

ANALYSES OF LIFESTYLE AND ENVIRONMENTAL FACTORS FOR CANCER  
PREVENTION USING DEEP LEARNING AND CONVENTIONAL  
MACHINE LEARNING FROM UK BIOBANK DATA

by

Jian Dai

---

Copyright © Jian Dai 2020

A Thesis Submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN STATISTICS

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2020

THE UNIVERSITY OF ARIZONA  
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by: Jian Dai

titled: Analyses of Lifestyle and Environmental Factors for Cancer Prevention using Deep Learning and Conventional Machine Learning from UK Biobank Data

and recommend that it be accepted as fulfilling the thesis requirement for the Master's Degree.

*Haiquan Li*

Haiquan Li

Date: Dec 23, 2020

*JOE WATKINS*

JOE WATKINS


Date: Dec 23, 2020

*Lingling An*

Lingling An

Date: Dec 23, 2020

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master's requirement. 

*Haiquan Li*

Haiquan Li

Thesis Committee Chair  
Biosystems Engineering

Date: Dec 23, 2020

ARIZONA

## ACKNOWLEDGEMENTS

I would like to express appreciation to my advisor Dr. Haiquan Li for all his support. His application for access to the UK Biobank data resource, his great idea of using deep learning techniques to analyze data, his critical check for the analysis method, and advice for the thesis greatly helped me to successfully complete this task in Fall semester, 2020. I thank Dr. Joseph Watkins for his important guidelines along my path on my master's degree. I thank Dr. Lingling An for her advice revising my thesis, and in particular, express my appreciation for her counsel and encouragement to pursue a master's degree in statistics which will steer my future career. I thank senior Ph.D. candidate Dillon Aberasturi for his generosity of time helping me in light of his busy schedule, and senior Ph.D. candidate Dailu Chen for her insights in review and revisions of my thesis. I thank our program coordinator Melanie Bowman for her excellent detailed instructions on logistics. I would also like to thank the faculty members in GIDP statistics program who gave me guidance during my study. And finally, I thank everyone in my family for their support, especially, my husband Dr. Richard Carmen who has fully supported me for this degree.

This research has been conducted using data from UK Biobank, a major biomedical database ([www.ukbiobank.ac.uk](http://www.ukbiobank.ac.uk)). The application number is 28979 for the access to UK Biobank data.

## Table of Contents

Abstract.....	6
Chapter 1: Introduction.....	8
1.1 Prevalence of cancers and previous study in environmental factors associated with cancer.	8
1.2 Brief introduction of data resource used in this study .....	9
1.3 Condition, purpose, and approaches for this study .....	10
1.3.1 Condition for this study .....	10
1.3.2 Purpose of this study.....	10
1.3.3 Background of machine learning and deep learning.....	10
1.3.4 Applications of deep learning in healthcare.....	11
1.3.5 Approaches implemented in this study .....	12
Chapter 2: Review of deep learning and conventional machine learning.....	13
2.1 Deep Neural Networks (DNNs).....	13
2.2 Convolutional Neural Networks (CNNs).....	17
2.3 Dropout and early stopping.....	19
2.4 Support vector machines (SVM).....	19
2.5 Random forest and extra trees.....	20
2.6 SelectKBest.....	21
2.7 SMOTE.....	22
2.8 Logistic Regression, ROC Curve, Lasso .....	23
2.8.1 Logistic regression.....	23
2.8.2 ROC Curve.....	23
2.8.3 Lasso Regression .....	25
Chapter 3: Methods.....	27
3.1 Methods overview.....	27
3.2 Data preprocessing.....	29
3.2.1. Raw data processing in Excel and SAS .....	29
3.2.2. Data further processing in Python.....	30
3.2.3 Split data for machine learning .....	30
3.3 Design of the Models.....	30
3.3.1 Deep Neural Networks (DNNs).....	30
3.3.2 Convolutional Neural Networks (CNNs).....	31
3.3.3 Conventional machine learning (SVM, random forest, extra trees, SelectKBest, logistic regression with lasso penalty).....	32
Chapter 4: Data analyses and results.....	34

4.1 Deep Neural Networks (DNNs) for analyzing imbalanced and balanced datasets .....	34
4.2 Convolutional Neural Networks (CNNs) with early stopping for analyzing imbalanced dataset and balanced dataset .....	35
4.3 Support vector machine (SVM) for analyzing both imbalanced dataset and balanced dataset .....	36
4.4 Implement of random forest classifier, extra trees classifier, SelectKBest for analysis and selection of important features .....	37
4.5 Logistic regression analysis and results .....	41
4.5.1 Combining the features selected from random forest classifier, extra trees classifier, SelectKBest to get candidate variables .....	41
4.5.2 Applying the selected variables in logistic regression with lasso penalty .....	42
4.5.3 Important environmental factors related to cancer prediction .....	43
4.6 Comparisons of the prediction accuracy and sensitivity of all the models .....	47
for the balanced dataset.....	47
Chapter 5: Discussion .....	49
5.1 Deep learning techniques offer better performance and prediction .....	49
5.2 Environmental factors associated with cancer .....	50
5.2.1 Risk factors for cancer .....	50
5.2.2 Protective factors for cancer among environmental factors.....	51
5.2.3 Some factors were not found to associate with cancer incidence .....	52
5.3. Lack of interpretability using machine learning .....	53
5.3.1 The reasons causing uninterpretable problem when applying machine learning in healthcare .....	53
5.3.2 Some uninterpretable scenarios in this study .....	53
5.3.3 Promising ideas may be considered to solve the uninterpretable problem .....	53
5.3.3.1 Integrating features may need to be considered.....	54
5.3.3.2 Collaborate with experts for the external knowledge .....	54
5.3.3.3 Automatic explanation is being researched in healthcare area .....	54
Chapter 6: Conclusion.....	55
References.....	57

# Abstract

Cancer may be a fatal illness which brings patients extreme pain and cause death. Cancer is the second leading reason behind death globally. There has been substantial research showing that cancer may be prevented by changing healthy environmental factors. Previous statistical studies investigated each of the risk factors separately, and so it failed to test multiple factors in tandem. The studies which did examine multiple risk factors simultaneously exhibited insufficient sample sizes for the large number of hypothesis tests made therein. With the emergence of enormous biobanks like the United Kingdom Biobank (UK Biobank) in recent years, it's possible to unveil this relationship between environmental factors and cancer occurrence from such big data using cutting-edge statistical and machine learning integration techniques. Deep learning processes data with multi-layer neural networks increases the predictive power. Deep learning techniques have made substantial advances in many domains including healthcare. Clinical application of deep learning has been most rapid in image-intensive fields. However, there wasn't much research reporting applications of deep learning in analyzing environmental factors for cancer prevention.

To explore risk factors among environmental factors associated with cancer incidence for cancer prevention, the data of environmental factors was extracted from the UK Biobank, and two deep learning techniques including deep neural networks (DNNs) and convolutional neural networks (CNNs) were applied to analyze the data. Meanwhile, conventional machine learning techniques including random forest, support vector machine, and logistic regression model were also applied for comparisons. All the accuracies and sensitivities were over 0.92 and 0.90 respectively from the analyses of deep neural networks, convolutional neural networks, support vector machine (SVM), random forest, and logistic regression. Overall, CNNs had the most effective prediction (sensitivity: 0.933; F1 score: 0.961). DNNs had the second-best prediction (sensitivity: 0.933; F1 score: 0.956). Eighty-four important features were selected by machine learning techniques and were analyzed in logistic regression with the lasso penalty model to further reduce the number of features. Twenty-seven important features were further screened and obtained from the logistic regression model. After applying these 27 features selected to logistic regression, the results showed that sensitivity was 0.900, F1 Score was 0.919, and the area under the curve was 0.974. "Age", "Number of operations, self-reported", "Other serious medical condition/disability diagnosed by doctor (Yes)", "Long-standing illness, disability or infirmity (Yes)" were significantly associated with cancer risk (p-values<0.01 or 0.05; odds ratios: 8.79, 1.76, 1.54, 1.33, respectively). Females were significantly more likely to get cancer than males (p-value

<0.01, odds ratio: 1.502). “Number of operations, self-reported” was found to have significant associations with the risk of cancer. However, the factor of “0 self-reported non-cancer illness” was a risk factor for cancer, while the factor of “1 self-reported non-cancer illness” was a protective factor for cancer, which is very interesting for further research and discussion with physiologists with respect to mechanisms. The factors of “Sleep duration of 6 hours”, “Water intake (2 glasses of water daily)”, “Overall health rating (Excellent)” and “Overall health rating (Good)” were found as protective factors for cancer (p-values: <0.01, 0.01, 0.06, 0.08 (close to 0.05); odds ratios: 0.834, 0.878, 0.534, 0.562, respectively). The factor of “Never/rarely sleeplessness /insomnia” was also helpful for cancer prevention (p-value: 0.196; odds ratio: 0.931). The findings in this study will be helpful for initiating early cancer screening and educating the general public about risk factors and protective factors among lifestyle-environmental factors in high risk populations for cancer prevention.

**Key words:** deep learning; DNNs; CNNs; random forest; extra trees; support vector machine; logistic regression; environmental factors; cancer risk

# Chapter 1: Introduction

## 1.1 Prevalence of cancers and previous study in environmental factors associated with cancer

Cancer is a very serious sickness which is likely to bring patients extreme pain and cause mortality. It is the second leading cause of death and is responsible for an estimated death of 9.6 million in 2018 globally. There is about 1 in 6 deaths due to cancer worldwide (World Health Organization, 2018). In 2019, there were more than 1.76 million new cancer cases and still 606,880 estimated deaths in the United States, even with significant breakthroughs in chemotherapy (American Cancer Society, 2019).

There have not been effective medical treatments to cure cancer heretofore. Many factors can cause cancer recurrences (Takeshi et al., 1993; Yamamoto et al., 1996; Aniket et al., 2012; Kim et al., 2018). Surgery is often not able to remove all cancer cells. Quiescent cancer cells continue to stay in patient's bodies and cause cancer's recurrence and resistance to chemotherapy. Hematogenous metastasis was the most common mode of recurrence after surgery (Takeshi et al., 1993), and the metastatic recurrence after surgery remains a major cause of morbidity and mortality (Aniket et al., 2012). Chemotherapeutic agents cause serious side effects and patients' unbearable discomfort to the treatment because they not only kill cancer cells, also kill normal cells eventually. For instance, even two natural compounds purified from *Ganoderma lucidum* which are ergosterol peroxide and ganodermanondiol only preferentially killed breast cancer cells (MCF7) compared to its non-transformed counterpart cells (MCF10A), and killing normal cells is unavoidable (Dai et al., 2017). Therefore, exploring the factors which are associated with cancer occurrence for early prevention would seem to be an ideal strategy.

Previous studies showed that cancer occurrence was related to not only genetic variants but also environmental factors. Amit Sud et al. showed that common genetic variation contributed substantially to the heritable risk of many common cancers. Over 450 genetic variants associated with increased risks have been identified (Amit et al., 2017). Generalized environmental factors including lifestyle factors (e.g., diet, physical activity, medication), health condition, psychosocial factors, social and nature local environment, etc. are also related to cancer risk. Illustrated in the "Cancer Epidemiology" book, environmental factors including physical activity, weight loss, alcohol use, tobacco use, tea drinking, and smoking are related to cancer (Mukesh et al., 2009). It has been estimated that 50% of cancer is preventable due to the substantial effect of modifiable environmental factors on the most prevalent cancers (Mukesh et al., 2009; Chung et al., 2000).

Medication is also a possible factor for cancer prevention. Daily aspirin use at low doses, such as 75-100 mg, may reduce the incidence of all cancers combined (Michael et al., 2012). High intake vitamin C conferred approximately a twofold protective effect compared with low-dose intake (Gladys et al., 1991). Vitamin D intakes have been found to prevent cancer occurrence and reduce case fatality rates by half in patients who had breast, colorectal, or prostate cancer (Cedric et al., 2009). Recent estimates suggest environmental factors may be major risk for cancer up to 95% (Anand et al., 2008). For instance, smoking increases the risk of lung cancer (Peto et al., 2000), and colon cancer (Slattery et al., 2000); alcohol drinking increases the risk of bladder cancer (Wakai et al., 1993), breast cancer, and colon cancer (Boffetta et al., 2006). Physical activity has been shown to reduce the risk of breast cancer and prostate cancer (Frattaroli et al., 2008). Therefore, changing lifestyle habits, social and local environments to improve health conditions could greatly reduce the prevalence of cancers, particularly for those with high genetic susceptibility. Thus, it is important to exhaustively explore environmental factors for cancer prevention and preventing cancer recurrences.

## **1.2 Brief introduction of data resource used in this study**

Previous statistical studies investigated each of the risk factors separately, and they were often underpowered due to their insufficient sample sizes for multiple corrections. Collective studies of the factors were usually not practical due to the difficulties of collecting all the necessary information. It is still elusive how accurately we can predict cancer from these environmental risk factors. With the emergence of large biobanks such as UK Biobank in recent years, it is now possible to unveil this relationship between environmental factors and cancer occurrence from such big data using cutting edge statistical and machine learning integration techniques.

UK Biobank is one of the major international resources of electronic health records. It followed the health information and well-being of 500,000 voluntary participants aged between 40-69 years in 2006-2010 from across the UK. It is composed of comprehensive electronic health records including hospital inpatient diagnoses (ICD-10 coded), primary care records, prescriptions, and other medication records. It also includes cancer and death registries and provides extra information about the cancer origin, types, age of onsets, histology, and death records. More importantly, UK Biobank provides detailed information for each environmental category, such as types and frequencies of physical activity, sleep patterns, smoking, drinking, and various dietary habits, health condition, psychological condition, early life factors, family history, and a variety of natural environmental exposures, such as home location, air pollution, sun exposure (UK Biobank, 2020).

## **1.3 Condition, purpose, and approaches for this study**

### **1.3.1 Condition for this study**

We have the access to UK Biobank data (application number: 28979). We extracted the data of environmental factors including sleep, smoking, diet, alcohol, physical activity, early life style (such as breastfeeding as a baby, maternal smoking around birth), family history (such as illnesses of father, illnesses of mother, father's age at death, mother's age at death), psychosocial factors (such as frequency of friend/family visits, leisure/ social activities, able to confide, mental health), health and medical history (such as pain, state of health, different medication), etc. from the UK Biobank for investigation of the relationship between environmental factors and cancer occurrence in order to provide references for cancer prevention.

### **1.3.2 Purpose of this study**

The purpose of this study is to establish an appropriate model to estimate the probability of cancer based on data from individuals to provide a reference for cancer screening and cancer treatment at early stage, and to explore the environmental factors associated with cancer risk for preventing cancer occurrence and recurrence. We hope the results can act as a reference for the development of new approaches for cancer prevention and diagnosis.

### **1.3.3 Background of machine learning and deep learning**

Machine learning has a long history stretching back decades but only has become applied in industry as recently as in the 1990s. After this time, it quickly became the most popular and most successful subfield of artificial intelligence (AI). Deep learning is a subfield of machine learning. Machine learning, and by extension deep learning, is related to mathematical statistics, but it is different from statistics. Unlike statistics, some extensions of machine learning, such as deep learning, tend to be applied to analyze large, complex datasets for which classical statistical analysis would be impractical. As a result, some methods within the field of machine learning, and especially deep learning, are applied and evaluated sometimes from engineering-based viewpoints, although extensive use of mathematical theory exists throughout most of the field. A machine learning model examines its inputted training data to discover its patterns between features and the studied output in a process commonly called "learning." Once machine learning models have learned the inherent relationship between the data features and the target, they can then be used to perform valuable tasks such as predicting the outcome of other observations and datapoints.

