Applications of Machine Learning in Predicting the Risk of Cancer in Adults: A Predictive Risk Model

Abiodun Ajanaku[a,*], Olufemi Isiaq[b]

[a] Department of Applied Artificial Intelligence and Data Science, Solent University, Southampton United Kingdom
[b] Solent AI Lab, Department of Computing, Solent University, Southampton, United Kingdom

ABSTRACT

Background: The incidence and mortality of cancer are increasing due to lack of awareness of the basic risk factors. The interaction of genes, lifestyle, comorbidities, and environment can influence a person's risk of cancer. This study is aimed at: (1) Identifying different risk factors and their effects in cancer risk classification for both symptomatic and asymptomatic adults using machine learning (ML) algorithms, (2) Identify, compare the performance of 6 machine learning models, and select the most suitable model for prediction, (3) synthesise existing knowledge to identify gaps to support the training and deployment a cancer risk predictive model.

Method: A quantitative analysis was carried out using primary dataset of 580 records with 58 features obtained through survey questionnaire from participants (aged 18 and above). The data were used to construct six machine learning models including decision tree (DT), random forest (RF), logistic regression (LR), support vector machine (SVM), naïve bayes (NB) and K-Nearest Neighbour (KNN) to predict and classify the risk of cancer in adults. A total of 10 relevant variables were input into these models. The performance of the models was measured using the area under receiver operating characteristic curve (AUC-ROC) and the confusion metrics.

Findings: The analysis identified 10 relevant risk factors used for the predictive risk model including: gender, sugar intake, frequency of exercise, smokes per pack year, level of alcohol intake, comorbid such as frequent cold, exposure to industrial pollution, domestic pollution, sun and co2 emissions. Two (2) cancer risk class (Low: 0 and High: 1) were predicted. Of the six machine learning models, the Random Forest ensemble learning methods had the best performance in predicting cancer risk (AUC: 0.93), outperformed conventional logistic regression (AUC: 0.83). While the decision tree, support vector machine, naïve bayes and KNN have AUC of 0.89, 0.87, 0.80, 0.80 respectively. With the Random Forest model, the number of people considered as having a high-risk of cancer (that is, RECALL, otherwise known as Sensitivity) was 93% with miss rate (false negative) of 7%.

Conclusion: The study showed that machine learning can accurately predict and classify the risk level of cancer by using a person's baseline information. Beyond the lifestyle and environmental risk factors, the study included chronic diseases (comorbidities) in a person in building the predictive model which have been overlooked in many previous studies. Risk factors such as gender, exposure to sun, chronic diseases, pollution resulting from domestic, co2 emissions, alcohol intake, and industrial practices contributed significantly to classifying the cancer risk level for Black African, Asian-Indian, Asian Pakistani, White British and White Irish. However, the sample size (n=580) used for the research was relatively small and limits the research in certain aspects. Hence, further research is necessary with a larger dataset than 580 and better threshold for each of the risk class to scale up the generalisability and feasibility of the predictive model in medical settings.

Keywords: Machine learning in healthcare, cancer risk factors, decision support system, predictive modelling, oncology, comorbidities

Abbreviations: AI, Artificial Intelligence; ML, Machine Learning; HPV, human papillomavirus.

* Corresponding author at: Solent University, East Park Terrace, Southampton, SO14 OYN, England, United King Kingdom. *E-mail address:biodunajan@gmail.com* (A. Ajanaku).

1. Introduction

Cancer is a leading cause of deaths globally and co-exist with other chronic diseases such as COVID-19 and tuberculosis. It is a non-communicable disease which has a high potential of morbidity and mortality; triggered by a rapid development of abnormal cell that grows beyond their usual boundaries which have the tendencies to invade adjoining part of the body and spread to other organs [1]. In other words, the disease could be called a malignant tumours or neoplasms. From a global perspective, cancer accounted for nearly 10 million deaths in 2020, representing one in six deaths. Around one-third(33.33%) of the global deaths from cancer were due to lifestyle (modifiable) risk factors including: use of tobacco, high body mass index, alcohol consumption, low fruit and vegetables intake, and lack of physical activities[1]. The combination of lifestyle and other risk factors such as family and past health history have influenced the development of many popular types of cancers such as breast, lung, colon, rectum, and prostate cancers. Furthermore, chronic infections as human papillomavirus (HPV) and hepatitis which are cancer causing infections are responsible for 30% of the cancer cases in low and lower-middle - income countries. In the UK for example, one in every two people will be told they have cancer at some point in their lifetime[2]. Based on recent available data, around 366,303 cases of cancer were diagnosed in 2017. Out of the total cases of 366,303, 51% are in men and the remaining 49% were in women. In contrast, 1,708,921 new cancer cases were reported and 599,265 people died of cancer in the United States in 2018**.** In addition, 436 new cancer cases were reported and 149 people died of cancer per 100,000 population in the US[3]. Whereas, In Africa Cancer is an emerging health problem, and it been projected that by 2030 there will be a 70% increase in the incidence and mortality rate due to accelerated population growth, aging, shortage of medical equipment, research resources and epidemiological expertise. Around 57% of all new cancer cases globally occurred in low-income countries, and significant part of this were from Africa [4].

A cancer risk is anything including modifiable behaviours and unmodifiable factors that can potentially increase a person's chance of getting cancer [5]. Although most of the risk factors do not directly cause cancer, however, they can influence the development of cancer [6]. Modifiable risk factors include but not limited to lifestyle behaviours such as alcohol consumption, use of tobacco, unhealthy diet, physical inactivity. This also includes air pollution, and occupational risk. The unmodifiable risks include age, genetic or other family medical history of a person. Beyond the modifiable and unmodifiable risk factors, some chronic infections associated with geographical locations are risk factors for cancer which are common in low-and-middle income countries and accounted for 13% of the global cancers diagnosed in 2018 [1]. This includes chronic infections such as Helicobacter pylori, human papillomavirus (HPV), hepatitis B virus, hepatitis C virus, and Epstein-Barr virus.

In terms of current state-of-the-art, the major approach in identifying the risk of cancer at the early stage is the population-wide asymptomatic screening, which is a pathological approach that is aimed at identifying individuals who do not show symptoms of cancer and could be at risk. This screening includes but not limited to mammography, cervical, and faecal occult blood testing for breast, cervical, and colorectal cancers respectively. There are numerous nation-wide screening programmes in the US and UK designed to specifically for early detections of the most common cancers including breast, colorectal, bowel and cervical. For instance, the US has the National Breast and Cervical Cancer Early Detection Programme (NBCCEDP). Despite the numerous nation-wide efforts and resources that have mobilized to identify early risk of cancer, majority of the people eligible for the screening programme do not participate due to fear, not creating time for the screening or they do not see any value in the programme. Secondly, both the World Cancer Research Fund and the Cancer Research UK[7,8] have an online web cancer risk quiz platform that people complete in understanding their risk of cancer. These platforms are not based on any machine learning models; merely informational without categorizing a person's risk of cancer accurately into class as either low or high. This research provides practical solution to the gap through the applications of machine learning algorithms in predicting the risk of a cancer and makes a far- reaching departure from existing state of art. It combines both accuracy and speed in learning from data, finds patterns, trains itself using the labelled data to predict an outcome: low, or high risk of cancer.

