2; ORC4-ORC1; TACC3-AURKA; BCOR-PCGF1; ORC4-ORC1; PENK-OPRM1; and significant hubs (more than two nodes interactions): TP53BP1-TOPBP1-BRAC1-ATR-TP53;TRAF3-MAVS-IFIH1-TRIM65; RAD23B-XPC-PSMD2-TDP2; C5-C7-C8A; DDX18-NOC3L-PES1-PDCD11

20. Genes: MDN1, MKI67, DNAH5, FAT2, FAT1, SYNE1, SYNE2, USH2A, and CMYA5 in GC; XIRP2, USH2A, OR51B6, FBN3, PARP4 in BC; NEB AND OBSCN were common in all the three cancer types found to hold high prospective to develop gastric, breast and head & neck cancer in Mizo population which were evident from machine learning and deep learning models.

Machine learning algorithms were able to evidently identify the gastric, breast and head & neck cancer risk factors from diet-lifestyle with accuracies greater than 95%. Deep learning algorithms have shown great efficiency in detecting the foremost genetic traits of germline variants from gastric, breast and head & neck cancers with precision and recall greater than 95%. The epidemiological risk pattern of the above-mentioned three cancers had been achieved with accuracy of 75% using computationally less intensive architecture designed using MLP. The genomic risk factors of these three cancers had given an accuracy of 84.18% and f1_score of 84.43% by AdaBoost-hyperparameter tuned algorithm. The algorithm had identified twenty genes by integrating germline cancer dataset which will be signature genes to develop multi-cancer genetic panel for this cancer-burden population. There were many node (gene) interactions between each cancer used in this study, they are showing very interesting and unique patterns which gives different insight to understand the cause of cancer and its propagation within the cells.