Older and more traditional machine learning techniques such as high-dimensional nonlinear projections (support vector machines (SVMs)) or decision trees are shallow learning only

involving transforming the input data into one or two successive representation spaces, usually through modeling simple patterns between features and target. But more refined and complicated patterns associated with complex problems generally cannot be identified by such techniques. Deep learning takes on learning representations from data through successive layers which continue to increase representations' meaning. Deep learning completely automates to learn all features in one pass. This has greatly simplified machine learning workflows with a single, end-to-end deep learning model (Francois Chollet, 2018). Deep learning processes data with multi-layer neural networks also increases the predictive power. Deep learning offers better performance on many problems (Miotto et al., 2017).

Feature selection techniques based on regression or kernel methods are implemented by calculating the statistical significance of each potential risk factor. These methods primarily select features based on their individual relative predictive capabilities. If they fail to examine the predictive performance of pairs or sets of features, these methods can fail to identify significant features when substantial correlations exist between them. While these shallow models may struggle in finding correlated risk factors, deep learning models can often discover and disentangle latent factors (Suo et al., 2016).

### **1.3.4 Applications of deep learning in healthcare**

Machine learning techniques including deep learning techniques have made substantial advances in many domains over the past decade. In healthcare, global interest in the potential of machine learning has increased (Cabitza et al., 2017). This increase in global interest with healthcare directly relates to the field's commonly large unstructured datasets that possess wide arrays of features and very diverse observations.

Modern biomedical data are often complex, heterogeneous, poorly annotated, and generally unstructured, such as electronic health records (EHR), imaging, omics, sensor data and text. Traditional data mining and statistical learning approaches typically begin with feature engineering so that excellent quality features can be utilized within the subsequent step of analysis. Both the careful construction of these features and the analysis of data can be particularly challenging within dataset if lacking previous information (Miotto et al., 2017). Deep learning is different from traditional machine learning. Deep learning can create accurate models directly from raw data using its multiple layers of neurons. Thus, deep learning provides an effective and efficient method for obtaining end-to-end learning models from complex data (Andrew et al., 2018).

Deep learning has strong appeal for health-related applications, due to its demonstrable strengths in intricate pattern recognition and predictive model building from big high dimensional data sets.

Thus far, clinical application of deep learning has been most rapid in image-intensive fields such as radiology, radiotherapy, pathology, ophthalmology, dermatology, and image-guided surgery (C. David Naylor et al., 2018). The application of deep learning started on image processing of clinical data. Brosch et al. applied deep learning on the analysis of brain Magnetic Resonance Imaging (MRI) scans to predict Alzheimer disease and its variations (Brosch et al., 2013). Gulshan et al. analyzed over 10,000 test images using CNNs to identify diabetic retinopathy in retinal fundus photographs and obtained high sensitivity and specificity (Gulshan et al., 2016). More recently deep learning has been applied to process aggregated EHRs, including both structured data (diagnosis, laboratory tests, medications) and unstructured data (e.g., free-text clinical notes). Deep learning has been applied to predict diseases from the patient clinical status. Cheng et al. applied a four-layer CNN to predict congestive heart failure and chronic obstructive pulmonary disease and showed significant advantages over baselines (Cheng et al., 2016). Choi et al. used RNNs with gated recurrent unit (GRU) to develop Doctor AI model, which used the data of patient history to predict diagnoses and medications for subsequent encounters. The evaluation showed significantly higher recall than shallow baselines (Choi et al., 2016).

### **1.3.5 Approaches implemented in this study**

The applications of deep learning to clinical datasets have been researched, and it will be critical in the future and have significant applications in medicine (Mamoshina et al., 2016). However, applications of deep learning to environmental factors have been rarely reported. This study not only applied deep learning techniques including DNNs and CNNs for analyzing the big dataset from UK Biobank and finding an optimized predicting model for cancer screening at an early stage, but also applied conventional machine learning techniques including random forest, extra trees, support vector machine, and logistic regression for comparisons. After important features being selected, logistic regression was also applied to find risk and protective factors for cancer.

# Chapter 2: Review of deep learning and conventional machine learning

Machine learning belongs to artificial intelligence. It is a subset of AI. Deep learning is part of a family of machine learning methods, including deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks, etc. (Francois Chollet, 2018).

## 2.1 Deep Neural Networks (DNNs)

Deep neural networks (DNNs) include the input layer, multiple hidden layers, and output layer. Figure 1 (Lozano-Diez A et al., 2017) shows the basic structure of DNNs. Mu et al. addressed that DNNs could deal with linear or nonlinear problems by computing the probability of each output layer by layer through appropriate activation function. In other words, inputs are weighted and fed into a series of neurons called a hidden layer, the neurons consist of appropriate activation functions which process the inputs, the subsequent output of hidden layer of neurons can then be fed into subsequent hidden layers until the final layer of neurons provides the target outputs. DNNs are essentially full-connected neural networks. Deep neural network is sometimes also called multi-layer perceptron (MLP) (Mu et al., 2019). Typically, DNNs are Feed Forward Neural Networks (FFNNs) in which data flows from the input layer to the output layer without going backward and the links between the layers are in the forward direction without touching a node again (SPRH LABS, 2019).

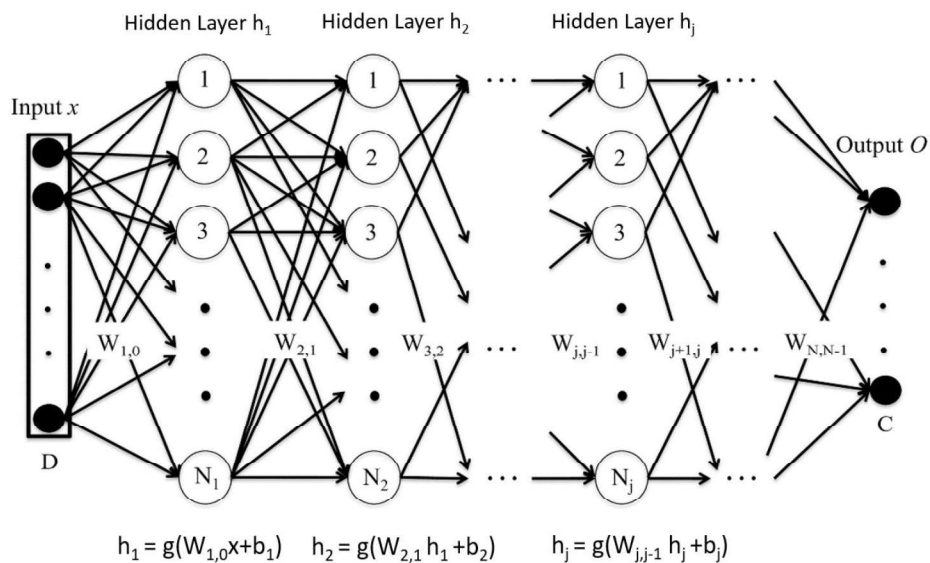


Figure 1: Pipeline of deep neural networks (Lozano-Diez A et al., 2017)

In more details, the input feature vectors are transformed by the hidden layer. The neurons perform a linear transformation on the input through weights and biases (Equation 1). These weights contain the information which are learned by the network from exposure to training data. After the linear transformation of the input, an activation function is applied (Equation 2).

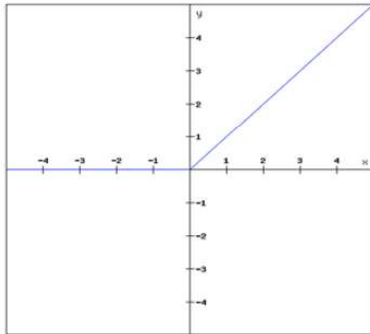
$$x = (\text{weight} * \text{input}) + \text{bias} \quad (1)$$

$$Y = \text{Activation} (\Sigma(\text{weight} * \text{input}) + \text{bias}) \quad (2)$$

Then, the output from the activation function moves to the next hidden layer and the same process is repeated (Dishashree Gupta, 2020). Finally, it gets the classification result. We can add the number of neurons and hidden layers to construct DNNs (Mu et al., 2019).

Two activation functions which are “activation= ‘sigmoid’ ” and “activation= ‘relu’ ”, are related to this study. The Rectified Linear Unit (ReLU) function is nonlinear activation function which has become very popular in the deep learning domain. The ReLU function does not activate all the neurons at the same time. The neurons are deactivated if the output of the linear transformation is less than zero. The plot below can help us understand it better. The result is zero for the negative input values. It indicates that the neuron does not get activated.  $f(x) = \max(0, x)$  is a function for ReLU activation. ReLU function should only be used in the hidden layers (Dishashree Gupta, 2020).

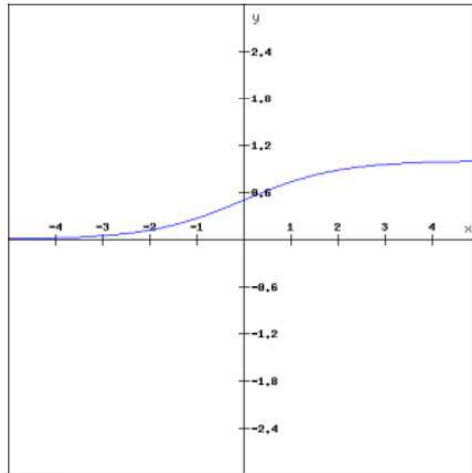
$$f(x) = x, x \geq 0 \\ = 0, x < 0$$



One of the most widely used nonlinear activation functions is Sigmoid function, which is a nonlinear function. Sigmoid transforms the values between the range 0 and 1. The function is as Equation 3 (Dishashree Gupta, 2020):

$$f(x) = 1 / (1 + e^{-x}) \quad (3)$$

$$f(x) = 1/(1+e^{-x})$$



Training a very large deep neural network can be very slow. Therefore, faster gradient descent optimizers are strongly desirable. Some of the most popular faster gradient descent optimizers are Momentum optimization, Nesterov Accelerated Gradient, AdaGrad, RMSProp, and Adam optimization. Optimizer RMSProp and Adam in model compilation are related to this study.

### **RMSProp Optimization**

The equation of RMSProp algorithm is as follows:

RMSProp algorithm equation (4)

$$s \leftarrow \beta s + (1 - \beta) \nabla_{\theta} J(\theta) \otimes \nabla_{\theta} J(\theta)$$

$$\theta \leftarrow \theta - \eta \nabla_{\theta} J(\theta) \oslash \sqrt{s + \epsilon}$$

$\theta$ : weights

$J(\theta)$ : cost function

$\nabla_{\theta} J(\theta)$ : partial derivative regarding weights

$\eta$ : learning rate

$\beta$ : a new hyperparameter called the momentum or decay rate,  $\beta$  must be set between 0 (high friction) and 1 (no friction). Typically, it is set to 0.9.

RMSProp optimizer almost always performs significantly better than AdaGrad. It was the preferred optimization algorithm of many researchers before Adam optimization was developed (Kingma, et al., 2015).

### Adam Optimization

Adam represents adaptive moment estimation. It integrates the ideas of Momentum optimization and RMSProp (Kingma, et al., 2015). Like Momentum optimization, Adam keeps track of an exponentially decaying average of past gradients. Like RMSProp, Adam keeps track of an exponentially decaying average of past squared gradients. The equation (Kingma, et al., 2015) is as follows:

Adam algorithm Equation (5)

$$\mathbf{m} \leftarrow \beta_1 \mathbf{m} + (1 - \beta_1) \nabla_{\theta} J(\theta)$$

$$\mathbf{s} \leftarrow \beta_2 \mathbf{s} + (1 - \beta_2) \nabla_{\theta} J(\theta) \otimes \nabla_{\theta} J(\theta)$$

$$\mathbf{m} \leftarrow \frac{\mathbf{m}}{1 - \beta_1^T}$$

$$\mathbf{s} \leftarrow \frac{\mathbf{s}}{1 - \beta_2^T}$$

$$\theta \leftarrow \theta - \eta \mathbf{m} \oslash \sqrt{\mathbf{s} + \epsilon}$$

Note: T represents the iteration number (starting at 1).

Aurélien Géron addressed the above equations in detail as follow. Steps 1, 2, and 5 of Adam are similar to both Momentum optimization and RMSProp. The only difference for Adam is that step 1 computes an exponentially decaying average rather than an exponentially decaying sum, however, these are equivalent except for a constant factor. Steps 3 and 4 help boost m and s at the beginning of training where the values of m and s will be biased toward 0 at the beginning of training, since m and s are initialized at 0.  $\beta_1$  is the momentum decay hyperparameter, which is typically initialized to 0.9.  $\beta_2$  is the scaling decay hyperparameter, which is usually initialized to 0.999. The smoothing  $\epsilon$  term is often initialized to a tiny number (e.g.,  $10^{-8}$ ) (Aurélien Géron, 2017).

### Binary crossentropy

The loss function used in binary classification tasks was binary crossentropy. The binary crossentropy loss function calculates the loss of an example by computing as the following equation:

$$Loss = - \frac{1}{output\ size} \sum_{i=1}^{output\ size} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \quad (6)$$

where  $\hat{y}_i$  is the  $i$ -th scalar value in the model output,  $y_i$  is the  $i$ -th corresponding target value, the ‘output size’ is the number of scalar values in the model output (Peltarion, 2020). Loss for deep learning in this study is “binary\_crossentropy”.

## 2.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs, or ConvNet) is one of deep neural networks. Figure 2 (Saha S, 2018) shows the structure of CNNs. It employs a mathematical operation called convolution. Convolution is a specialized kind of linear operation. Convolutional networks are simply neural networks that use convolution instead of general matrix multiplication in at least one of their layers (Jeff Heaton, 2018). CNNs are composed of convolutional layers, pooling layers, and fully connected layers except for the input layer and the output layer. These networks are set up such that each input image is to pass through a series of convolution layers with filters (kernels), pooling layers, fully connected layers. Mu et al. addressed CNNs could reduce the complexity and parameters through sharing weights to promote the generalization ability of neural network and reduce neurons through pooling operation to be more robust (Mu et al., 2019).