Machine Learning (ML) is a core part of Artificial Intelligence and extension of traditional statistical techniques that uses computational resources to detect underlying patterns in high-dimensional data, and it is increasingly being used in different areas in healthcare and other domains requiring predictions. The goal of machine learning models/algorithms is to facilitate and establish the best way in mapping of features (inputs) to output(labels) through learning from the dataset[9]. The six models selected for this research project includes Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, Logistic Regression Classifier, Naïve Bayes Classifier and K-Nearest Neighbour.

The accelerating increase in the global cancer burden and mortality has resulted into more research on the risks of cancer in the past few years. The recent demands for machine learning in healthcare and the growing trends towards personalised predictive medicine have facilitated the research process. Many recent studies have focused on the applications of machine learning in predicting the different types of cancer with emphasis on lifestyle and environmental factors, only a few research has taken cancer risk factors combined with comorbidities into consideration. Hence, this study provides the opportunity for further investigation. In recapping related works on the research topic, the PubMed database was searched from inception to September 1 using the search string: *(("Supervised Machine Learning"[Mesh] OR "machine learning in healthcare" OR "prediction of cancer" OR "ML") AND ("Neoplasms"[Mesh] OR "cancer" OR "Cancers" OR "Oncology" OR "prognosis")) AND ("Risk Assessment"[Mesh] OR "risk factors" OR "comorbidities" OR "symptoms").* The search returned 594 unique articles of which 554 articles were excluded because they did not meet the inclusion criteria, were duplicates or were related to other studies. Full-text reviews were conducted for 40 articles and a final set of 6 articles were included for literature review because they deployed ML to predict cancer risk or cancer related diseases in individuals.

Moncada-Torres and colleagues, for example, suggests that explainable machine learning models can outperform cox regression predictions (including other standard for survival analysis in oncology) and provide insights in breast cancer survival if the process of the decision making are transparent and explainable [10]. Ye and colleagues deployed the ensemble feature learning with 20 variables in identifying risk factors for predicting secondary cancer with due consideration for class imbalance and patients' heterogeneity. Three ensemble models: decision tree (DT), support vector machine (SVM) and k-nearest neighbour (KNN) were used for the classification problem. Amongst the three classifiers, DT produced the best result for predicting secondary cancer with 0.72 and 0.38 in terms of AUC and F1-score when the patients were divided into 15 and 20 groups respectively [11]. Achilonu and colleagues argued that accurate prediction of patients at risk of cancer can enhance clinical expectations and decisions. The authors applied supervised machine learning approach in predicting the colorectal cancer recurrence and patients' survival: a South-African population-based approach. For higher predictive performance and interpretability, six supervised machine learning models were evaluated. The models included logistic regression(LR), naïve bayes(NB), decision tree C5.0, random forest (RF), support vector machine (SVM) and artificial neural network(ANN). Although the six algorithms produced high accuracy in terms of AUC-ROC and without much significant difference, however, ANN outperformed the other models with highest AUC-ROC for recurrence (87.0%) and survival (82.0%). The variables used as inputs for the modelling includes patients' age, histology, radiology stage which are relevant features for both cancer recurrence and survival [12]. Cruz and Wishart carried out a detailed systematic review of the different machine learning techniques, the data types being used, and the performance of the various models in predicting cancers and prognosis. The main findings from the review were a bias towards the application of older technologies such as artificial neural networks (ANNs) in place of more interpretable and recent machine learning models such as logistic regression, naïve bayes, decision tree and other ensemble methods. In addition, the authors stated there were absence and lack of appropriate level of testing and validation of the models in several published studies. The authors concluded that the well designed and validated studies show a substantial improvement of around 15-25% accuracy in predicting the cancer susceptibility, recurrence, and mortality [13]. Alfayez and colleagues, carried out a systematic scoping review on the application of machine learning in predicting the risk of cancer in adults. The authors used a scoping review based on population, context and concepts to identify and understand the various machine learning that have been deployed in predicting the risk of cancer using clinical, demographic, and lifestyle behaviour datasets. The finding of the research showed that wide variety of ML models were used in different studies including Artificial Neural Networks(ANN: 8 out of 10 studies), Logistic Regression (2/10 studies), Gaussian naïve Bayes (1 out of 10 studies), Bayesian network inference (1/10 studies), DTs (1/10 studies) and RFs (2/10

studies), linear discriminant analysis (LDA) (1/10 studies), and SVMs (1/10 studies). The scoping review did not identified a single 'best' method. This is because not all models generalised well to the validation datasets[14]. Bundazak and colleagues analysed and predicted the risk of cancer using Decision Tree Classifier using 1,030 datasets obtained from a survey. The research used 10 variables including participants smoking behaviour, dietary, alcohol consumption, exercise, occupational risk, environmental risk, medical, personal history, and other demographic profile of participants. The authors deployed two machine learning models: decision tree classifier and neural network. The results of the research showed that decision tree (DT) model performed better that than the Neural Network model. The DT model produced 85.63%, 16.7%, and 35% for accuracy, mean-squared error, and root mean squared error respectively. Whereas the neural network produced a score 45.3, 0.5013 and 0.5017 for accuracy, mean-squared error, and root mean squared error respectively[15]. The aim of this study is to investigate different risk factors in adults and identify relevant variables for machine learning predictive model using 9 risk categories: *anthropometric, demographic, environmental, personal history, family history, social and lifestyle, occupational, economic class, and co-existing diseases in a person*. Consistent with the aim, the study addresses the below main research question:

- What is the effect of different risk factors in predicting and classifying the level of cancer risk using machine learning techniques?

Furthermore, the study insights and outcomes of the study will provide significant benefits to the below beneficiaries in the following way:

- **Individuals**: the cancer risk predictive system: which is one of the outcomes of the research will help individuals at the comfort of their homes to detect the risk of cancer early and gain insights into the relevant cancer risk factors; this will in turn drive informed decision about going for screening and seeing a health professional.
- **Health Professionals**: can adopt the predictive system as a preliminary fact-finding and vulnerability assessment tool to form the basis for referring patients for cancer screening and diagnosis.
- **Global health community and government**: the findings and insights from the research project will serve as a contribution to the body of knowledge especially on cancer epidemiology and new datasets.

2. Methodology

The study is quantitative research aimed at creating a cancer risk prediction system using evidence-based data obtained through a primary data collection method. The quantitative methodology was selected in line with the aim of the study to produce generalisable knowledge about the different risk factors for cancer, enhance personalised predictive healthcare, and achieve higher interpretability in findings to accelerate adoption in the medical settings. One key issue considered extensively throughout the research was to establish the effect and magnitude of relationship between the different cancer risk factors (predictors) and the cancer risk level (target).

