ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

Colorectal Cancer Screening: Simulating Haemoglobin Values for MISCAN-Colon Using Machine Learning[1]

Author

Yoëlle Kilsdonk (513530)

Supervisors

E. P. O'Neill (EUR)

dr. I. Lansdorp-Vogelaar (EMC)

R. van den Puttelaar (EMC)

D. van den Berg (EMC)

Second assessor

TBD

May 30, 2022

[1]The views stated in this thesis are those of the author and not necessarily those of the supervisors, second assessor, Erasmus School of Economics, Erasmus University Rotterdam or Erasmus Medical Centre.

Abstract

Keywords— MISCAN, Machine Learning

Table of contents

List of Figures

List of Tables

1 Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer-related deaths in Western countries, while also being one of the most preventable types of cancer (Loeve et al., 1999; Sung et al., 2021; Torre et al., 2015). An important determinant for CRC prevention is screening, but how do we determine what policies work best? Clinical trials often only last a couple of years, while we are most interested in the (cost-)effectiveness of screening policies over a lifetime. For example, to answer the question 'can we prevent CRC mortality through changes in the current policy?', one would have to follow individuals over all their life, which is infeasible in practice. Additionally, it would be impossible to simultaneously implement and evaluate multiple policies using a real-life population. To this end, the Erasmus Medical Center (EMC) developed the MISCAN-Colon (MIcrosimulaten SCreening ANalysis) model – a microsimulation model for the evaluation of CRC screening.

The data for this research is provided by EMC, and contains information on the Dutch national CRC screening programme from 2014-2019. For each of the 3.5 million individuals in the data set, a maximum of four screening rounds are available in the data set, along with their age and sex[2].

We can distinguish a demography part, a natural history part and a screening part in the MISCAN-Colon model. In the natural history part, life histories are generated for the demography, during which colorectal adenomas may develop and sometimes may cause death. The second part overlays screening policies. In this model, we extrapolate results from the EMC data on a simulated screening population, and use this information to assess incidence and mortality rates with, and without screening. Using this MISCAN-Colon model, we can evaluate different screening policies by comparing their costs and effectiveness (Loeve et al., 1999).

Before, the MISCAN-Colon model simulated a positive or negative faecal immunochemical test (FIT) result based on the sensitivity and specificity of the FIT. Recently, however, the Public Health department of EMC explored the extension of predicting FIT results in the MISCAN-Colon model with a simulation model for the haemoglobin (Hb) values in a patient's stool. With this new extension, MISCAN-Colon needs accurate simulations of the Hb concentration to evaluate the benefits of personalised screening strategies. To predict these concentrations, this research uses black-box machine learning methods on the longitudinal data set provided by the EMC. Unfortunately, however, the assumption of iid observations – necessary for most machine learning methods – is often violated in longitudinal data due to correlations within observations[3]. To overcome this issue, Ngufor et al. (2019) propose an approach which incorporates random-effects in machine learning algorithms for efficient analysis of longitudinal data.

The current method to simulate Hb values for MISCAN-Colon is a mixed-effect zero-inflated negative binomial model (ZINB). However, van den Berg (2021) finds that mixed-effect machine learning (MEml) models outperform the mixed-effect ZINB model significantly for this purpose, based on the approach by Ngufor et al. (2019). The optimal MEml model was chosen to be a decision tree due to its interpretability, as more complicated models attained similar performance. It is unclear, however, whether the increase

[2]The method of data imputation will follow in the coming days, after more in depth evaluation of the data set.

[3]In this case, patients with positive FITs participate in multiple rounds, which allows for such correlation.

in predictive accuracy found in van den Berg (2021) is specifically due to the inclusion of random-effects, or due to the use of machine learning methods in general. Therefore, this research investigates the contribution of the inclusion of random-effects to the predictive accuracies of black-box machine learning methods. We implement artificial neural networks (ANNs) and support vector machines (SVMs), both with, and without mixed-effects, using the approach of Ngufor et al. (2019). This leads to the following research questions:

RQ1a Do MEml models outperform 'regular' machine learning models?

RQ1b Which model is best suited for predicting the Hb concentration in CRC screening?

This research consists of two phases, the first being outside of MISCAN-Colon, where we predict Hb concentrations using four different models. These models are trained, tested, and validated using a longitudinal data set provided by the EMC. Based on phase one, we answer RQ1a and RQ1b. In phase two, we implement the most promising model in MISCAN-Colon, and calibrate this model such that the simulated Hb concentrations resemble the observed concentrations of real-life Dutch population screening data as closely as possible.