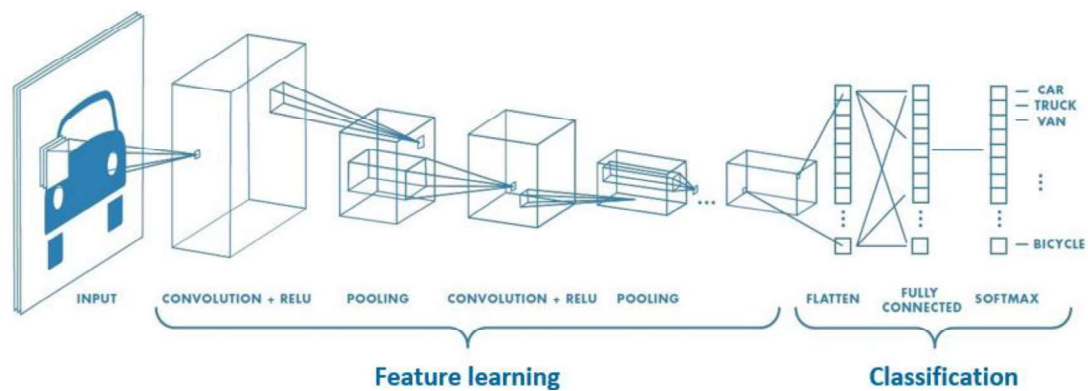
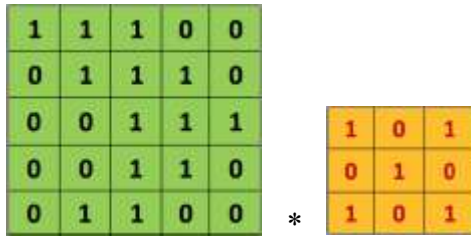


Figure 2: Structure of convolutional neural networks (Saha S, 2018)

The primary purpose of convolution in case of a ConvNet is for features' extraction from the input image. Ujjwal Karn states convolution preserves the spatial relationship between pixels by learning image features using small squares of input data (Ujjwal Karn, 2016).

Figure 3 and 4 (Ujjwal Karn, 2016; Prabhu, 2018) shows the computation. The orange matrix called a 'filter' or 'kernel' was slid over the original image (green) by 1 pixel which is also called 'stride' each time. For every position, we compute element wise multiplication between the two matrices (orange matrix and green matrix), then, sum the multiplication outputs to get the final integer which forms a single element of the output matrix (pink, Figure 4) (Ujjwal Karn, 2016; Prabhu, 2018).



5\*5-image Matrix      3\*3-filter matrix

Figure 3: Image matrix multiplies kernel or filter matrix (Ujjwal Karn, 2016; Prabhu, 2018)

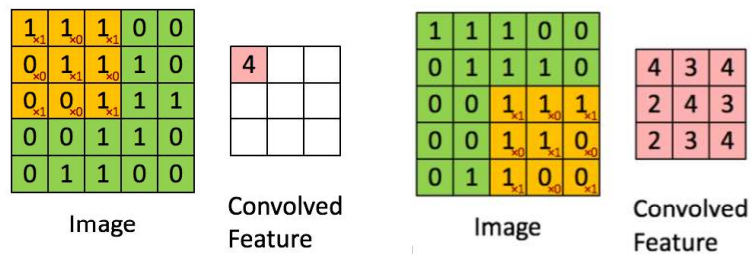


Figure 4: 3 x 3 Output matrix (Ujjwal Karn, 2016; Prabhu, 2018)

After the convolutional layer, the pooling layer is used for reducing the spatial size of the convolved feature. The pooling layer is to decrease the computational power required to process the data through dimensionality reduction. Additionally, it is useful for extracting dominant features which are rotational and positional invariant, thus it maintains effectively training of the model. There are two types of pooling which are maximum pooling and average pooling. Max pooling and average pooling return the maximum value and average value, respectively, from the portion of image covered by the kernel (Sumit Saha, 2018). Figure 5 shows how the max pooling works (Prabhu, 2018).

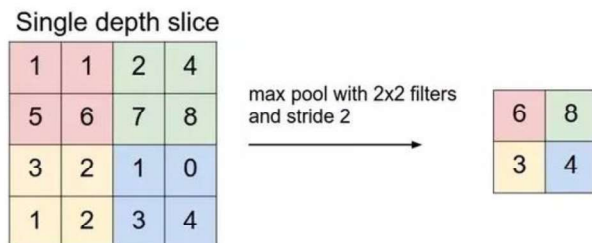


Figure 5: Max pool with 2\*2 filters and stride 2 (Prabhu, 2018)

Kernel\_size determines the dimensions of the kernel which Figure 3 shows an example of 3\*3 kernel size. The default value of strides is set to (1, 1). GeeksforGeeks website explains the details as the following statement. The given Conv2D filter is applied to the current location of the input volume; then, the filter takes a 1-pixel step to the right; again, the filter is applied to the input volume; this is performed until reaching the far-right end border of the volume in which we are moving out filter. Kernel\_initializer is a parameter which controls the initialization method. It is used to initialize all the values in the Conv2D class before training the model.

Kernel\_initializer is the initializer for the kernel weights matrix. It is usually glorot\_uniform by default. Each matrix element in the convolution kernel is the weights which are trained from training dataset. After getting the resulting features which are representations, then, the predicted outputs can be obtained. The predicted outputs can be subsequently used in backpropagation to train the weights in the convolution filter (GeeksforGeeks, 2020).

### **2.3 Dropout and early stopping**

Dropout is a regularization method which approximately train a mass of neural networks with different ways in parallel. Dropout prevents overfitting. It approximately combines exponentially many different neural networks efficiently. This suggests that dropout probably breaks up situations where network layers co-adapt and correct mistakes from prior layers, in order to make the model more robust (Jason Brownlee, 2019).

“Epochs” is a parameter set up in deep neural networks. An epoch refers to one cycle through the full training dataset. Usually, training data is fed into a neural network for more than one epoch in different patterns for a better generalization when given a new validation data. A problem with training neural networks is how many epochs should be set for training. Brownlee suggests that it can lead to overfitting of the training dataset if too many epochs, whereas it may result in an underfit model if too few. The method of early stopping allows us to set up an arbitrary large number of training epochs and stop training once the model performance stops improving on a validation dataset (Jason Brownlee, 2018).

### **2.4 Support vector machines (SVM)**

A support vector machine (SVM) is a very powerful machine learning model. SVMs particularly perform well for classification of complex dataset with small or medium size of the dataset. SVM works by constructing a hyperplane or set of hyperplanes in a high- dimensional or infinite-dimensional space. SVM can be used for performing linear or nonlinear classification, regression, and even outlier detection (Aurélien Géron, 2017). Intuitively, a good separation is achieved by the hyperplane that should have the largest distance to the nearest training data point of any class

which is called functional margin. In general, the larger the margin is, the lower generalization error the classifier has. Figure 6 (Wikipedia, 2020) helps us to understand it better.

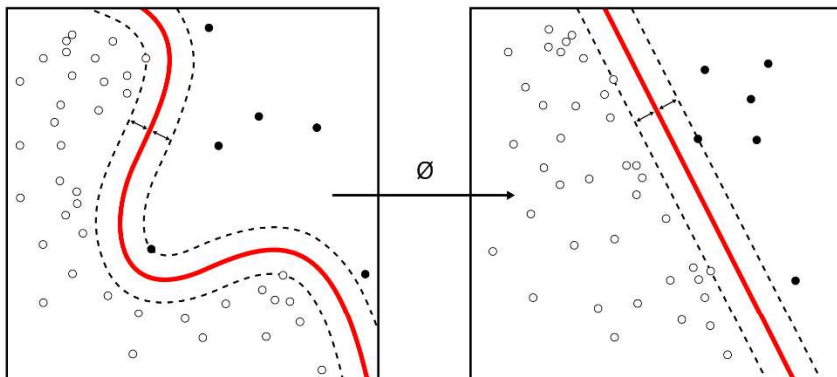


Figure 6 : Demo of support vector machine (Wikipedia, 2020)

## 2.5 Random forest and extra trees

Géron discusses the strategy of random forest and defines that a random forest is an ensemble of decision trees. It's generally trained via the bagging method with maximum samples set which is the size of the training set (Aurélien Géron, 2017). Many individual decision trees operating together comprise a random forest. To classify a new object for a given input, we put the input vector down each of the trees in the forest. Each tree gives a result for classification, which is just like the tree votes for the class. The random forest finally chooses the classification having the most votes among the trees in the forest (Aurélien Géron, 2017). Figure 7 (Misra S et al., 2019) shows how it works.

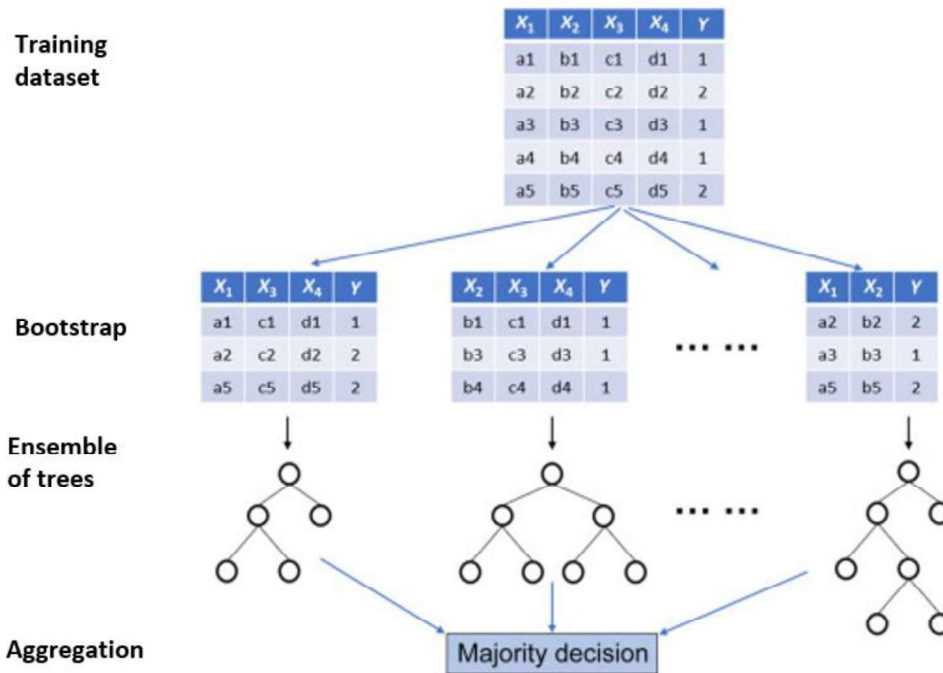


Figure 7: Demo of random forest (Misra S et al., 2019)

The method of extra trees is called an extremely randomized trees ensemble. It trades more bias for a lower variance (Aurélien Géron, 2017). While random forest develops each decision tree from a bootstrap sample of the training dataset, in contrast, extra trees algorithm fits each decision tree on the whole training dataset. The extra trees algorithm randomly samples the features at each split point of a decision tree, which is like random forest. Random forest searches for an optimal cut-point from the randomly chosen features at each node. However, the method of extra trees selects a cut-point at random, which can help to perform faster (Geurts et al., 2006).

## 2.6 SelectKBest

Feature engineering is the process of systematically choosing the set of features in the dataset which are relevant and useful for training data. Irrelevant features often negatively affect the performance of the model by adding extraneous noise with little signal. One of the techniques implemented in Scikit-learn for selecting relevant features from a dataset is to perform statistical tests for selecting the best k features using the SelectKBest module. There are the following statistical tests for selection using with SelectKBest: ANOVA F-value, `f_classif` (classification); Chi-squared stats of non-negative features, `chi2` (classification); F-value, `f_regression`

(regression); mutual information for a continuous target, `mutual_info_regression`. The choice depends if the dataset target variable is numerical or categorical (Ekaba Bisong, 2019).

## 2.7 SMOTE

Chawla et al. explained that synthetic minority oversampling technique (SMOTE) is "an over-sampling approach. In SMOTE, the minority class is over-sampled by creating 'synthetic' examples rather than by over-sampling with replacement." This process of over-sampling is necessary to correct large imbalances between the number of observations in each class. Often any extremely large difference in the size of the classes severely affects the performance of the classifier and results in the trained classifier misclassifying many of the minority class observations in the test dataset (Chawla et al., 2002). The following Figure 8 (DISQUS, 2017) shows how SMOTE works. The minority class is over-sampled by taking samples in each minority class. SMOTE creates line segments joining any/all the  $k$  minority class nearest neighbors. The number of  $k$  nearest neighbors depends on the amount of over-sampling required. Neighbors from the  $k$  nearest neighbors are randomly chosen. New synthetic observations are then generated in the minority class and with features of a randomly selected points along the line segments (Chawla et al., 2002). `Class_weight` usually may cause overfitting. SMOTE can reduce an imbalanced situation.

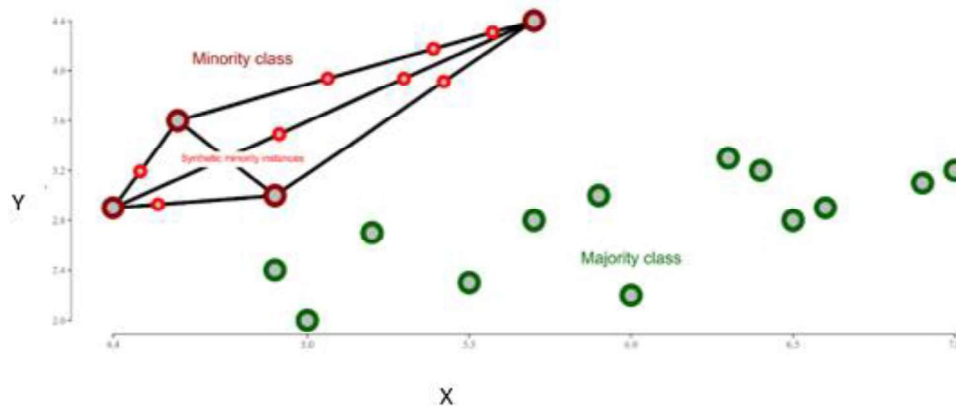


Figure 8: Address class imbalance problems of machine learning via SMOTE: synthesizing new dots between existing dots (DISQUS, 2017).

## 2.8 Logistic Regression, ROC Curve, Lasso

### 2.8.1 Logistic regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine a binary outcome (MedCalc, 2020). The goal of logistic regression is to find the best fitting model to describe the relationship between log odds of the event (presence of the interesting characteristic) and a set of independent (predictor) variables. The formula is as Equation 7 shows. Once a logistic regression model is fitted, it generates the coefficients, its standard errors and significance levels which describes if predictor variables have significant impacts on the odds of presence of the interesting characteristic (MedCalc, 2020).

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (7)$$

$p$ : the probability of presence of the characteristic of interest

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ : regression coefficients

### 2.8.2 ROC Curve

A Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate, which helps diagnostic test evaluation. Figure 9 (Karen GM, 2008) shows a ROC curve. Each point on the ROC curve represents a pair of a sensitivity and (1-specificity) corresponding to a particular decision threshold. The area under the ROC curve (AUC) is to measure how well a predictive model can distinguish between true positive and true negative which usually represent two diagnostic groups (diseased/normal) in healthcare. The best decision rule is high on sensitivity and low on 1-specificity. A test with perfect performance has a ROC curve that passes through the upper left corner where sensitivity is 100% and specificity is 100%. Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test will be (Karen GM, 2008).

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Note: TP: True Positive; FN: False Negative; FP: False Positive; TN: True Negative

Sensitivity is also known as recall, which measures the proportion of positives that are correctly identified. It is calculated as Equation 8 shows. For this study, the sensitivity would be the percentage of cancerous people who were correctly identified as having cancer.

$$\text{Sensitivity} = TP / (TP + FN) \quad (8)$$

Specificity summarizes the proportion of negatives that are correctly identified. It is calculated as Equation 9 shows. For this study, the specificity would be the percentage of non-cancerous people who were correctly identified as not having cancer.

$$\text{Specificity} = TN / (TN + FP) \quad (9)$$

The sensitivity might be more interesting than the specificity for imbalanced classification.