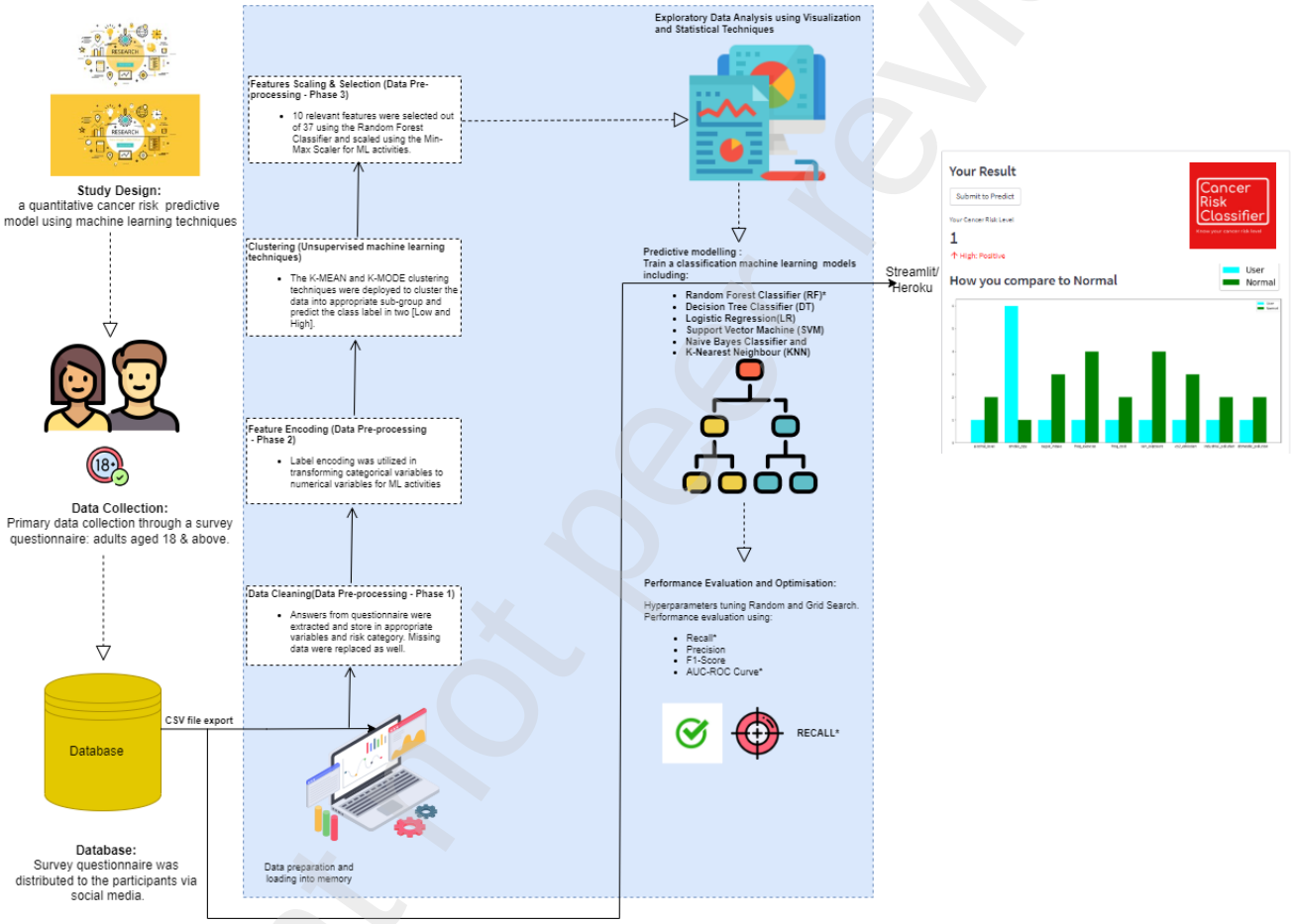

Fig. 1: Schematic representation of research methodology

2.1. Study design and sampling

Based on the study objectives, a range of different risk factors were included in the survey questionnaire distributed to the target population (Adults aged 18 and above). The questionnaire was designed using a combination of multiple choice, Likert-scale, matrix choice, slider scale and short answers questions. There were 27 main questions (including 31 sub-questions) as detailed in table 1 below:

Table 1

Primary data: survey questionnaire design

Type of Questions	Number/Frequency of Questions
Short (answer) question	5
Closed-ended multiple choice questions	9
Matrix Choice (one answer)	1
Slider Scale	1
Likert scale (slider, matrix, and rating)	11

The multiple-choice questions were used for the nominal variables because they lack numerical significance and can be measured using the frequency distribution table to depict each option and the frequency at which these options were selected. Whereas the Likert-scale questions were used for the ordinal variables because they depict the order of value. There are quantitative values because one rank is higher than the other. The Likert-scale questions were measured using Mode, Median and cross tabulation analysis. In similar vein, the short (answer) questions were used for the ratio variables. The ratio variables depict the order and differences between values and can take true zero point. It is quantitative and the absence of a value can still provide information. These variables were measured using mode, median and mean. In addition, the scale can be analysed using t-tests, ANOVA and correlation analyses. ANOVA tests the significance of the survey results. While t-tests and correlation establish variables relationship [16]. Overall, there were 55 categorical variables (20 nominal and 35 ordinal) and 3 numerical variables (ratio). This implies that the data table contains 57 independent variables and one initial (target) variable called cancer_status.

The risk factors(variables) were grouped into the following 9 categories using survey questionnaire as shown in Table 2.

Table 2

Category of cancer risk factors and variables

Risk Category	Variables
Demographic	age, gender, ethnicity, country of origin, country of residence.
Social and Lifestyle	alcohol intake, smokes per pack year, intensity of physical exercise, frequency of exercise, fruits and veggies intake, sugar intake from sugary drinks and high fat processed foods.
Anthropometry	height, weight.

Personal history	personal history of cancer: existing medical conditions.
Family history	family history of cancer: existing medical conditions.
Comorbidities	chronic diseases such as lung cancer, heart problems, dry cough, fatigue, shortness of breath, wheezing, swallowing difficulty, clubbing of fingernails, frequent cold or urination, dry cough, snoring, pain and bloody discharge from any part of the body.
Environmental risks	sun exposure, pollution from domestic, water, co2 emissions and industrial activities.
Occupational risks	measured by skill levels:
Economic class	measured by economic class:

2.2. Data collection

The research project utilised primary data collected through survey questionnaire distributed online through social media platforms including Facebook, Twitter, Instagram, LinkedIn, WhatsApp group amongst others. In addition, participants who completed the questionnaire also referred others to complete the survey. Hence, word of mouth/referrals had the highest reach of 46.01% followed by WhatsApp group with 36.64% . The method has proven to be very effective for collecting data from people about possible exposure to risk factors, symptoms, and co-existing medical conditions. Hence, it is widely used for epidemiology research in the healthcare and amongst others [16]. Furthermore, it is the preferred method to reach a wider population and provides the flexibility in collecting data that reflects the characteristics and demographic of the target population, which is important for quantitative analysis and generalisation of the research results.