2 Literature

2.1 Colorectal cancer

Colorectal cancer (CRC) is the development of cancer from the colon or rectum, which usually starts as a benign adenoma. CRC is one of the most commonly diagnosed and most deadly cancers worldwide (Torre et al., 2015; Sung et al., 2021). According to the Dutch Rijksinstituut voor Volkgezondheid en Milieu, 5% of people will develop CRC in the Netherlands. Nearly nine in ten cases occur in people older than 55. Risk factors for CRC include age, gender, genetics, environment, diet, physical activity, and smoking (Botteri et al., 2008; Thanikachalam and Khan, 2019). Moreover, the worldwide burden of CRC is expected to further increase due to, *inter alia*, the growth and aging of the population (Jiang et al., 2022).

Adenomas of the colon are estimated to be present in 20-53% of the U.S. population older than 50 years of age, with a prevalence of 0.2-0.6% for adenocarcinomas (CRC) (Strum, 2016). Hence, given that only a small percentage of adenomas becomes cancerous, we distinguish between progressive and non-progressive adenomas, where non-progressive adenomas do not develop into an adenocarcinoma (see Figure 1). We also distinguish between clinical and preclinical stages, where preclinical indicates that the cancer is not yet diagnosed. Preclinical cancer can then progress from stage I to stage IV, where symptoms may develop in each stage, which in turn may lead to disease diagnosis (Compton and Greene, 2004). Once the cancer has been diagnosed, the cancer is referred to as clinical.

Figure 1 shows the progression of CRC in five stages. In stage 0, the adenoma is *in situ* and has not grown beyond the mucosa (i.e., the inner lining) of the colon or rectum. Stage I is when the adenoma has grown beyond the mucosa, without spreading to the lymphatic system or distant organs. In stage II

the adenoma has invaded the colonic or rectal wall, with possible infection of nearby organs. Finally, in stages III and IV, the metastatic adenocarcinoma has spread to lymph nodes and distant organs.

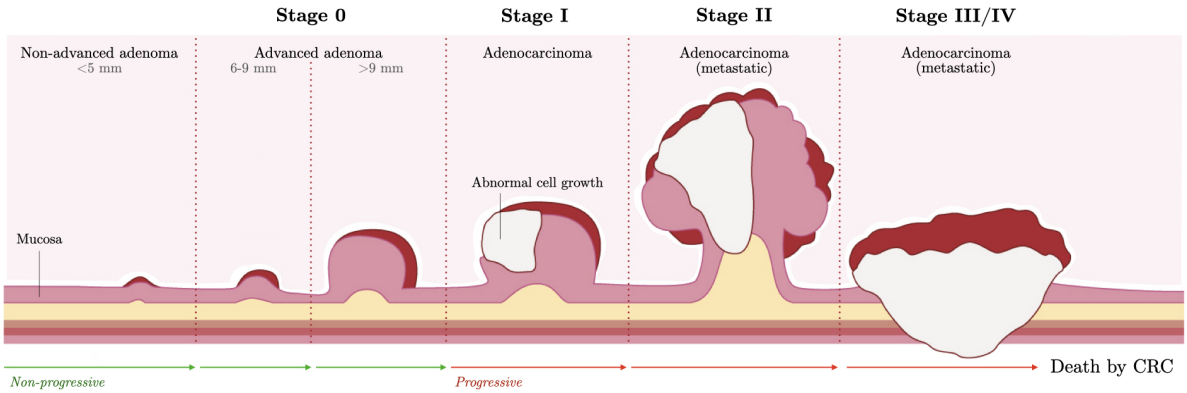

Figure 1: Progression of colorectal cancer in stages

2.1.1 Screening

The effect of screening is twofold. First, research indicates that over 90-95% of CRCs develop from (benign) adenomas (Bronner and Haggitt, 1993; Morson, 1974). Hence, early detection and removal might prevent CRC (Loeve et al., 1999). Second, early detection of an (a)symptomatic cancer may result in an improvement in prognosis. More specifically, a large body of literature finds that screening results in a reduction in mortality, as cancers can be detected at an early and curable stage (Jiang et al., 2022; Levin et al., 2008; Whitlock et al., 2012; Toribara and Sleisenger, 1995).