Precision summarizes the proportion of examples assigned the positive class that belong to the positive class. It is calculated as Equation 10 shows (Wikipedia, 2020).

$$\text{Precision} = TP / (TP + FP) \quad (10)$$

Recall measures how well the positives were predicted and is the same calculation as sensitivity.

It is calculated as Equation 11 shows.

$$\text{Recall} = TP / (TP + FN) \quad (11)$$

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset as Equation 12 shows.

$$\text{Accuracy} = (TP+TN) / (TP+TN+ FP +FN) \quad (12)$$

In statistical analysis for binary classification, the F1 score (also F-measure or F-score) is a measure of a test accuracy. It is calculated from the precision and recall of the test as Equation 13 shows.

$$\text{F1 score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (13)$$

The F1 score measures the harmonic mean of the precision and recall. The highest possible value of F1 score is 1, which indicates perfect precision and recall. The lowest possible value of F1 score is 0, and this occurs if either the precision or the recall is zero (Wikipedia, 2020).

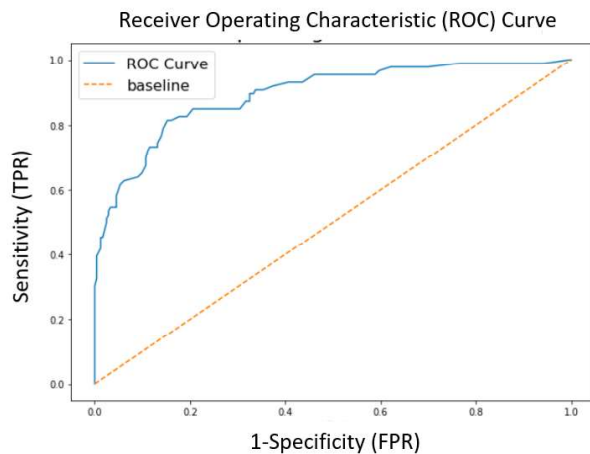


Figure 9: Demo of receiver operating characteristic curve (Karen GM, 2008)

### 2.8.3 Lasso Regression

The Least Absolute Shrinkage and Selection Operator (LASSO), defined by Wikipedia (2020) is “a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.”

Lasso is a penalized least squares method imposing an  $L1$  regularization on the regression coefficients. This means that a penalty equal to the absolute value of the magnitude of coefficients is added to the least squares model. The goal of the algorithm is to minimize Equation 14:

$$\sum_{i=1}^n (y_i - \sum_j x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (14)$$

Minimizing the above Equation 14 is the same as minimizing the sum of squares with constraint  $\sum |\beta_j| \leq s$ .

A tuning parameter,  $\lambda$  controls the strength of the  $L1$  penalty, thereby,  $\lambda$  is the amount of shrinkage of the regression coefficients. When  $\lambda = 0$ , none of the estimated coefficients are eliminated. As  $\lambda$  increases, the absolute values of the estimated coefficients are increasing shrunk, so that more and more coefficients are set to zero and eliminated. Theoretically, when  $\lambda = \infty$ , all coefficients are set to 0 and eliminated. As  $\lambda$  increases, bias increases and variance decreases; as  $\lambda$  decreases, bias decreases and variance increases (Stephanie, 2015).

$L1$  regularization can result in sparse models with few coefficients. This means some coefficients can become zero and be eliminated from the model. Therefore, lasso is useful when simpler

models are desirable (Stephanie, 2015). Lasso regression has been applied for variables' selection in this study.  $L_2$  regularization adds an  $L_2$  penalty, which equals the square of the magnitude of coefficients. One popular example of  $L_2$  regression is ridge regression. All coefficients are shrunk by the same factor, and none of the coefficients are eliminated from the model. Thus, the type of  $L_2$  regularization (e.g. ridge regression) cannot result in elimination of coefficients or sparse models (Stephanie, 2017).

# Chapter 3: Methods

## 3.1 Methods overview

The dataset in this study were extracted from the UK Biobank. It included lifestyle/environmental factors as input and cancer occurrences as the output. We selected 50,000 out of ~ 500,000 observations (participants) and 3,175 columns in the original dataset for this study. Figure 10 shows the overall workflow of our methodology. The study focused on designing the optimal models of deep neural networks and convolutional neural networks by tuning different numbers of neurons, layers and epochs, with the combination of the techniques of dropout and early stopping for different models, to achieve the highest predictive accuracy and lowest loss. Meanwhile, this study also compared deep learning models to conventional machine learning techniques including random forest and extra trees, support vector machine, and logistic regression. The original dataset in this study was highly imbalanced between noncancer class and cancer class in the output. Since the predictive accuracy is known to be affected by the imbalance of training data in supervised machine learning, we applied the synthetic minority oversampling technique to generate balanced dataset and mitigate the issue. The techniques of SelectKBest and lasso penalty were applied to select important features. We applied various programming language to analyze the data, such as Python, SAS, and SQL.

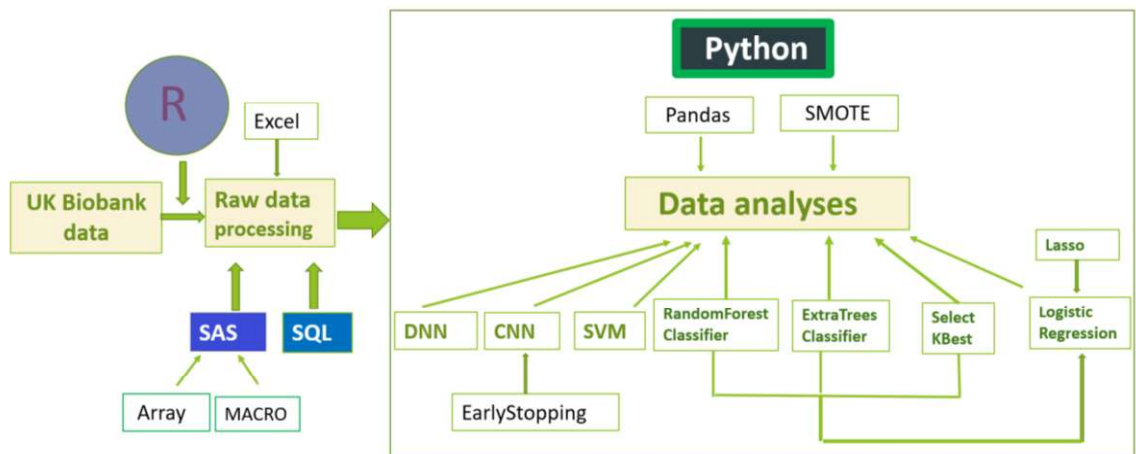


Figure 10: The overall workflow of our study, including data extraction, raw data manipulation, deep learning (DNN, CNN), conventional machine learning including random forest, extra trees, SVM, SelectKBest, and logistic regression.

The workflow of the experiments was as follows:

1) Data preprocessing

- Extract raw data from UK Biobank database.
- Preprocess data including removing rows and columns with excessive missing values, combining features with multiple instances, standardizing data, etc.
- Apply SMOTE to balance the cancer and noncancer classes and generate a new balanced dataset.
- Split the analysis dataset into training, validation, testing datasets as 75%, 12.5%, 12.5%, respectively.

2) Deep learning model building

- Build model sequential of deep neural networks model, train the model with training dataset and validation dataset, test the model with testing data, optimize hyper-parameters, and get the model with highest predictive accuracy value and lowest loss value.
- Build model sequential of convolutional neural networks model, train the model with training dataset and validation dataset, test the model with testing dataset, optimize hyper-parameters, and get the model with highest predictive accuracy value and lowest loss value.

3) Conventional machine learning implementations and results comparison

- Apply SVM to analyze the data and get the predictive accuracy from the model.
- Apply random forest to analyze the data, get the predictive accuracy from the model, get each of most important 30 features from the original imbalanced and balanced datasets.
- Apply extra trees to analyze the data, get the predictive accuracy from the model, get each of most important 30 features from the original imbalanced and balanced datasets.
- Apply SelectKBest to analyze the data and get each of most important 30 features from the original imbalanced and balanced datasets.
- Combine all features from the 6 groups with each having 30 important features selected by random forest, extra trees and SelectKBest from balanced and imbalanced datasets together without duplicated features, get 84 unique important features.
- Apply these 84 unique important features in logistic regression with lasso penalty to get the 27 most important features related with cancer occurrence.
- Compare the results analyzed by different methods and interpret the results.

## 3.2 Data preprocessing

### 3.2.1. Raw data processing in Excel and SAS

The dataset was extracted from the UK Biobank with lifestyle/environmental factors as features and the output (target) being “Cancer diagnosed by doctor”. We chose 50,000 rows and 3,175 columns for our study. Each row represents a participant with a unique de-identified ID. Each column represents a UK biobank data field or an instance of a field to be investigated as multiple instances (repeat measurement) for a data field if possible. The fields with multiple instances were generated by data collection in different stages of the project. These fields were preprocessed into one or two instances by taking average, minimum, maximum values according to their specific implications, in order to analyze data easily. Features with excessive missing values (over 50%) were removed from the dataset. Rows with more than 9 missing numbers were removed. The fields correlated with the output (i.e., f\_2453: cancer diagnosed by doctor) were removed from the dataset, such as “number of self-reported cancers”. After removing missing values, combining multiple instances for some fields, and removing the columns correlated with the output, 147 features were left, which included 116 categorical and 31 numerical variables in the input dataset (X). These 116 categorical variables were transformed into dummy variables; the 31 numerical variables were standardized by SAS with the mean equal to 0, standard deviation equal to 1.

The output variable Y (No.2453 field) had 4 categories of values. They were coded as follows according to the UK Biobank:

1 Yes

0 No

-1 Do not know

-3 Prefer not to answer

The values of coding number “-1” and “-3” meant “Do not know” and “Prefer not to answer”, respectively. They were not meaningful for this study. Thus, the observations of these values were excluded. Thereby, the final Y output only had value 1 (Yes) and 0 (No), from the field of “Cancer diagnosed by doctor”, which resulted a binary classification problem.

After the raw data processing, the final input dataset X for machine learning had 1,433 columns of the input features and 40,548 rows of the observations. The output Y had 1 column and 40,548 observations. The input dataset for final analysis included relatively comprehensive data of common lifestyle factors and environmental exposures, such as baseline characteristics, body size measures, physical activity, diet, alcohol intake, smoking, early life factors, family history, sleep,

mental health, general health, medical conditions, blood pressure, ethnicity, and residential air pollution.

### **3.2.2. Data further processing in Python**

To use machine learning package in Python, the dataset was imported to Python for further cleaning up. For example, usually only people aged 40-69 years were recruited into the project in the UK Biobank between 2006-2010, and people usually cannot live over 100 years, so, the observations were deleted with the age over 100 years old. Then, the dataset was split into output dataset (Y, f\_2453 feature) and input dataset (X, the rest of features). A new table with index and variables' names was created for final checking.

### **3.2.3 Split data for machine learning**

The training, validation, test datasets were created by randomly splitting both X and Y datasets into 75%, 12.5%, 12.5%, respectively, according to the following two steps. First, split the datasets into 75% versus 25%. The random seed was set as 42. Second, split 25% dataset into 50% versus 50%. The random seed was also set as 42. The training datasets were named as X\_train and Y\_train. The validation datasets were named as X\_val and Y\_val. The test independent datasets were named as X\_test and Y\_test.

Imbalanced data were converted into balanced data through the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002). Because these datasets had seriously imbalanced outcome for the values of 1 and 0, good results of predictive accuracy could hardly be obtained by analyzing the imbalanced data. Therefore, the datasets needed to be converted into balanced datasets. However, class\_weight adjustment for imbalanced datasets often causes overfitting. SMOTE was applied for transforming these imbalanced training, validation, and testing datasets into balanced datasets accordingly.

## **3.3 Design of the Models**

### **3.3.1 Deep Neural Networks (DNNs)**

The analysis workflow for deep neural networks followed these steps: import all the packages used in the program; created model sequential; define model compile; use the training datasets and validation datasets to train model; use the test datasets to test the accuracy and loss value of the model. In this study, I set 1,433 neurons for the input layer based on the number of features of the input dataset. There were 3 hidden layers, which had 928, 128, 16 neurons, respectively, based on tuning results. The node number of output layer was 1 since it was a binary classification. ReLU was the activation function in all layers, except the last layer which used sigmoid activation function. I compared the results of RMSProp optimization with Adam

optimization (Kingma, et al., 2015). Switching between the two methods of optimization had little effect on the performance of the model. Therefore, RMSProp optimization were used in DNNs. The parameter epochs equal to 3, 7, 10, 15 were experimented. The code for the pipeline of deep neural networks is shown as Table 1.

Table 1: Pipeline of deep neural networks

---

```

model=Sequential()
model.add(Dense(928, activation= "relu", input_shape=(1433,)))
model.add(Dense(128,activation= "relu"))
model.add(Dense(16,activation= "relu"))
model.add(Dense(1,activation= "sigmoid"))
model.summary()
model.compile(optimizer= "rmsprop",
              loss= "binary_crossentropy",
              metrics=["accuracy"])
history=model.fit(X_train,Y_train,
                 batch_size=128,
                 epochs=7,
                 verbose=2,
                 validation_data=(X_val,Y_val))
score=model.evaluate(X_test,Y_test,verbose=0)
predictions_DnnImb=model.predict_classes(X_test)
confusion_matrix(predictions_DnnImb,Y_test)

```

---

### 3.3.2 Convolutional Neural Networks (CNNs)

This CNNs model included 10 layers. The kernel\_size that determines the dimensions of the kernel was set to (1, 3). The numbers of filters for the first, third and fifth layer, were 32, 64, and 64, respectively. Max pooling was used and the pool\_size was set to (1,2). To avoid overfitting, dropout techniques were applied in two layers. The proportion of dropout was 0.25 for the two dropout functions. To find the optimal epochs automatically, early stopping technique was also used by the kernel.Tf.keras.callbacks.EarlyStopping function. The author imported all the packages for this CNNs model (such as Conv2D, MaxPooling2D, Early Stopping, Dropout), and then created model Sequential, defined model compile, and defined monitor for early stopping.

ReLU was the activation function in all layers, except the last layer which used the sigmoid activation function. The detailed model sequential and pipeline are shown in Table 2.

Before using the data to train the model defined, the shape of data needs to be reshaped to fit the structural requirement of CNNs training. So, the datasets of 2 dimensions were reshaped into the datasets of 4 dimensions for all the training, validation, and test datasets. After the datasets were reshaped, the training dataset from the original imbalanced input dataset became (30411, 1, 1433, 1), the shape of the test dataset from the original imbalanced input dataset became (5069, 1, 1433, 1), the shape of the validation dataset became (5068, 1, 1433, 1).

The same CNNs model was applied to train balanced datasets transformed by SMOTE. After reshaping the input datasets, the shapes of the training, test, and validation datasets became (55774, 1, 1433, 1), (9254, 1, 1433, 1), and (9334, 1, 1433, 1), respectively.

### **3.3.3 Conventional machine learning (SVM, random forest, extra trees, SelectKBest, logistic regression with lasso penalty)**

Conventional machine learning including SVM, random forest, extra trees and logistic regression with lasso penalty were also applied to analyze the data for comparisons, including the original imbalanced dataset and balanced dataset after transformed by SMOTE. Since this is a large dataset, `sklearn.svm.LinearSVC` function was used for SVM analysis.