2.3. Data analysis

The dataset collected was anonymised and each instance (row) was assigned a serial number and represents the cancer risk data of each participant. The analysis was based on the cross industry standard process for data mining (CRISP-DM) model [17]. The four analysis steps were: *1) Data cleaning(data pre-processing - phase one):* this included extracting answers from questionnaire and storing in appropriate variables and risk category, replacing missing values using the most frequent strategy and constant such as 0 to represent N/A, constraining outliers to the median value, 11 out of the initial 58 features were dropped because they contained question heading. Additional cleaning of the data was carried to correct for inconsistent input by users and reduced the

features further from 47 to 37; *2) Feature encoding (data pre-processing – phase two)* including label encoding in transforming categorical variables to numerical variables for machine learning activities, the variables encoded included: ethnicity, gender, economic, class, occupational risk, smoke, family cancer history and cancer status; *3) Clustering using unsupervised machine learning techniques*: this involved establishing the optimal number of class label using the ELBOW method and Silhouette Score and the use of K-MEAN an K-MODE clustering techniques to cluster the datapoints in appropriate sub-group and predict the class label: Low and High; *4) Features scaling and selection (data pre-processing – phase three)* involving the deployment of Min-Max scaler to transform each feature to a given range (feature range (0,1)) for machine learning activities. The process of feature selection was carried out using five feature selection techniques including Random Forest Classifier, Chi-Square, Forward Features Selection, Backward Features Elimination and Mutual Information. Random Forest Classifier was selected as the preferred feature selection technique and was deployed in selecting the 10 relevant features in building the machine learning predictive model.

Fig. 2: Word Cloud Visualisation of Cancer Risk Factors with Connection to Personal and Family History of Chronic Medical Conditions (Comorbidities)

The analytics and visualisation in Fig. 3 below show the prevalence of different risk factors based on gender, and ethnicity. The first figure shows that female has high outliers with regards to intake of alcohol, and the consumption of alcohol tends to decrease further down the gender class. However, the cancer risk level is higher in female than male with a count of 350 and 200 respectively. The remainder are attributed to participants who did not indicate their gender. The age bracket of 25-47 has the highest distribution of cancer risk. Interestingly, the ethnicity: Asian-Indian(0) has the highest class of high cancer risk attributed to exposure to co2 emission followed by Mixed: White & Black African and Mixed: White & Black Caribbean. Whereas, Asian- Pakistani[0], the Asian-Indian[1], Black African [2], White British [9] and White Irish[10] ethnicity group have relatively high risk class of cancer attributable to sun exposure.

The research project support existing literature that the incidence and mortality of cancer can be reduced when the risk factors are better understood, detected, and treated on early; making early detection of risk vitally important. The insights and results from the research project provide practical solution to existing barriers to cancer screening resulting from fear or concerns about medical procedures, lack of knowledge of risk factors.

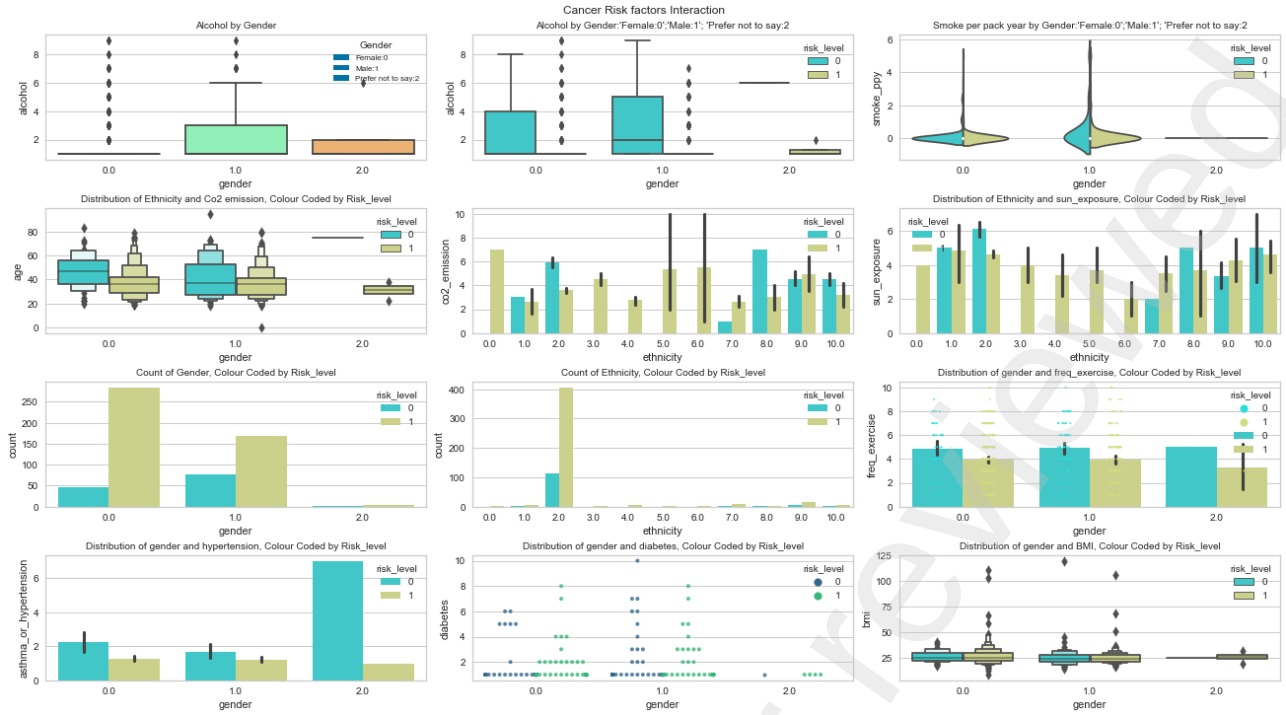

Fig. 3: Data analytics and visualisation of cancer risk factors:

3. Results

Consistent with the objectives of the study, the dataset was divided into training (n=371), validation (n=93) and testing (n=116) for building, training, and testing the model respectively. The Kruskal-Wallis hypothesis testing as outlined in Table 6 was used to test the hypothesis that different risk factors have effect on the level of cancer risk. Apart from age, height, and weight; other risk factors were recorded on a Likert-rating scale of 1 to 10 as the predictors; and the level of cancer risk was classified into two categories (Low risk: 0 and High Risk: 1) as the target variable. The Kruskal-Wallis hypothesis showed a t-statistics (95% confidence level) of 54.9 with p-value of 0.000 which is less than the critical value of 0.05. This contradicts the initial hypothesis(H0) that the different risk factors are not statistically significant and do not have any effect on the level of cancer. Both the random forest classifier and the Chi-square test of association (otherwise known as test of independence) in Fig.6 and Table 8 were deployed to find the relevant (significant) features with strong correlation and information about the target variable. Similarly, the machine learning algorithms were tuned using random search in Table 4 to select the best hyperparameters and to optimise the model performance as well as customise it. Whereas Table 3 displayed the models results using the default model without hyperparameters tuning. To establish the average generalisation performance of the model, the 10 folds cross validation was deployed to validate how well the model generalises its learning Table 4. Fig 4 summarizes the K-MEANS cluster analysis carried out to show number of clusters and patterns in each cluster. The AUC-ROC (Fig.5) shows the separability power of the model in classifying positives and negatives risk of cancer.