To find the important features in X dataset which could efficiently influence the outcome of output Y, 3 methods of random forest classifier, extra trees classifier, and SelectKBest were applied respectively to select 30 important features from both original imbalanced dataset and balanced dataset. Then, all the selected important features were combined without taking duplicates and applied in logistic regression with lasso penalty. Since there was a large number of life/environmental factors in the logistic model, ridge regression was not appropriate for removing relatively unimportant features in this study and so was not used. We applied Lasso penalty instead to select important features.

Table 2: Pipeline of convolutional neural networks

---

```

model=Sequential()
model=models.Sequential()
model.add(layers.Conv2D(32,(1,3),activation= "relu",input_shape=(1,1433,1)))
model.add(layers.MaxPooling2D(pool_size=(1,2)))
model.add(layers.Conv2D(64,(1,3),activation= "relu"))
model.add(layers.MaxPooling2D(pool_size=(1,2)))
model.add(layers.Conv2D(64,(1,3),activation= "relu"))
model.add(Dropout(0.25))
model.add(layers.Flatten())
model.add(layers.Dense(128, activation= "relu"))
model.add(Dropout(0.25))
model.add(layers.Dense(1,activation= "sigmoid"))
model.summary()
model.compile(loss= "binary_crossentropy",
              optimizer= "adam",
              metrics=["accuracy"])
monitor = EarlyStopping(monitor= 'val_loss', min_delta=1e-3, patience=5, verbose=1,
mode='auto',restore_best_weights=True)
hist=model.fit(X_train_shaped,Y_train,
              # batch_size=25,
              callbacks=[monitor],
              verbose=1,
              epochs=100,
              validation_data=(X_val_shaped,Y_val))
score=model.evaluate(X_test_shaped,Y_test,verbose=0)

```

---

# Chapter 4: Data analyses and results

## 4.1 Deep Neural Networks (DNNs) for analyzing imbalanced and balanced datasets

We applied DNN using parameters described in Section 3.3.1. After the parameter epochs equal to 3, 7, 10, 15 being experimented, epoch 7 yielded the best results by this DNNs model for balanced dataset and imbalanced dataset.

First, the DNNs model constructed was applied to analyze original imbalanced datasets. The shapes (rows and columns) of training, validation, test datasets were (30411, 1433), (5068, 1433), and (5069, 1433), respectively. The loss of the model for the test dataset was 0.315 and the accuracy was 0.924. Although it could predict noncancer cases well, it could not predict cancer cases well. The sensitivity was only 0.283 for the imbalanced test dataset and the precision was 0.644. See Table 3 for details of the confusion matrix (or contingency table) of the classifier.

The same DNNs model structure was applied to train balanced dataset transformed by SMOTE. The training, validation, and test datasets were (55774, 1433), (9334, 1433), and (9254, 1433), respectively. The loss of the model for the test dataset was 0.303 and the accuracy was 0.957. It not only could predict noncancer cases well, but it also predicted cancer cases well. The sensitivity, which for this study corresponds to the accuracy of cancer prediction among true cancer cases, was 0.933. The testing precision was 0.981. F1 score was calculated as equation 13. The F1 score was 0.956. Table 4 shows the details of the testing results.

Table 3: Confusion matrix of the imbalanced test dataset using DNNs model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4558 (TN)	317 (FN)
	Cancer (1)	69 (FP)	125 (TP)

Table 4: Confusion matrix of the balanced test dataset using DNNs model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4543 (TN)	312 (FN)
	Cancer (1)	84 (FP)	4315 (TP)

## 4.2 Convolutional Neural Networks (CNNs) with early stopping for analyzing imbalanced dataset and balanced dataset

The imbalanced datasets were used to train the CNNs model defined in Section 3.3.2. From the training results, 10 was determined to be the optimal number of epochs for early stopping. After the model was trained, the imbalanced test dataset was used to test the model accuracy. The loss of the model for the imbalanced test dataset was 0.200 and the accuracy was 0.931 (Table 5).

Although it could predict noncancer cases well, it could not predict cancer cases well. The sensitivity was only 0.287 for the imbalanced test dataset, which means it had difficulty to predict cancer among true cancer cases. The precision was 0.794 (Table 5).

From the training results of balanced data using this CNNs model, the optimal number of epochs for early stopping was 12. The loss of the model from the balanced test dataset was 0.111, and the accuracy was 0.962. It could predict both noncancer cases well and cancer cases very well. The sensitivity was 0.933. The precision was 0.991, and the F1 score was 0.961 (Table 6).

Figure 11 shows the number of epochs was equal to 12 yielding the optimal loss and accuracy among training and validation datasets.

Table 5: Confusion matrix of the imbalanced test dataset using CNNs model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4594 (TN)	315 (FN)
	Cancer (1)	33 (FP)	127 (TP)

Table 6: Confusion matrix of the balanced test dataset using CNNs model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4586 (TN)	312 (FN)
	Cancer (1)	41 (FP)	4315 (TP)

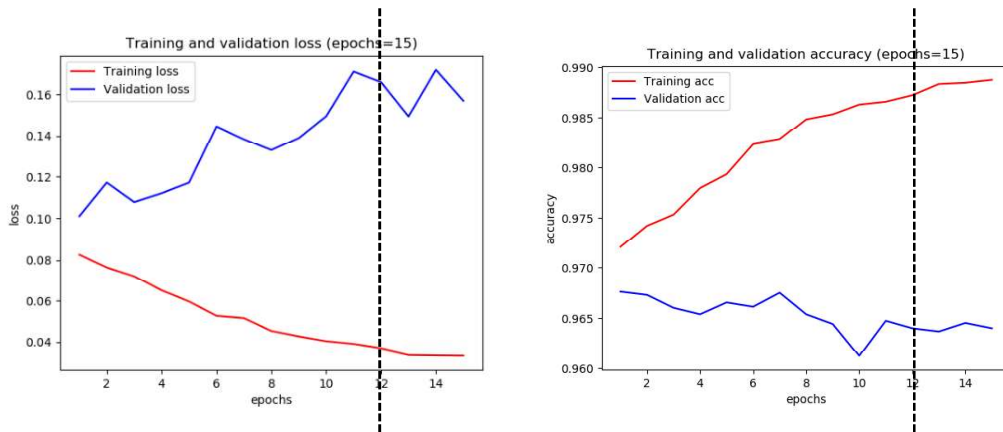


Figure 11: The loss and accuracy of training and validation

### 4.3 Support vector machine (SVM) for analyzing both imbalanced dataset and balanced dataset

Support Vector Machine (SVM) was applied to analyze original imbalanced dataset as described by Section 3.3.3. The accuracy of the imbalanced test dataset was 0.927, and the precision of test dataset was 0.987. However, it only could predict noncancer cases well, but not cancer cases. The testing sensitivity was only 0.170. Table 7 shows the details of the confusion matrix of the classifier.

Table 7: Confusion matrix of the imbalanced test dataset using SVM model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4626 (TN)	367 (FN)
	Cancer (1)	1 (FP)	75 (TP)

Support Vector Machine (SVM) was applied to analyze balanced test dataset. The accuracy of test dataset was 0.959, and the precision was 1.0. It could predict both cancer cases and noncancer cases well. The sensitivity was 0.918. Table 8 shows the details of the confusion matrix.

Table 8: Confusion matrix of the balanced test dataset using SVM model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4626 (TN)	378 (FN)
	Cancer (1)	1 (FP)	4249 (TP)

From the comparisons of the results of imbalanced dataset and balanced dataset analyzed by DNNs, CNNs, SVM (Table 9), we could conclude that the balanced dataset transformed by SMOTE had much better results for predictions than the original imbalanced data.

Table 9: Comparisons of prediction accuracy and sensitivity using DNN, CNN and SVM for imbalanced data and balanced data

Model	Data type	Accuracy	Sensitivity (Recall)	Specificity	Precision	F1 score
DNNs	Imbalance data	0.924	0.283	0.985	0.644	0.393
	Balanced data	0.957	0.933	0.982	0.981	0.956
CNNs	Imbalance data	0.931	0.287	0.993	0.794	0.422
	Balanced data	0.962	0.933	0.991	0.991	0.961
SVM	Imbalance data	0.927	0.170	1.000	0.987	0.290
	Balanced data	0.959	0.918	1.000	1.000	0.957

#### 4.4 Implement of random forest classifier, extra trees classifier, SelectKBest for analysis and selection of important features

Random forest classifier model was set up as described in Section 3.3.3, then the X training dataset was used to train the model. Similarly, as aforementioned, using the original imbalanced to train the model, the results were not good for predicting cancer cases. But using the balanced dataset to train the model, better results of predictions were obtained. Table 10 shows the confusion matrix of test dataset. The testing accuracy of prediction was 0.948, the sensitivity was 0.896. The precision was 1. The top 30 importance features were selected from both imbalanced dataset and balanced dataset. Figure 12 shows the 30 most important features from balanced dataset by random forest classifier model. Table 11 shows the feature's IDs and names.

Table 10: Confusion matrix of the balanced test dataset using random forest model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4627 (TN)	479 (FN)
	Cancer (1)	0 (FP)	4148 (TP)

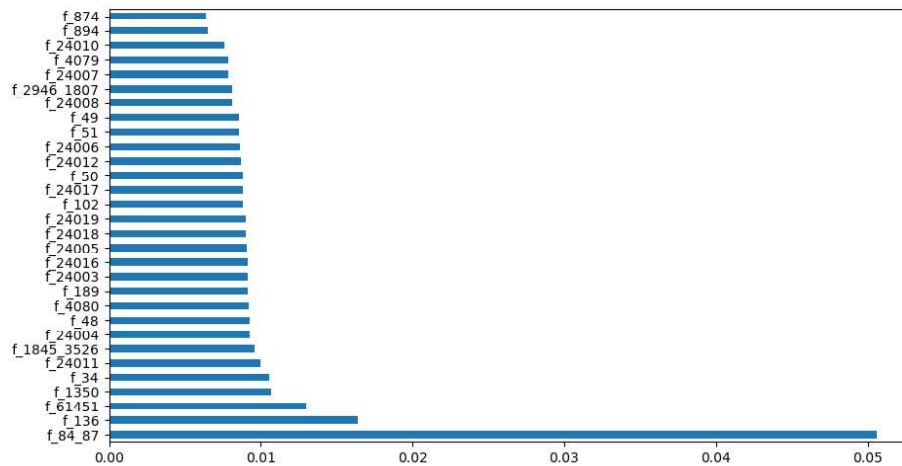


Figure 12: 30 most important features selected by random forest classifier from the balanced dataset

An extra trees classifier model was applied as described in Section 3.3.3, and then the X training dataset was used to train the model. Using the original imbalanced data to train the model, the extra trees classifier only could predict noncancer cases well, but not cancer cases. When using the balanced dataset to train the model, it could predict both noncancer cases and cancer cases well. Table 12 shows the confusion matrix of test dataset. The overall testing accuracy of predictions was 0.951, the sensitivity of prediction for cancer cases was 0.902. The precision was 1. The top 30 importance features were selected from both imbalanced data and balanced data. Figure 13 shows the 30 most important variables from the balanced dataset by extra trees classifier model.

The author applied SelectKBest class to select the top 30 important features. The chi-square method, which is used in many feature selection problems, could not be applied in this dataset because numerical variables had been standardized and there were negative values. Thus, ANOVA F-value was appropriate. The author applied `sklearn.feature_selection.f_classif` for SelectKBest in each of original imbalanced dataset and the balanced dataset. The features of top 30 importance values were selected from both imbalanced dataset and balanced dataset by SelectKBest using the ANOVA F-value as the scoring function. Figure 14 shows the 30 most important features selected by SelectKBest from balanced dataset.

Table 11: Important environmental factors selected by random forest, extra trees and SelectKBest from balanced dataset

---

Field code	Description of lifestyle factor
f_84_87	Age
f_1350	Number of self-reported non-cancer illnesses (0 self-reported non-cancer illness)
f_136	Number of operations, self-reported
f_13092	Fresh fruit intake ( Quantity level 2 daily )
f_12001	Sleeplessness / insomnia (Never/rarely)
f_20107_08	Illnesses of father (High blood pressure)
f_11606	Sleep duration (6 hours)
f_61450D07	Illness, injury, bereavement, stress in last 2 years (None of the above)
f_201270D2	Neuroticism score (missing data)
f_1351	Number of self-reported non-cancer illnesses (1 self-reported non-cancer illness)
f_61453	Illness, injury, bereavement, stress in last 2 years (Death of a close relative)
f_21880	Long-standing illness, disability or infirmity (No)
f_24730	Other serious medical condition/disability diagnosed by doctor (No)
f_18351	Mother still alive
f_15380	Major dietary changes in the last 5 years (No)
f_21781	Overall health rating (Excellent)
f_15382	Major dietary changes in the last 5 years (Yes, because of other reasons)
f_1370	Number of treatments/medications taken (0 treatments/medications taken )
f_1371	Number of treatments/medications taken (1 treatments/medications taken )
f_9432	Frequency of stair climbing in last 4 weeks (6-10 times a day)
f_10702	Time spent watching television (TV) (2 hours/day)
f_11002	Drive faster than motorway speed limit (Sometimes)
f_1371	Number of treatments/medications taken (1 treatments/medications taken )
f_9432	Frequency of stair climbing in last 4 weeks (6-10 times a day)
f_10702	Time spent watching television (TV) (2 hours/day)
f_12101	Snoring (Yes)
f_12991	Salad / raw vegetable intake (1 tablespoon/day)
f_13190	Dried fruit intake (0 pieces/day)
f_13291	Oily fish intake (Less than once a week)
f_13391	Non-oily fish intake (Less than once a week)
f_16180D06	Alcohol usually taken with meals (It varies)
f_16770	Breastfed as a baby (No)
f_16871	Comparative body size at age 10 (Thinner)
f_16973	Comparative height size at age 10 (About average)
f_17871	Maternal smoking around birth (Yes)
f_19201	Mood swings (Yes)
f_19401	Irritability (Yes)

---

Table 11 (continued).

f_61451	Illness, injury, bereavement, stress in last 2 years (Serious illness, injury or assault to yourself)
f_61529	Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor (Hayfever, allergic rhinitis or eczema)
f_61621	Types of transport used (excluding work) (Car/motor vehicle)
f_874	Duration of walks (minutes/day)
f_894	Duration of moderate activity (minutes/day)
f_34	Year of birth (years)
f_48	Waist circumference (cm)
f_49	Hip circumference (cm)
f_50	Standing height (cm)
f_51	Seated height (cm)
f_102	Pulse rate, automated reading (bpm)
f_189	Townsend deprivation index at recruitment
f_4079	Diastolic blood pressure, automated reading (mmHg)
f_4080	Systolic blood pressure, automated reading (mmHg)
f_2946_1807	Father's age
f_1845_3526	Mother's age
f_24003	Nitrogen dioxide air pollution; 2010 (micro-g/m3)
f_24004	Nitrogen oxides air pollution; 2010 (micro-g/m3)
f_24005	Particulate matter air pollution (pm10); 2010 (micro-g/m3)
f_24006	Particulate matter air pollution (pm2.5); 2010 (micro-g/m3)
f_24007	Particulate matter air pollution (pm2.5) absorbance; 2010 (per-metre)
f_24008	Particulate matter air pollution 2.5-10um; 2010 (micro-g/m3)
f_24010	Inverse distance to the nearest road (1/metres)
f_24011	Traffic intensity on the nearest major road (vehicles/day)
f_24012	Inverse distance to the nearest major road (1/metres)
f_24016	Nitrogen dioxide air pollution; 2005 (micro-g/m3)
f_24017	Nitrogen dioxide air pollution; 2006 (micro-g/m3)
f_24018	Nitrogen dioxide air pollution; 2007 (micro-g/m3)
f_24019	Particulate matter air pollution (pm10); 2007 (micro-g/m3)