Table 3

Summary of machine learning model results: using the default model without hyperparameters tuning

S/N	Model	Target	Accuracy Score	Precision	Recall	F1-Score
1	Random Forest Classifier (RF)	Low risk: 0	0.70	0.71	0.82	0.76
		High risk: 1		0.68	0.52	0.59
2	Decision Tree Classifier (DT)	Low risk: 0	0.61	0.65	0.75	0.70
		High risk: 1		0.67	0.55	0.60
3	K-Nearest Neighbour (KNN)	Low risk: 0	0.65	0.70	0.82	0.75
		High risk: 1		0.46	0.31	0.37
4	Support Vector Machine (SVM)	Low risk: 0	0.53	0.57	0.75	0.65
		High risk: 1		0.41	0.24	0.31
5	Naïve Bayes Classifier (NB)	Low risk: 0	0.75	0.58	0.34	0.43
		High risk: 1		0.78	0.90	0.84
6	Logistic Regression (LR)	Low risk: 0	0.83	0.43	0.16	0.23
		High risk: 1		0.85	0.95	0.90

Table 4

Random Search hyperparameters tuning

Hyparameters	**Declared dictionary of hyperparameters for tuning**	**Best Estimator Hyperparameters**
max_depth	[int(x) for x in np.linspace(10, 120, num = 12)]	120
min_samples_leaf	[1,2,4,6,8,9]	4
min_samples_split	[2, 6, 10]	6
bootstrap	[True, False]	FALSE
n_estimators	[5,20,50,100]	50
max_features	['auto', 'sqrt']	sqrt
criterion	["gini", "entropy"]	gini
random_state	[42, 35, 10, 0]	35
Cross-validation scores:[0.87234043 0.82978723 0.80851064 0.78723404 0.82608696 0.82608696 0.86956522 0.84782609 0.82608696 0.82608696]		
Cross val mean: 0.832 (std:0.024) Best Score: %s 0.931; Number of run was set at 10.		

Table 5

Summary of machine learning model results: using the best hyperparameters

S/N	**Model**	**Target**	**Accuracy Score**	**Precision**	**Recall**	**F1-Score**
1	Random Forest Classifier (RF)	Low risk: 0	0.92	0.96	0.91	0.93
		High risk: 1		0.88	0.93	0.91
2	Decision Tree Classifier (DT)	Low risk: 0	0.85	0.75	0.91	0.82
		High risk: 1		0.74	0.82	0.88
3	Support Vector Machine (SVM)	Low risk: 0	0.83	0.90	0.88	0.89
		High risk: 1		0.63	0.68	0.66
4	Naïve Bayes Classifier (NB)	Low risk: 0	0.76	0.56	0.33	0.42
		High risk: 1		0.80	0.91	0.85
5	Logistic Regression (LR)	Low risk: 0	0.72	0.77	0.76	0.76
		High risk: 1		0.64	0.65	0.65
6	K-Nearest Neighbour (KNN)	Low risk: 0	0.67	0.74	0.71	0.72
		High risk: 1		0.58	0.61	0.60

Table 6

Kruskal-Wallis Hypothesis Testing Results

t-statistics	54.9112
p-value	0.0000
Alpha	0.05
Confidence Level	0.95
Sample size: C1- Low	364
Sample size: C2-High	216
Total Sample size	580

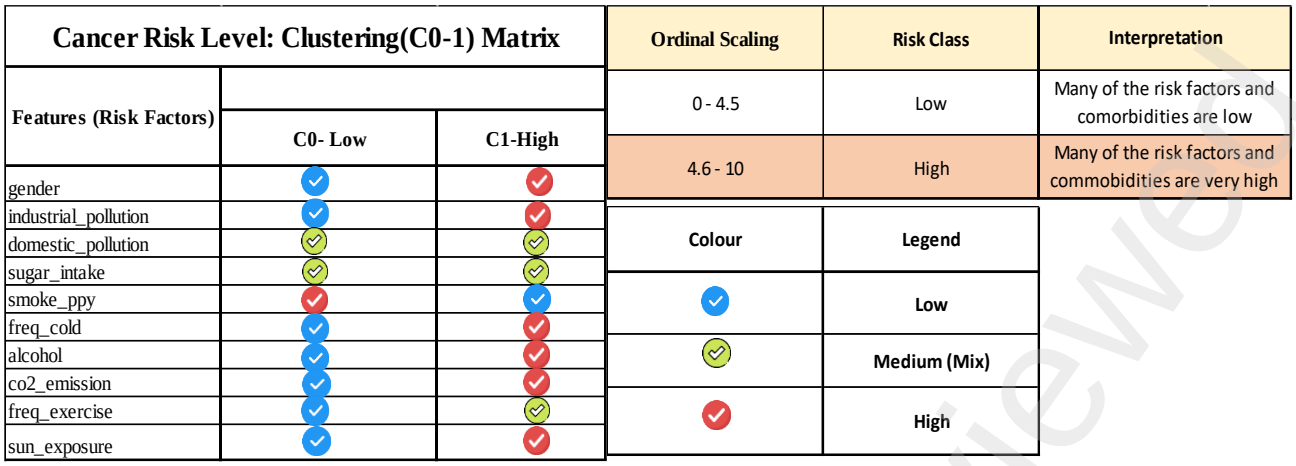

Fig. 4: Cancer risk level clustering (C0 – C1) matrix

Table 7
Summary of top 10 variables Interactions using Linear Regression and corresponding R^2

S/N	Variable Interaction	R^2
0	Baseline	0.187
1	['industrial_pollution', 'domestic_pollution']	0.230
2	['sugar_intake', 'gender']	0.220
3	['sugar_intake', 'smoke_ppy']	0.220
4	['gender', 'domestic_pollution']	0.216
5	['sugar_intake', 'interaction']	0.212
6	['industrial_pollution', 'gender']	0.209
7	['sugar_intake', 'freq_exercise']	0.208
8	['sugar_intake', 'domestic_pollution']	0.206
9	['sun_exposure', 'freq_exercise']	0.205
10	['gender', 'co2_emission']	0.204