Table 12: Confusion matrix of the balanced test dataset using extra trees model

		Actual Values	
		Noncancer (0)	Cancer (1)
Predicted Values	Noncancer (0)	4627 (TN)	455 (FN)
	Cancer (1)	0 (FP)	4172 (TP)

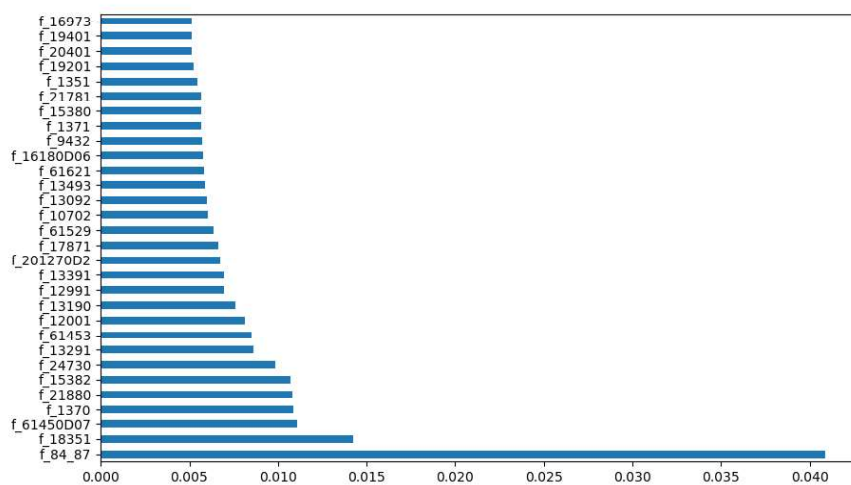


Figure 13: 30 most important features selected by extra trees classifier from the balanced dataset

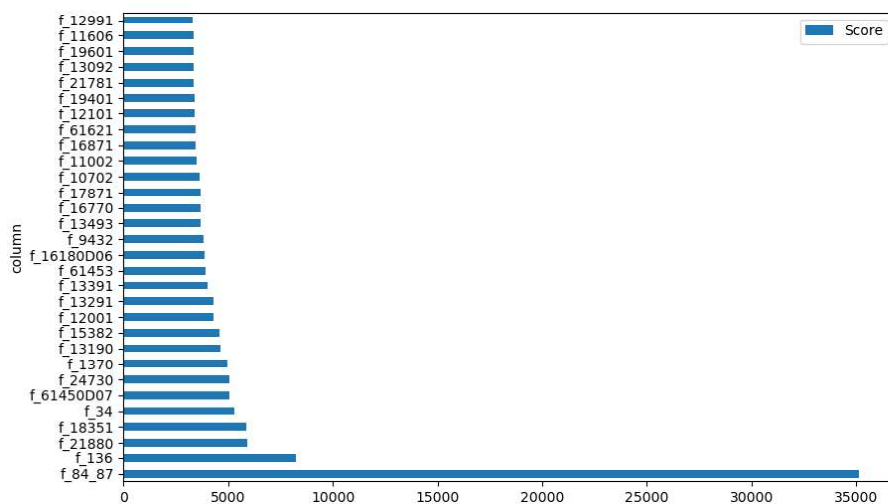


Figure 14. 30 most important features selected by SelectKBest from the balanced data

## 4.5 Logistic regression analysis and results

### 4.5.1 Combining the features selected from random forest classifier, extra trees classifier, SelectKBest to get candidate variables

The author combined the 30 most important features selected by three methods of random forest, extra trees, and SelectKBest from the two datasets (the original imbalanced dataset and the balanced dataset by SMOTE). In total, there were 6 groups with each group having 30 features.

The total number of selected features were 180. Eighty-four important features were uniquely selected after merging all these 180 variables together without duplicates using SAS.

These 84 variables were extracted from input X dataset to get a subset whose dimensions were (40548, 84). Then, SMOTE was used to balance the number of noncancer and cancer cases for this subset. The resulting dataset was split randomly into training dataset and test dataset with proportion of observations between them set as 75%:25%, Within the code, random\_state was set as 42.

#### 4.5.2 Applying the selected variables in logistic regression with lasso penalty

The author applied SelectFromModel function for logistic regression model with lasso penalty. Some features with low coefficients had been removed from the model by lasso. Get\_support() function was used in the code to get the selected features from the trained model. Twenty-seven features were selected, and 57 features were dropped from the logistic regression with a lasso penalty.

Finally, the author applied these 27 features into logistic regression again. After the logistic regression model was trained by this X training dataset with 27 selected variables, the author applied this model to the test datasets with these 27 selected features. The test accuracy for general prediction was 0.920, and the precision was 0.937. The recall (or sensitivity) for predicting cancer cases was 0.900. The F1 Score was 0.919 and the area under the curve (AUC) was 0.974 (Figure 15). All these results demonstrated that these selected important 27 features played important roles for cancer prediction.

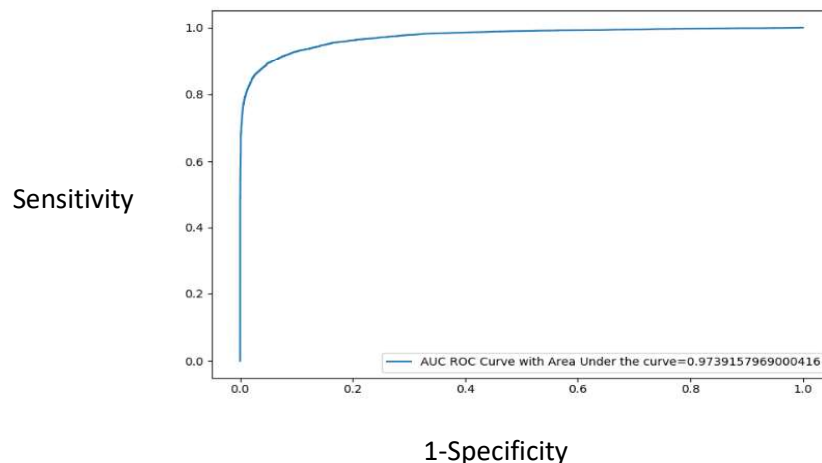


Figure 15: AUC ROC curve from logistic regression analyzing the final 27 important selected features

### 4.5.3 Important environmental factors related to cancer prediction

After applying 84 important features selected in logistic regression with lasso, 27 important features were obtained. These 27 important features as predictor variables were analyzed by logistic regression with “Cancer diagnosed by doctor” as the response variable using SAS. Eleven lifestyle/environmental factors were significantly associated with incidence of cancer, and three lifestyle/environmental factors were strongly associated with incidence of cancer. Tables 13 and 14 show maximum likelihood estimates, odds ratio estimates of the environmental factors which had significant or close to significant associations with incidence of cancer.

From Table 13 and Table 14, we can see there is a statistically significant association between age and incidence of cancer since p-value was less than 0.0001, which is widely known. The participants who were one year older had 8.788 times the odds of cancer than those who were one year younger among the subjects with middle or old age in the UK Biobank data. All else being equal, older people were more likely to get cancer than younger people. This indicates that the aged population is a more vulnerable group for cancer occurrence. The factor “Number of operations, self-reported” had a statistically significant association with incidence of cancer, since its p-value was less than 0.0001. The participants whose number of operations was increased 1 had 1.763 times the odds of cancer than participants whose number of operations was not increased 1. This means that the participants who had a larger number of operations were more likely to get cancer. There was a statistically significant association between “Other serious medical condition/disability diagnosed by doctor ” and incidence of cancer since p-value is 0.01. The participants who had “Other serious medical condition/disability diagnosed by doctor” had 1.541 times the odds of cancer than the participants who did not have any other serious medical condition/disability diagnosed by doctor. There was a statistically significant association between gender and incidence of cancer (p-value <0.0001). Females had 1.502 times the odds of cancer than males, probably due to enrichment of breast cancers in the biobank. The data were collected in England which is a highly developed country. In less developed countries, the odds of females developing cancer may not be higher than that of males. The factor of “Long-standing illness, disability or infirmity (Yes)” was found to have a statistically significant association with incidence of cancer, since its p-value is 0.0467, which is less than 0.05. The participants with “Long-standing illness, disability or infirmity” had 1.333 times the odds of cancer than those without “long-standing illness, disability, or infirmity.” Field ID f\_135 represented the number of self-reported noncancer illnesses. This factor was regarded as categorical variables and had been transformed into a series of dummy variables. So, f\_1350 meant whether the participants had 0 self-reported illness or not, f\_1351 meant whether the participants had 1 self-reported illness or

not. It was interesting that the factor of “0 self-reported non-cancer illness” was significantly associated with incidence of cancer since its p-value was less than 0.0001. The participants who had “0 self-reported non-cancer illness” had 30.907 times the odds of cancer than the participants who had “self-reported non-cancer illness/illnesses.” This means the participants who never had a non-cancer illness were more likely to get cancer than those having noncancer illness/illnesses. It is probably because their bodies had never received any stimulation for improving their immunity.

Table 13: Analysis of maximum likelihood estimates of environmental factors selected

<b>Analysis of Maximum Likelihood Estimates</b>					
<b>Parameter</b>	<b>DF</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>Wald Chi-Square</b>	<b>Pr &gt; ChiSq</b>
<b>Intercept</b>	1	-3.5684	0.8847	16.2700	<.0001
<b>Number of self-reported non-cancer illnesses (0 self-reported non-cancer illness)</b>	1	3.4310	0.0919	1392.4744	<.0001
<b>Age</b>	1	2.1734	0.0382	3238.5312	<.0001
<b>Number of operations, self-reported</b>	1	0.5672	0.0198	818.9063	<.0001
<b>Other serious medical condition/disability diagnosed by doctor (Yes)</b>	1	0.4323	0.1683	6.6013	0.0102
<b>Female</b>	1	0.4066	0.0454	80.2538	<.0001
<b>Long-standing illness, disability or infirmity (Yes)</b>	1	0.2877	0.1447	3.9550	0.0467
<b>Sleeplessness / insomnia (Never/rarely)</b>	1	-0.0718	0.0556	1.6691	0.1964
<b>Water intake (2 glasses water)</b>	1	-0.1307	0.0507	6.6385	0.0100
<b>Sleep duration (6 hours)</b>	1	-0.1816	0.0577	9.9181	0.0016
<b>Number of self-reported non-cancer illnesses (1 self-reported non-cancer illness)</b>	1	-0.3412	0.0518	43.3581	<.0001
<b>Illness, injury, bereavement, stress in last 2 years (None of the above)</b>	1	-0.5171	0.0474	119.1259	<.0001
<b>Illness, injury, bereavement, stress in last 2 years (Death of a close relative)</b>	1	-0.5353	0.0689	60.4413	<.0001
<b>Overall health rating (Good)</b>	1	-0.5767	0.3245	3.1593	0.0755
<b>Overall health rating (Excellent)</b>	1	-0.6269	0.3304	3.6015	0.0577

In addition, some lifestyle/environmental factors which had statistically significant effects for preventing cancer were found. The factors of “Overall health rating (Excellent)” and “Overall

health rating (Good)” were the most important prevention factors, and their negative associations with incidence of cancer were close to statistically significant. Although the p-values (0.06 and 0.08) were not less than 0.05, but they were close to 0.05. The odds ratios of “Overall health rating (Excellent)” and “Overall health rating (Good)” were lowest at 0.534 and 0.562, respectively. These odds ratios indicate that participants who had excellent overall health rating or good overall health rating were less likely to get cancer. There was a statistically significant association between “1 self-reported non-cancer illness” and incidence of cancer since its p-value was less than 0.0001. The participants having one self-reported non-cancer illness had 0.711 times the odds of cancer than those who did not have one self-reported non-cancer illness. It means that the participants who had one self-reported non-cancer illness were less likely to get cancer than those who did not have any non-cancer illnesses or had more than one non-cancer

Table 14: Odds ratio estimates of environmental factors selected

<b>Odds Ratio Estimates</b>			
<b>Effect</b>	<b>Point Estimate</b>	<b>95% Wald Confidence Limits</b>	
<b>Number of self-reported non-cancer illnesses (0 self-reported non-cancer illness)</b>	30.907	25.810	37.010
<b>Age</b>	8.788	8.155	9.471
<b>Number of operations, self-reported</b>	1.763	1.696	1.833
<b>Other serious medical condition/disability diagnosed by doctor (Yes)</b>	1.541	1.108	2.143
<b>Female</b>	1.502	1.374	1.641
<b>Long-standing illness, disability or infirmity (Yes)</b>	1.333	1.004	1.771
<b>Sleeplessness / insomnia (Never/rarely)</b>	0.931	0.835	1.038
<b>Water intake (2 glasses water)</b>	0.878	0.794	0.969
<b>Sleep duration (6 hours)</b>	0.834	0.745	0.934
<b>Number of self-reported non-cancer illnesses (1 self-reported non-cancer illness)</b>	0.711	0.642	0.787
<b>Illness, injury, bereavement, stress in last 2 years (None of the above)</b>	0.596	0.543	0.654
<b>Illness, injury, bereavement, stress in last 2 years (Death of a close relative)</b>	0.585	0.512	0.670
<b>Overall health rating (Good)</b>	0.562	0.297	1.061
<b>Overall health rating (Excellent)</b>	0.534	0.280	1.021

illness. This was an interesting finding which supported the result of the factor “0 self-reported non-cancer illness”, which was a cancer risk factor, to some extent. The reason is probably because one noncancer illness may help to stimulate the body and increase immunity. The factors of “Sleep duration (6 hours)” and “Water intake (2 glasses of water daily)” were also found to have significantly negative associations with incidence of cancer since their p-values were 0.0016 and 0.0100, respectively. The odds ratios were 0.834 and 0.878, respectively. These evidences indicated that the participants who had 6 hours of sleep duration every day were less likely to get cancer than the participants who did not have 6 hours of sleep duration; the participants who had water intake of 2 glasses daily were less likely to get cancer than the participants who did not have water intake of 2 glasses daily. Interestingly, the other hours of sleep duration had not been found out as important environmental factors related to cancer incidence. This may indicate that sleep duration of 6 hours every day is an optimal sleep duration for middle-aged and elderly people to prevent cancer. Although the factor of “Sleeplessness/insomnia (Never/rarely)” did not have significant association with cancer prevention, the p-value was 0.196 which was not quite big. It indicated that this factor had very close association with cancer prevention. The odds ratio was 0.931, which suggested the participants who never/ rarely had sleeplessness/ insomnia were less likely to get cancer. Therefore, good quality sleep may enhance cancer prevention.

The factor of “Illness, injury, bereavement, stress in last 2 years (None of the above)” means there were not any conditions listed on Table 15 except code “-7” which means “None of the above.” Its odds ratio was 0.596, which indicated that the participants who did not have any situations listed on Table 15 were less likely to get cancer. The factor of “Illness, injury, bereavement, stress in last 2 years (Death of a close relative)” was also found to have significantly negative association with incidence of cancer. The factor of “Illness, injury, bereavement, stress in last 2 years (None of the above)” may mean to have some other type of “Illness, injury, bereavement, stress in last 2 years” which was not listed in the field f\_6145. Utilizing the UK Biobank for detailed information of recording this field data may be helpful for better interpretation of the results.

Table 15: field f\_6145 coding and its meaning

Coding	Meaning
1	Serious illness, injury or assault to yourself
2	Serious illness, injury or assault of a close relative
3	Death of a close relative
4	Death of a spouse or partner
5	Marital separation/divorce
6	Financial difficulties
-7	None of the above
-3	Prefer not to answer

#### **4.6 Comparisons of the prediction accuracy and sensitivity of all the models for the balanced dataset**

The dataset must be converted to a balanced dataset for noncancer and cancer classes in the output. Otherwise, none of the models can effectively predict cancer among actual cancer cases. Comparing the results of the experiments (Table 16), we know that CNNs and DNNs had outperform than the other methods, because they not only had the highest values of sensitivity, but also had the better F1 score. After selecting important features and removing the relatively unimportant features, logistic regression still could make good predictions, which the general accuracy was 0.920, the sensitivity was 0.900, and the F1 score was 0.919, and more importantly, providing insights to cancer-related risk factors or associated phenotypes.

Table 16: Comparisons of prediction accuracy and sensitivity among all the models

Model Name	General Accuracy of predictions	Sensitivity of Cancer Case predictions (Recall)	F1 score
DNNs	0.957	0.933	0.956
CNNs	0.962	0.933	0.961
SVM	0.959	0.918	0.957
Random Forest Classifier	0.948	0.896	0.945
Extra Trees Classifier	0.951	0.902	0.948
Logistic Regression for final selected variables	0.920	0.900	0.919

# Chapter 5: Discussion

## 5.1 Deep learning techniques offer better performance and prediction

Conventional machine learning techniques such as SVMs or decision trees are shallow learning. SVMs and decision trees can exhibit good performance on simple classification. But the refined representations required by complex problems generally cannot be achieved by such techniques. As such, researchers had to manually make the initial input data more amenable to process by these methods. Deep learning, on the other hand, completely automates this step: all features are learned in one pass with a single, end-to-end deep-learning model rather than having to engineer them manually for complex multistage pipelines (Francois Chollet, 2018).

Nature does not have any assumption for anything before an event occurs. So, the less assumptions in a predictive model, the higher the predictive power will be. Machine learning works on iterations where computer learns to discover patterns hidden in data. Because machine works on comprehensive data and is independent of all the assumptions, the predictive power is generally very strong in these models (Tavish Srivastava, 2015).

A statistical model uses statistics to build a representation of the data and then conducts analysis to infer any relationships between variables or to discover new insights. Inferring interesting conclusions about real populations almost always requires some background assumptions. Those assumptions must be made carefully because incorrect assumptions can give rise to inaccurate conclusions. It requires the modeler to understand the relationship between variables before putting it in (Tavish Srivastava, 2015).

Due to the advantages of deep learning described above, this study applied deep learning to analyze the data containing environmental factors from the UK Biobank, and meanwhile, the author also analyzed the same data using SVM, random forest, extra trees, and logistic regression for comparisons. Results from this study showed better performance and better predictions using deep learning via deep neural networks and convolutional neural networks than other methods. However, the advantages of deep learning were not prominent since the differences of accuracies were not very large. Potential explanations and future directions are as follows:

- Because of the limitation of computer's CPU, the data in this study did not include all the observations from the United Kingdom Biobank. Deep learning techniques show better performance in complex problems than traditional machine learning techniques. For simple classification problems, SVMs also can perform well. Extracting more raw data from the UK Biobank may be considered for further analysis if computer's CPU permits.

- Deep learning can involve tens or even hundreds of successive layers and they are all learned automatically from exposure to training data. Typically, there are 10 or more intermediate layers. More hidden layers may be added to explore better accuracy for this study.
- Deep learning models were designed expressly to create accurate models directly from raw data by complex networks of artificial neurons (Andrew et al., 2017). In this study, raw data was manipulated, in which case not all information from the raw data was included before being applied to deep learning for analysis. The advantage of using deep learning to analyze raw data directly may be considered for further exploration.

## **5.2 Environmental factors associated with cancer**

### **5.2.1 Risk factors for cancer**

In this study, age was found as the most important cancer risk factor. As age increased, people are more likely to get cancer. This finding is consistent with previous studies. K McPherson et al. reviewed the clinical data and found that the incidence of breast cancer increased with age, doubling about every 10 years until menopause (McPherson et al., 2000). Michael F Leitzmann et al. also demonstrated that prostate cancer incidence strongly increased with age (Leitzmann et al., 2012). Midlife is a period of life when the prevalence of multiple cancer risk is high and incidence begin to increase in many types of cancer (White et al., 2013). U.S. Cancer Statistics Working Group concluded that cancer could be considered as an age-related disease because the incidence of most cancers increases with age and rises more rapidly beginning in midlife based on U.S. cancer statistics from 1999 to 2009 (White et al., 2013; U.S. Cancer Statistics Working Group). However, adults with the greatest longevity are less likely to develop cancer. This indicates that interventions which support healthy environments, assist chronic conditions management, and promote healthy behaviors may help people reduce the likelihood of developing cancer (Terry et al., 2004; Christensen et al., 2012).

Having past surgical operations was also found as a factor increasing cancer risk in this study. Several other studies also supported this finding. J E Rigby et al. studied factors on the risk of invasive breast carcinoma in women age 50-65 years and discovered women with breast cancer were more likely to report physical trauma to the breast within the previous 5 years than the controls (Rigby et al., 2002). They found a strong association between reported trauma to the breast and subsequent development of invasive breast carcinoma (Rigby et al., 2002). Miguel Angel Pérez et al. developed an experimental model in the cheek pouch of hamster and found oral chronic traumatic ulcer acted as a tumor promoter, which demonstrated the carcinogenic action of

oral chronic traumatic ulcer. Wound healing and carcinogenesis have certain biological characteristics in common. In wounds, cells proliferate to allow for healing, and cells cease to proliferate when healing has been completed. This process is regulated by the interaction among the cells which make up the newly formed structures. Normal cells respond appropriately to the controlling signals. In malignant tumors, cell proliferation is deregulated (Pérez et al., 2005; Dai et al., 2017). Cancer cells grow and divide without any control. With continuously unregulated proliferation, cancer cells invade normal tissues and organs and eventually spreading throughout the whole body (Geoffrey M Cooper, 2000; Dai et al., 2017). Cancer cells contain multiple genetic defects including mutations, translocations, and amplifications of oncogenes. All these phenomena require cell division for their occurrence and fixation. The pathogenesis of cancer may result from molecular genetic errors induced during the process of cell division (Susan et al., 1990). Increased cell division stimulated by external or internal factors such as physical or mechanical trauma increases the risk of genetic errors occurring while wound healing. This might be the reason why the number of operations is positively related to cancer incidence.

Another interesting result is that participants who never had self-reported noncancer illness were more likely to get cancer. In other words, healthy people who almost never went to hospitals were suddenly stricken with cancers. This result was rarely reported. We speculate that the immunity may not be activated if people never have any noncancer illness. Another interesting finding is that participants who had one self-reported noncancer illness were less likely to get cancer, which supports the assumption of the role of immunity activation. However, this phenomenon needs to be further researched.

### **5.2.2 Protective factors for cancer among environmental factors**

Sleep duration of 6 hours and never or rarely having sleeplessness are two important factors for cancer prevention. It is general knowledge that it is not good for health if sleep duration is less than 6 hours. Sleep duration of less than 6 hours was found to be a risky factor for the development of chronic diseases, particularly stroke and cancer (Ruesten et al., 2012). However, is it healthy to sleep too long? Although some previous studies found no convincing evidence for an association between sleep duration and incidence of breast cancer (Pinheiro et al., 2006; Girschik et al., 2013), in this study, “Sleep duration of 6 hours” and “Never or rarely having sleeplessness” were found to be two of the important factors for cancer prevention. These results are consistent with some other studies that analyzed big data and/or different populations. Susan Hurley et al. studied sleep duration and cancer risk in 101,609 adult females from California. Results from their analyses suggested that longer sleep might be associated with increased risks of estrogen-mediated cancers such as breast cancer (Hurley et al., 2015). Zhang et al. studied

associations of self-reported sleep duration and snoring with colorectal cancer risk in a total of 30,121 of men aged 41 to 79 years and 76,368 women aged 40 to 73 years. They concluded that longer sleep duration was associated with an increased risk of developing colorectal cancer among individuals who were overweight or snored regularly. This observation raised the possibility that sleep apnea and its attendant intermittent hypoxemia may contribute to cancer risk (Zhang et al., 2013). According to a review of epidemiological studies presented on June 9 at Sleep 2019, more than a third of people in the Americas may have obstructive sleep apnea (Rob Goodier, 2019). Therefore, sleeping too long may not be protective from cancer occurrence for people in middle and old age. Sleep duration of 6 hours may be optimal for middle age and older people.

Overall excellent or good health rating were found to be very important factors for cancer prevention. The odds ratios were lowest, and their association with cancer prevention almost reached the significant level.

### **5.2.3 Some factors were not found to associate with cancer incidence**

This study analyzed the environmental factors for all types of cancer that appeared in 40,548 participants of the UK Biobank. Different types of cancers were not analyzed individually, which may explain why some important risk factors associated with special type of cancer were not found to associate with cancer risk in general in this study. For example, smoking is one of the major risk factors for lung cancer and gastric cancer (Yu et al., 2015; Wah Kit Lam, 2005; Guo et al., 2020), however, smoking was not found as a risk factor for cancer in general in this study. Some other factors which are commonly concerned as risk factors for cancer were also not found in this study. Intake of total meat, red meat, dairy products, different types of fat, and cholesterol were not found to be related to the risk of cancer. These results are consistent with the results of pancreatic cancer risk study conducted by Dominique S. Michaud et al. They did not find those were cancer risk factors either (Dominique et al., 2003).

Vitamin supplementation was not found to be effective for cancer prevention in this study. This result was also consistent with some other studies. Guo et al. demonstrated that a greater preventive effect of vitamin supplementation was only seen among those with low fresh vegetable and fruit intake (Guo et al., 2020). The findings from Theodore M. Brasky et al. did not support the use of most of the supplements studied for prostate cancer prevention (Brasky et al., 2011).

Pollution of the surrounding natural environment (e.g., air, water and soil contamination) are obviously important factors associated with cancer occurrence. However, natural environmental

factors were not found as significant risk factors for cancer in this study. It is maybe because the environment in England is not polluted enough to be a risk factor for cancer, but it might be a different result if studies were conducted in more polluted countries.

### **5.3. Lack of interpretability using machine learning**

#### **5.3.1 The reasons causing uninterpretable problem when applying machine learning in healthcare**

Machine learning including deep learning has achieved great success in many fields. There are promising results of using machine learning in the healthcare field, with the application of discovering disease phenotypes, identifying risk factors, and predicting diseases. However, unlike the imaging data, healthcare data is more diverse, complicated, and irregular without obvious spatial or sequential structure. Most diseases are highly heterogeneous, and there is still no complete knowledge on the causes and how they progress for most of the diseases (Miotto et al., 2018). There are several unresolved challenges when applying machine learning on the practical healthcare. For instance, machine learning sometimes produces uninterpretable problems.

Machine learning is often referred to as “black box models”, in which the rationale for generated outputs is inscrutable not only to physicians but also to engineers who develop them. Because accuracy-driven performance metrics are now pushing toward more opaque models, subtle shortcomings of machine learning may be difficult or impossible to prevent or detect. Thus, use of machine learning may have unintended consequences (Cabitza et al., 2017).

#### **5.3.2 Some uninterpretable scenarios in this study**

There are some uninterpretable scenarios in this study, which are the same problem as what Miotto et al. found when reviewing deep learning for healthcare and what Cabitza et al. found in reviewing unintended consequences of applying machine learning in medicine (Miotto et al., 2017; Cabitza et al., 2017). This is a common problem when applying machine learning to analyze data in the healthcare sector currently.

In this study, the factor of “Illness, injury, bereavement, stress in last 2 years (None of the above)” had significant negative association with incidence of cancer. However, the factor of “Illness, injury, bereavement, stress in last 2 years (Death of a close relative)” was also found to have significantly negative association with incidence of cancer. Both factors being significantly negatively associated with cancer incidence appears to be difficult to interpret.

#### **5.3.3 Promising ideas may be considered to solve the uninterpretable problem**

Interpretation might not be a problem in other more deterministic domains such as image annotation, because the end user can objectively validate the tags assigned to the images. In

healthcare, not only the quantitative algorithmic performance is important, but why the algorithms work is also important. In fact, model interpretability sometimes can be of key importance for convincing the medical professionals to take actions recommended by the predictive system. For example, prediction models can provide information regarding potential high risk of developing a certain disease or prescription of a specific medication (Miotto et al., 2018).

#### **5.3.3.1 Integrating features may need to be considered**

Potential risk factors may not be independent. They may have correlations with others because of the shared reasons behind it. It is possible that a single risk factor is not important or does not have a direct causal relation to the target disease, but its combinations with other factors may be the triggering or causal factor of that disease (Suo et al., 2016). Therefore, integrating features may need to be considered in the process of feature selection.

#### **5.3.3.2 Collaborate with experts for the external knowledge**

The existing expert knowledge for medical problems is invaluable in the health field. Because of limitation on the amount of medical data, various quality problems, and medical complexity, it is very important and helpful to incorporate expert knowledge into the machine learning process to guide it toward the optimal outcome. The quality of machine learning and subsequent regulatory decisions regarding adoption should be subject to proof of clinically significant improvements in relevant outcomes compared with usual care. It should also be satisfied by patients and physicians (Cabitza et al., 2017).

#### **5.3.3.3 Automatic explanation is being researched in healthcare area**

Research to alleviate challenges between accuracy and interpretability is being conducted. Machine learning is expected to automatically provide explanations and offer physicians rich interactive visualization tools in exploring the implications of potential exposure variables (Cabitza et al., 2017).

Therefore, to solve uninterpretable problems in this study, utilizing the UK Biobank for detailed information of those specific fields may be helpful for more accurate interpretations. The uninterpretable results may also come from interactions among these environmental factors and other environmental factors associated with cancer incidence. We may discuss with physicians or nutritionists about how to set up potential interactions among some factors before analyzing data in a future study.

## Chapter 6: Conclusion

This study extracted the data of environmental factors as features and “Cancer diagnosed by doctor” as outcome from the UK Biobank. In the original raw dataset, there were 50,000 participants and 3175 columns which were environmental factors or multiple instances for some factors. After the raw dataset was preprocessed, there were 40,548 participants and 147 features. After categorical variables were transformed into dummy variables, numerical variables were standardized, there were 40,548 rows and 1433 columns in the dataset for analysis. Deep learning techniques including deep neural networks and convolutional neural networks, and conventional machine learning techniques including random forest, extra trees, support vector machine (SVM), and logistic regression were applied to analyze the data for establishing a model to predict cancer and explore cancer risk factors among environmental factors.

All the sensitivities and F1 scores were over 0.900 and 0.919 respectively from the analyses of DNNs, CNNs, SVM, random forest, extra trees, and logistic regression. Overall, CNNs had the best prediction performance, DNNs had the second-best.