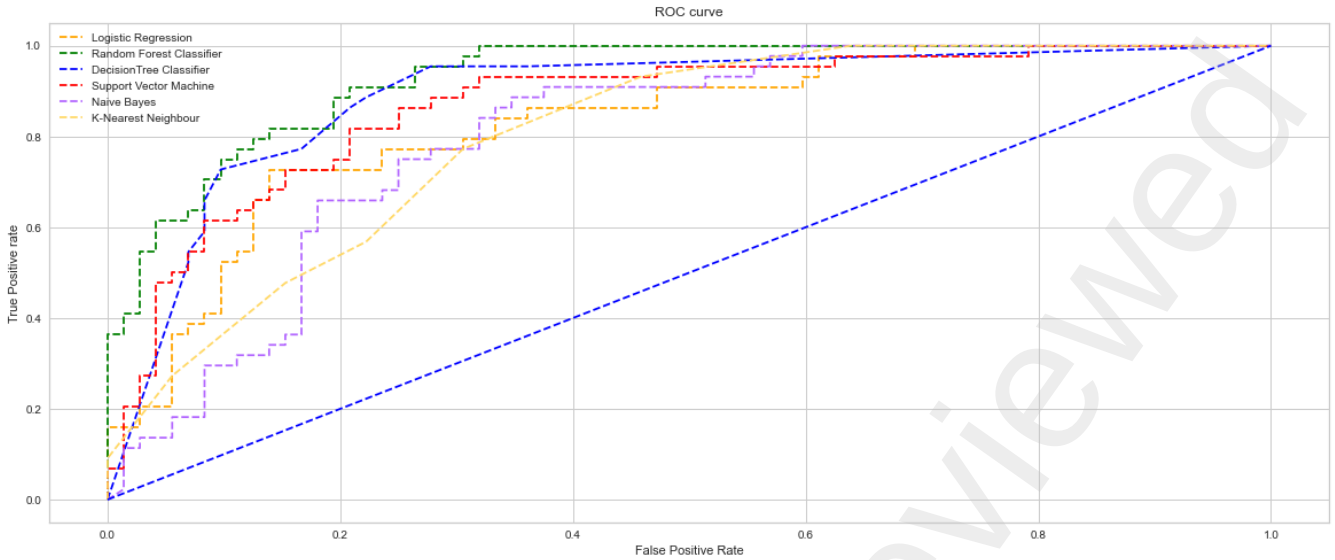

Fig. 5: The AUC and ROC curve for the six models

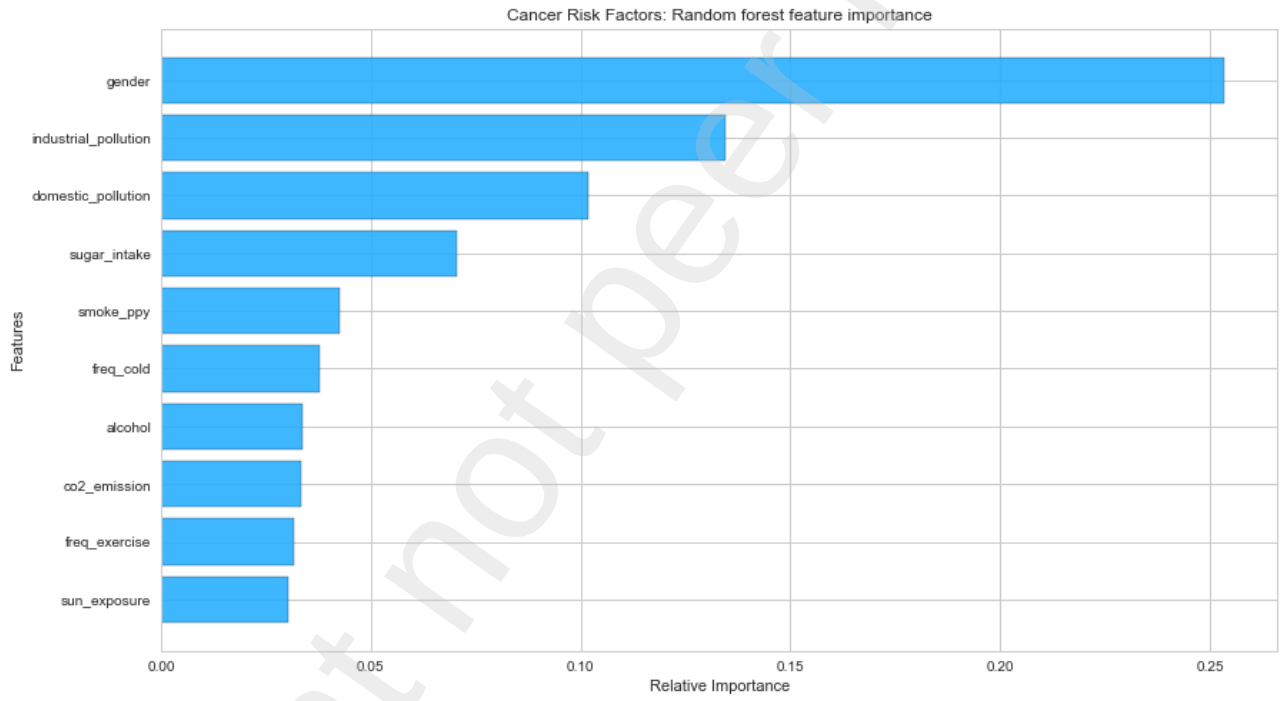

Fig. 6: Cancer risk factors: Random Forest features importance

Table 8:
Multi-variate analysis using Chi-Square test of independence (otherwise, known as test of association)

S/N	Features	p-value	Decision Rule: (Significant if p-value < 0.05)
1	age	7.97E-02	Not Significant
2	*gender	5.26009E-12	Significant
3	eclass	0.8838855	Not Significant
4	occupational_risk	0.6250767	Not Significant
5	intensity_pa	0.9250262	Not Significant
6	*freq_exercise	0.03914621	Significant
7	fruit_veg	0.4378758	Not Significant
8	wholegrains	0.6020934	Not Significant
9	*meats_intake	0.0290362	Significant
10	*sugar_intake	0.000211764	Significant
11	*alcohol	0.001734447	Significant
12	*smoke_ppy	0.005070948	Significant
13	pain	0.9560743	Not Significant
14	bloody discharge	0.8185542	Not Significant
15	fatigue	0.8249327	Not Significant
16	shortness_of_breath	0.3511865	Not Significant
17	wheezing	0.1468028	Not Significant
18	swallowing_difficulty	0.5246173	Not Significant
19	*clubbing_of_nails	0.02632377	Significant
20	*freq_cold	3.94769E-08	Significant
21	*dry_cough	0.01608948	Significant
22	*snoring	0.000123999	Significant
23	*diabetes	0.03436149	Significant
24	heart_problems	0.6108036	Not Significant
25	*asthma_or_hyperten-sion	0.01170021	Significant
26	lung_problem	0.05588973	Not Significant
27	Loss_weight_or_appe-tite	0.182712	Not Significant
28	family_cancer_history	0.1392305	Not Significant
29	sun_exposure	0.7424564	Not Significant
30	co2_emission	0.1467027	Not Significant
31	*industrial_pollution	1.9042E-07	Significant
32	*domestic_pollution	0.000201513	Significant
33	*water_pollution	0.001187506	Significant
34	ethnicity	0.1913963	Not Significant
35	*bmi	0.01233962	Significant

- Means significant risk factors

4. Discussion

The overall aim of this research project is to understand and establish how different risk factors impact the classification of the cancer risk level using machine learning. To achieve the objectives, six machine learning models were selected, this includes Random Forest Classifier(RF), Decision Tree Classifier(DT), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB) Classifier and Logistic Regression (LR). The result indicates that Random Forest model is the most suitable machine learning model with an overall accuracy and AUC score of 92% and 93% respectively.

The top 10 important and significant features utilized in building the model as detailed in Fig.6 and Table 8 includes gender, industrial pollution, domestic pollution, sugar intake, smoke per-person-year, frequency of cold, alcohol intake, co2 emission, frequency of exercise, and sun exposure. Several previous studies each using different approaches validated by the work of [1, 6, 7, 8, 15] concluded that cancer risks are inherently associated with lifestyle behaviours and environmental factors such as alcohol intake, sugar intake, smoke per-person-year, sun exposure and various pollution. This research supports existing studies, however, the discovery from the research was that other cancer risk factors such as snoring, body mass index (bmi), water pollution; co-existing medical condition like diabetes, dry cough, clubbing of nail as depicted in Table 8 can individually or through interaction with other factors significantly increase a person's risk of cancer if not detected early or untreated; are destined to develop into cancer. For instance, snoring and frequent cold in a person can lead to sleeping disorder which has a connection with several cancer types. Hence, the knowledge synthesis enhanced the efficiency of the model in accurately predicting and classifying the risk of cancer of the target population [adults]. In the context of the research problem, RECALL (Sensitivity of the model) is the most important success/evaluation metric.

The RF model outperformed the other models because it leverages the power of multiple trees for making decisions. In addition, the ensemble tree-based model considers all the variables sequentially and takes all interactions amongst the variables in making decision. Hence, both the RF and DT models proved to be the two most effective models for cancer risk prediction and classification. The results support existing studies that tree-based models are most widely used machine learning models for decision support system in healthcare/medical setting [13]. On the other hand, the DT, SVM, NB, LR and KNN models show accuracy [85%, 83.%, 76%, 72% and 67%] and AUC scores of [89%, 87%, 80%, 83%, 80%] respectively. However, they are not effective to match the predictive performance of the RF classifier. The KNN and logistic regression models produced the least accuracy and AUC score respectively. These are obviously not the most effective model considering for the classification problem classification of cancer risk level.

4.1. Evaluation and efficiency of the model

The performance measurement(success metric) of the model was based on the AUC-ROC and RECALL (as shown in Table 3. The metrics give a vivid visualisation of how well the models performed on the basis of separating low(negative) and high (positive) risk of cancer. The higher the AUC-ROC score, the better the performance of the model in separating between positive and negative class. The Random Forest classifier produced the highest AUC score of 93%. RECALL depicts the sensitivity of the model in how many of the actual positive cases (Low and High Class) the models can predict correctly. Random Forest Classifier produced the highest recall score of 91% and 93% for the low and high risk respectively. The random forest model performance was evaluated based on the confusion matrix classification report summary in Table 5. The classification results show a summary of the predictive results of the risk classifier as will now be discussed below:

- **Low Risk**: True positive(TP) was 64, False Negative(FN) was 6, False Positive(FP) was 3 and True Negative (TN) was 43. This implies a 96%, 91% and 93% for precision, recall and F1-score respectively.
- **High Risk**: True positive(TP) was 43, False Negative(FN) was 3, False Positive(FP) was 6 and True Negative (TN) was 64. This implies a 88%, 93% and 91% for precision, recall and F1-score respectively.

Recall are the actual positive cases the model can predict correctly which is 91% and 93% for the Low risk and High risk respectively. This is because False Negative(FN) - Type II error is of higher concern than False Positive(FP) - that is Type 1 error. Hence, True Positive(TP) should not go undetected for the two level of cancer risk. Despite the many advantages of the Random Forest classifier, the model has been labelled as a "Black box" because it there is no specific ways of interpreting the results of the model. This makes the model complex.

4.2. Validation of hypothesis, research questions and variable interactions

In line with the hypothesis, the different risk factors are statistically significant in establishing and classifying a person's level of cancer risk. The random forest features importance in Fig. 6 validates that the top 10 important features combined explained approximately 77% variation in the target variable. In addition, the overall LRR p-value was 0.0000 much less than the 0.05 (alpha: critical value) as shown in Table 4. Therefore, we can reject the null hypothesis (H0) and accept the alternate hypothesis (H1) that the predictors (risk factors) are statistically significant. In addition, the linear regression was used to evaluate and establish the top 10 variable interactions. The first three interactions resulted in slight improvement in the R^2 of 23%, 22% and 22% respectively.

4.3. Main Findings

The cluster analysis results Fig.4 provides new insights and contribution of the study in establishing the relationship between different risk factors, comorbidities, ethnicity, and level of cancer risk as detailed below:

- **Cluster C0**: Low risk – individuals in this cluster have many of the risk factors and comorbidities as relatively low.
- **Cluster C1**: High Risk – Individuals in this cluster have many of the risk factors and comorbidities as very high.

Previous research has overlooked the impact of existing medical conditions in people when predicting and classifying risk cancer. Most of these studies have applied machine learning on cancer risk prediction based on lifestyle and environmental. Although, this study shows that comorbidities have a modest relationship with other risk factors as depicted in Fig.6 , however, long-term chronic/long-term medical conditions such as high blood pressure (hypertension), obesity, mental health problems, and kidney disease are common(shared) risk factors that affect people without cancer as well as people with cancer; lifestyle factors such as lack of exercise, and poor diet can aggravate the comorbidities. Hypertension in Men may be associated with the increased risk of developing prostate cancer. Similarly, hypertension in women can be linked to endometrial and breast cancer as well as renal cancer [18].

5. Conclusion and Recommendations

This research clearly demonstrated the effect of the different risk factors in the prediction and classification of the level of cancer risk; provides insights into the general characteristics of the target population such as ethnicities, country of origin, country of residence amongst others. Furthermore, the study made significant contributions in the identification of different risk factors which are the major influencers of the many types of cancer including lungs, breast cancers and many more. The target population (adults) were clustered into **Low risk**: individuals in this cluster have many of the risk factors and comorbidities as relatively low and **High risk:** individuals in this cluster have many of the risk factors and comorbidities as very high.
The results showed that the different risk factors when combined with co-existing diseases (comorbidities) in a person are statistically significant in predicting and classifying cancer risk level. These insights were completely overlooked by previous studies.

Given the rapid demand for personalised predictive healthcare; individuals, health professionals and the global health community can enhance capability in spotting the risk of cancer early by adopting the predictive system created as one of the major outcomes of this study; utilize it at the comfort of their homes and in health facilities as a preliminary fact-finding system for cancer risk assessment and vulnerability. Hence, data-driven decisions can be made for cancer screening and diagnosis for different types of cancer based on the results from the predictive system. Early detection can save lives, reduce cancer incidence and mortality, and the huge associated cost with the treatment of the disease. The research was focused on predicting and classifying the level of cancer risk without much information on the threshold of the risk class. In addition, it does not predict the actual type of cancer; this limits the work in certain aspects. Improving on this aspect will provide for future work in predicting cancer types. In terms of the sample size, the target was to obtain 1000 dataset for the research, however, only 580 dataset was gathered which is another limitation of the study. Hence, further research can be done with primary data source with relatively larger dataset than 580 and better threshold for each of the risk class.