This study found some environmental factors had statistically significant associations with incidence of cancer. “Age”, “Number of operations, self-reported”, “Other serious medical condition/disability diagnosed by doctor (Yes)”, “Long-standing illness, disability or infirmity (Yes)” were significant risk factors for cancer. Females were significantly more likely to get cancer than males. “Number of self-reported non-cancer illness” was found to have significant associations with cancer incidence. However, the factor of “0 self-reported non-cancer illness” was a significant risk factor for cancer, while the factor of “1 self-reported non-cancer illness” was a significant protective factor for cancer, which is a very interesting finding requiring further confirmation and discussion among physiologists with respect to the physiology and pathology. This is probably related to the human immune system. The factors of “Sleep duration of 6 hours” and “Water intake (2 glasses of water daily)” were found to be significant protective factors for cancer. The factors of “Overall health rating (Excellent)”, “Overall health rating (Good)” and “Sleeplessness / insomnia (Never/rarely)” were also very important factors for cancer prevention. One of the common unsolved problems, which is uninterpretable for some factors in healthcare data using machine learning, appeared in this study. The factors of “Illness, injury, bereavement, stress in last 2 years (None of the above)” and “Illness, injury, bereavement, stress in last 2 years (Death of a close relative)” both had significantly negative associations with incidence of cancer. Collaborating with experts in healthcare and exploring interaction terms among appropriate environmental factors may be considered for future research.

The analysis models of highly accurate prediction in this study will be helpful for initiating early cancer screening. The findings of risk factors and protective factors in this study will be helpful for educating the public about the risk of environmental factors in a high-risk population for cancer prevention. Sleep duration of 6 hours being optimal was further demonstrated by analyzing this big data in this study, while there were many academic disputes in previous research regarding the length of sleep duration associated with cancer prevention. Based on the author's knowledge regarding published research, it has not been found that the factor of "0 self-reported non-cancer illnesses" was a risk factor for cancer, and in contrast, the factor of "1 self-reported non-cancer illness" was a cancer preventive factor. It is an interesting finding for further research and uncovering of mechanisms.

## References

- Anand P, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, Sung B & Aggarwal BB. Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical Research*. 2008; 25:2097-2116.
- Beam AL & Kohane IS. Big Data and Machine Learning in Health Care. *Viewpoint*. 2018; 319 (13):1317-1318.
- Bisong E. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. New York. *Apress*. 2019.
- Block G. Vitamin C and Cancer Prevention: the epidemiologic evidence. *American Society for Clinical Nutrition*. 1991; 53:270S-282S.
- Boffetta P & Hashibe M. Alcohol and cancer. *The Lancet Oncology*. 2006; 7:149-156.
- Brasky TM, Kristal AR, Navarro SL, Lampe JW & Peters U. Specialty supplements and prostate cancer risk in the Vitamins and lifestyle (VITAL) cohort. *Nutrition and Cancer*. 2011; 63 (4):573-582.
- Brosch T, Tam R. Manifold Learning of brain MRIs by deep learning. *Med Image Comput Assist Interv*. 2013; 16:633-640.
- Cabitza F, Rasoini R & Gensini GF. *Viewpoint*. August 8, 2017. <https://jamanetwork.com/journals/jama/article-abstract/2645762>
- Cancer Facts & Figures 2019. American Cancer Society. <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2019.html>
- Chollet F. Deep Learning with Python, New York. *Manning Publications Co*. 2018.
- Cancer. World health organization. 12 September 2018. <https://www.who.int/news-room/fact-sheets/detail/cancer>
- Chawla NV, Bowyer KW, Hall LO & Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002; 16:321-357.
- Cheng Y, Wang F, Zhang P & Hu J. Risk Prediction with Electronic Health Records: A Deep Learning Approach. *Society for Industrial and Applied Mathematics*. 2016; 432-440. <https://epubs.siam.org/doi/abs/10.1137/1.9781611974348.49>
- Choi E, Bahadori MT, Schuetz A, Stewart WF & Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. Cornell University. 2016; 56:1-18.

- Christensen K, Pedersen JK, Hjelmberg JvB, Vaupel JW, Stevnsner T, Holm NV & Skytthe A. Cancer and longevity—is there a trade-off? A study of co-occurrence in Danish twin pairs born 1900–1918. *The Journals of Gerontology*. 2012; 67A (5):489-494.
- Cooper GM. *The Cell: A Molecular Approach*, 2nd edition. Sunderland. MA. 2000. <https://www.ncbi.nlm.nih.gov/books/NBK9963/>
- Dai J, Miller MA, Evertts NJ, Wang Xi, Li P, Li Y, Xu J-h & Yao G. Elimination of Quiescent Slow-Cycling Cells via Reducing Quiescence Depth by Natural Compounds Purified from *Ganoderma Lucidum*. *Oncotarget*. 2017; 8(8):13770-13781.
- DISQUS. SMOTE explained for noobs - Synthetic Minority Over-sampling Technique line by line. Nov 2017. [https://rikunert.com/SMOTE\\_explained](https://rikunert.com/SMOTE_explained)
- Frattaroli J, Weidner G, Dnistrian AM, Kemp C, Daubenmier JJ, Marlin RO, Crutchfield L, Yglecias L, Carroll PR & Ornish D. Clinical events in prostate cancer lifestyle trial: results from two years of follow-up. *Urology*. 2008; 72(6):1319-1323.
- GeeksforGeeks. Title: Keras.Conv2D Class. 2020. <https://www.geeksforgeeks.org/keras-conv2d-class/>
- Géron A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media. CA. 2017.
- Geurts P, Ernst D & Wehenkel L. Extremely randomized trees. *March Learn*. 2006; 63:3-44.
- Girschik J, Heyworth J & Fritschi L. Self-reported sleep duration, sleep quality, and breast cancer risk in a population-based case-control study. *American Journal of Epidemiology*. 2013; 177 (4):316-327.
- Guo Y, Li Z-X, Zhang J-Y, Ma J-L, Zhang L, Zhang Y, Zhou T, Liu W-D, Han Z-X, Li WQ, Pan K-F & You W-C. Association between lifestyle factors, vitamin and garlic supplementation, and gastric cancer outcomes. *JAMA Network*. 2020; 3(6):e206628.
- Gupta D. Fundamentals of Deep Learning – Activation Functions and When to Use Them? January 30, 2020. <https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/>
- Heaton J. (Book Review). Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning, in Genetic Programming and Evolvable Machines*. 2018; 19:305-307.
- Hurley S, Goldberg D, Bernstein L & Reynolds P. Sleep duration and cancer risk in women. *Cancer Causes Control*. 2015; 26:1037-1045.
- Jason Brownlee. A Gentle Introduction to Dropout for Regularizing Deep Neural Networks. August 6, 2019. <https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/>

- Jason Brownlee. Use Early Stopping to Halt the Training of Neural Networks at the Right Time. December 10, 2018. <https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/>
- Kim R. Effects of surgery and anesthetic choice on immunosuppression and cancer recurrence. *Journal of Translational Medicine*. 2018; 16(8):1-13.
- Kingma DP & Ba J. Adam: A Method for Stochastic Optimization. Cornell University. 2015; 9:1-15.
- Karen GM. What Is an ROC Curve? The Analysis Factor. 2008. <https://www.theanalysisfactor.com/what-is-an-roc-curve/>
- Lam WK. Lung cancer in Asian women—the environment and genes. *Respirology*. 2005; 10:408-417.
- Leitzmann MF & Rohrmann S. Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates. *Clin Epidemiol*. 2012; 4:1-11.
- Lozano-Diez A, Zazo R, Toledano DT & Gonzalez-Rodriguez J. An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition. *PLOS ONE*. August 10, 2017 (online).
- Makuuchi M. Recurrence of hepatocellular carcinoma after surgery. *BJS European colorectal congress*. 1996; 83(9):1219-1222.
- Mamoshina P, Vieira A, Putin E & Zhavoronkov A. Applications of deep learning in biomedicine. *Mol Pharm* 2016; 13(5):1445-1454.
- McPherson K, Steel CM & Dixon JM. Breast cancer—epidemiology, risk factors, and genetics. *ABC of breast diseases*. 2000; 321(7261): 624-628.
- MedCalc. Logistic regression. 2020. [https://www.medcalc.org/manual/logistic\\_regression.php](https://www.medcalc.org/manual/logistic_regression.php)
- Michaud DS, Giovannucci E, Willett WC, Colditz GA & Fuchs CS. Dietary meat, dairy products, fat, and cholesterol and pancreatic cancer risk in a prospective study. *American Journal of Epidemiology*. 2003; 157:1115-1125.
- Miotto R, Wang F, Wang S, Jiang X & Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*. 2018; 19(6):1236-1246.
- Misra S, Li H & He J. Machine Learning for Subsurface Characterization (first edition). Houston. Gulf Professional Publishing. 2019; 9:266.
- Mu R & Zeng X. *A Review of Deep Learning Research*. 2019; 13(4): 1738-1764.

- Verma M (Editor) & Walker JM (Series Editor). *Cancer Epidemiology*. New Jersey. *Humana Press*. 2009.
- Naylor CD. On the Prospects for a (Deep) Learning Health Care System. *Viewpoint*. 2018; 320 (11):1099-1100.
- Peltarion. Binary crossentropy. 2020. <https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/binary-crossentropy>
- Pérez MA, Raimondi AR, Itoiz ME. An experimental model to demonstrate the carcinogenic action of oral chronic traumatic ulcer. *Journal of Oral Pathology & Medicine*. 2005; 34:17-22.
- Peto R, Darby S, Deo H, Silcocks P, Whitley E. & Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ*. 2000; 321:323-329.
- Pinheiro SP, Schernhammer ES, Tworoger SS & Michel KB. A Prospective Study on Habitual Duration of Sleep and Incidence of Breast Cancer in a Large Cohort of Women. *Cancer Research*. 2006; 66 (10):5521-5525.
- Prabhu. Understanding of Convolutional Neural Network (CNN) — Deep Learning. Mar 4, 2018. <https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148>
- Preston-Martin S, Pike MC, Ross RK, Jones PA & Henderson BE. Increased Cell Division as a Cause of Human Cancer. *Cancer Research*. 1990; 50:7415-7421.
- Rigby JE, Morris JA, Lavelle J, Stewart M & Gatrell AC. Can physical trauma cause breast cancer? *European Journal of Cancer Prevention*. 2002; 11:307-311.
- Ruesten AV, Weikert C & Fietze I. Association of Sleep Duration with Chronic Diseases in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. *PLOS ONE*. January 25, 2012; 7(1):e30972.
- Saha S. A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way. Dec 15, 2018. <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>
- Slattery ML, Curtin K, Anderson K, Ma K-N., Ballard L, Edwards S, Schaffer D, Potter J, Leppert M & Samowitz WS. Associations between cigarette smoking, lifestyle factors, and microsatellite instability in colon tumors. *Journal of the National Cancer Institute*. 2000; 92:1831-1836.
- SPRH LABS. Understanding Deep Learning: DNN, RNN, LSTM, CNN and R-CNN. March 21, 2019. <https://medium.com/@sprhlab/understanding-deep-learning-dnn-rnn-lstm-cnn-and-r-cnn-6602ed94dbff>

- Srivastava T. Difference between Machine Learning & Statistical Modeling. Analytics. Vidhya. July 1, 2015. <https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/>
- Stephanie. Lasso Regression: Simple Definition. *Statistics How To*. September 2015. <https://www.statisticshowto.com/lasso-regression/>
- Stephanie. Ridge Regression: Simple Definition. *Statistics How To*. July 2017. <https://www.statisticshowto.com/lasso-regression/>
- Sud A, Kinnersley B & Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer*. 2017; 17(11):692-704.
- Suo Q, Xue H & Xue H. Risk Factor Analysis Based on Deep Learning Models, Proceedings of the 7th ACM International Conference on Bioinformatics. *Computational Biology and Health Informatics*. 2016; 10:394-403.
- Takeshi S, Mitsuru S, Kinoshita T & Maruyama K. Recurrence of early gastric cancer. Follow-up of 1475 patients and review of the Japanese literature. *Cancer*. 1993; 72(12):3174-3178.
- Tavare AN, Perry NJS, Benzoana LL, Takata M & Ma D. Cancer recurrence after surgery: Direct and indirect effects of anesthetic agents. *International Journal of Cancer*. 2012; 130(6):1237-1250.
- Terry DF, Wilcox MA, McCormick MA, Pennington JMY, Schoenhofen EA, Andersen SL & Perls TT. Lower all-cause, cardiovascular, and cancer mortality in centenarians' offspring. *The American Geriatrics Society*. 2004; 52: 2074-2076.
- Thun MJ, Jacobs EJ & Patrono C. *Natural Reviews Clinical Oncology*. 2012; 9(5):259-267.
- Ujjwal Karn. An Intuitive Explanation of Convolutional Neural Networks. August 11, 2016. <https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/>
- UK biobank. Last updated Jun 17, 2020. <https://www.ukbiobank.ac.uk/>
- U.S. Cancer Statistics Working Group. U.S. cancer statistics: 1999-2009 incidence and mortality web-based report. 2013. [www.cdc.gov/uscs](http://www.cdc.gov/uscs)
- Varun, G, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, & Webster DR. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016; 316(22):2402-2410.
- Wakai K, Ohno Y, Obata K. & Aoki K. Prognostic significance of selected lifestyle factors in urinary bladder cancer. *Japanese Journal of Cancer Research*. 1993; 84:1223-1229.

White MC, Holman DM, Boehm JE, Peipins LA, Grossman M & Henley SJ. Age and Cancer Risk: A Potentially Modifiable Relationship. *Elsevier*. 2014; 46(3):S7-S15.

Wikipedia. Support vector machine. October 2020.  
[https://en.wikipedia.org/wiki/Support\\_vector\\_machine](https://en.wikipedia.org/wiki/Support_vector_machine)

Wikipedia. Sensitivity and specificity. October 2020.  
[https://en.wikipedia.org/wiki/Sensitivity\\_and\\_specificity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity)

Wikipedia. October 2020. F-score. [https://en.wikipedia.org/wiki/F1\\_score](https://en.wikipedia.org/wiki/F1_score)

Wikipedia. Lasso (statistics). October 2020. [https://en.wikipedia.org/wiki/Lasso\\_\(statistics\)](https://en.wikipedia.org/wiki/Lasso_(statistics))

Yamamoto J, Kosuge T, Takayama T, Shimada K, Yamasaki S, Ozaki H, Yamaguchi N, Makuuchi M. Recurrence of hepatocellular carcinoma after surgery. *BJS European colorectal congress*. 1996; 83(9):1219-1222.

Yang CS, Chung JY, Yang GY, Chhabra SK & Lee MJ. Tea and Tea Polyphenols in Cancer Prevention. *The Journal of Nutrition*. 2000; 130(2):472S-478S.

Yu S, Yang CS, Li J, You W, Chen J, Cao Y, Dong Z & Qiao Y. Cancer prevention research in China. *Cancer Prevention Research*. 2015; 8(8):662-674.

Zhang X, Giovannucci EL, Wu K, Gao X, Hu F, Ogino S, Schernhammer ES, Fuchs CS, Redline S, Willett WC & Ma J. Associations of self-reported sleep duration and snoring with colorectal cancer risk in men and women. *Sleep*. 2013; 36(5):681-688.