Summary Table

What was already known about the topic:

- Machine learning/AI personalised predictive healthcare is on high demand.
- Previous studies have overlooked the impact of co-existing diseases (comorbidities) in a person's when predicting the risk of cancer.
- Most common cancer risk factors used in predictive models are lifestyle and environmental factors.

What this study added to our knowledge:

- Beyond the lifestyle and environmental risk factors, the study included chronic diseases (comorbidities) in a person in building the predictive model which have been overlooked in many previous studies.
- The study considered the general characteristics of the target population ethnicities, country of origin, country of residence amongst others. Hence, these were very well represented in the relevant variables used in building the predictive model and ultimately contributed to the body of knowledge especially on cancer epidemiology and new datasets.
- Provided practical and novel solution to people's apathy to cancer screening. Individuals, health professionals and the global health community can use the predictive system in real-time: at the comfort of their homes and at health facilities to establish cancer risk level and vulnerability, which in turn supports decision-making for cancer screening and diagnosis.

Acknowledgements
I am heartily thankful to Mrs. Emmanuella Ajanaku for her encouragement and the material support that have made this work successful.

Data Availability Statement
The primary data used for this study was obtained through a survey questionnaire. Participants responses were anonymized, and personal identifier information was not collected.

Ethical Statement
Participants gave their consent to participate in the survey and were given the opportunity to withdraw from the survey at any time, without reason. The study did not engage any less privileged, physically challenged individuals nor animal subjects.

Ethics Approval
Ethical approval was applied and granted for this research by Solent University.

References

[1] World Health Organization, Cancer. https://www.Who.Int/News-Room/Fact-Sheets/Detail/Cancer, 2022 (accessed 3 February 2022).

[2] Gov.uk, Closed consultation: 10-year cancer plan: Call for evidence. https://www.Gov.Uk/Government/Consultations/10-Year-Cancer-Plan-Call-for-Evidence, 2022(accessed 29 May 2022).

[3] Centers for Disease Control and Prevention. Cancer Data and Statistics. https://www.Cdc.Gov/Cancer/Dcpc/Data/Index.Htm, 2022(accessed 3 June 2022).

[4] Y. Hamdi, I. Abdeljaoued-Tej, A. Zatchi Afzal, S. Abdelhak, B. Samir, J.S. Brown ., A. Benkahla; Cancer in Africa: The untold story, Frontiers in Oncology: Cancer Epidemiology and Prevention. 11 (2021). https://www.frontiersin.org/articles/10.3389/fonc.2021.650117/full (accessed 3 June 2022).

[5] M. Ghanizada, K. K. Jakobsen, J.S.Jensen, I. Wessel, J.F. Tvedskov, C. Grønhøj, C. Buchwald, 2021. The impact of comorbidities on survival in oral cancer patients: a population-based, case-control study, Acta Oncologica. 60:2, 173–179. https://www.tandfonline.com/doi/full/10.1080/0284186X.2020.1836393(accessed 13 June 2022).

[6] American Society of Clinical Oncology, Understanding cancer risk. https://www.cancer.net/navigating-cancer-care/prevention-and-healthy-living/understanding-cancer-risk, 2022 (accessed 29 May 2022).

[7] Cancer Research UK, Can I reduce my cancer risk? https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/cancer-risk-health-quiz, 2020 (accessed 29 May 2022).

[8] World Cancer Research Fund, Cancer health check. https://www.wcrf-uk.org/health-advice-and-support/health-checks/cancer-health-check/?gclid=CjwKCAjwyryUBhBSEiwAGN5OCDqALq0IXJm78bsanppTz9cUcneTYWnICXc8mm0xzbaxY69b9Uj19hoCc5cQAvD_BwE, 2022(viewed 29 May 2022).

[9] A.C. Muller, S. Guido, Introduction to machine learning with python. First ed., Sebastopol: O'Reilly Media, Inc, 2016.

[10] A. Moncada-Torres, M.C. van Maaren, M.P. Hendriks, S. Siesling, G. Geleijnse. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep. 2021 Mar 26;11(1):6968. doi: 10.1038/s41598-021-86327-7. PMID: 33772109; PMCID: PMC7998037. https://pubmed.ncbi.nlm.nih.gov/33772109/ (accessed 29 July 2022).

[11] X. Ye, H. Li, T. Sakurai, P.W. Shueng, Ensemble feature learning to identify risk factors for predicting secondary cancer. Int J Med Sci. 2019 Jun 7;16(7):949-959. doi: 10.7150/ijms.33820. PMID: 31341408; PMCID: PMC6643128. https://pubmed.ncbi.nlm.nih.gov/31341408/ (accessed 18 August 2022).

[12] O.J. Achilonu, J. Fabian, B. Bebington, E. Singh, M.J.C. Eijkemans, E. Musenge, Predicting colorectal cancer recurrence and patient survival using supervised machine learning approach: a South African population-based study. Front Public Health. 2021 Jul 7;9:694306. doi: 10.3389/fpubh.2021.694306. Erratum in:

Front Public Health. 2021 Oct 29;9:778749. PMID: 34307286; PMCID: PMC8292767 Available from: https://pubmed.ncbi.nlm.nih.gov/34307286/ (accessed 22 August 2022).

[13] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis. Cancer Informatics. 2006;2. doi:10.1177/117693510600200030.
https://journals.sagepub.com/doi/10.1177/117693510600200030 (accessed 4 August 2022).

[14]A. Alfayez, H. Kunz, A.G. Lai. Predicting the risk of cancer in adults using supervised machine learning: A scoping review. BMJ Open 2021;11:e047755. doi: 10.1136/bmjopen-2020-047755.
https://doi.org/10.1136/bmjopen-2020-047755 (viewed 29 May 2022).

[15] S. Bundasak, K. Neatpana, J. Jittayanun, Analysis of cancer risk system using decision tree. Conference: International Conference on Business and Industrial Research, 2016. https://www.researchgate.net/publication/342159102_Analysis_of_Cancer_Risk_System_Using_Decision_Tree (viewed 4 June 2022).

[16] Scribbr. Research methods: definitions, types and examples. https://www.scribbr.co.uk/category/research-methods/, 2022 (accessed 15 September 2022).

[17] Data Science Process Alliance, What is CRISP DM? Available from: https://www.datascience-pm.com/crisp-dm-2/, 2022 (accessed 10 April 2022).

[18] M. Khatib, N. Badillo, P. Kahar, D. Khanna, The risk of chronic diseases in individuals responding to a measure for the initial screening of depression and reported feelings of being down, depressed, or hopeless. Cureus. 2021 Sep 1;13(9):e17634. doi: 10.7759/cureus.17634. PMID: 34646682; PMCID: PMC8486358. https://pubmed.ncbi.nlm.nih.gov/34646682/ (accessed 14 September 2022).