Screening tests can be subdivided into two categories: stool-based tests and visual exams. The guaiac-based fecal occult blood test (gFOBT) and fecal immunochemical test (FIT), e.g., belong to the first category, in which the stool is tested for Hb. If high Hb values are present, this could be an indicator for the presence of CRC. The two most common visual exams are (flexible) sigmoidoscopy, and colonoscopy, which investigate the structure of the colon and rectum for abnormal tissue. According to the review by Ding et al. (2022), colonoscopies are most effective in reducing CRC-related deaths, at an approximate 68% decrease (Brenner et al., 2014). As for the stool-based tests, the FIT test reduces CRC-related deaths by 22% on average, which is approximately 7% more effective than the gFOBT test (Hewitson et al., 2008; Zorzi et al., 2015). The FIT test also has a higher participation rate and positivity rate compared to gFOBT in CRC screening programs, while reporting fewer false negatives (Mousavinezhad et al., 2016). Moreover, the FIT test is relatively close in effectiveness compared to flexible sigmoidoscopies, with reported reduction of approximately 28%, while being considerably less invasive (Holme et al., 2013). When screening with a test (other than a colonoscopy) leads to abnormal test results, the general advice is to follow up with a colonoscopy in due time (Ding et al., 2022).

In the Netherlands, each person between the age of 55-75 is asked to participate in the biennial population screening for CRC once every two years since January of 2014[4]. The participants receive a FIT, which is sent back to the hospital after taking a stool sample. In the event of an aberrant result, a

[4]For more information see: https://www.rivm.nl/darmkanker.

referral is made for a colonoscopy and, if necessary, treatment. If any abnormalities are present during the colonoscopy, small amounts of tissue can be removed for analysis (i.e., a biopsy), and abnormal growths, or adenomas, can be identified and removed. This way, CRC can be detected at an early stage. According to the Integraal Kankercentrum Nederland, patients diagnosed with CRC through the population screening had a more favorable stage distribution than patients without screening (see Figure 2). Also, patients who were diagnosed through population screening were more likely to receive less invasive treatments.

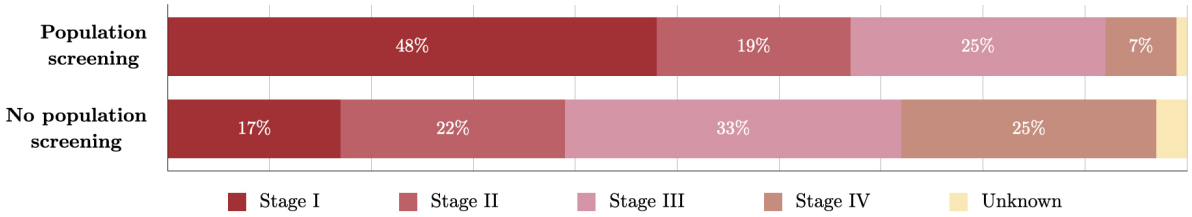

Figure 2: Distribution of diagnosed cancers in patients with, and without screening (source: https://iknl.nl/)

2.2 MISCAN-Colon

Unfortunately, screening is not a silver bullet in healthcare, as it could lead to, e.g., overdiagnosis or false positives, while also being costly and invasive. Welch and Black (2010) provide a summary of current evidence that early detection leads to overdiagnosis in breast, lung, and prostate cancer, where overdiagnosis is defined as the diagnosis of a medical condition or disease that would not cause symptoms or death during a patient's lifetime. Overdiagnosis is associated with long-term psychosocial harm, lower quality of life, and unwanted/unnecessary usage of (follow-up) tests, treatment, and healthcare facilities (Barton et al., 2001; Brodersen and Siersma, 2013; Jenniskens et al., 2017; van der Steeg et al., 2011). On the other hand, Brasso et al. (2010) and Wardle et al. (2003) find no adverse psychological effects due to cancer screening, although they do not specifically investigate the effects of overdiagnosis. That said, overdiagnosis could be particularly harmful if it leads to unnecessary treatments, each of which comes with their specific risk[5].

Given the previously stated disadvantages to screening, it is of high importance to identify the most cost-efficient and the most effective screening policy. Using the adapted version of Habbema et al. (1985)'s MISCAN (MIcrosimulaten SCreening ANalysis) microsimulation model – called MISCAN-Colon – we can overlay screening scenarios on a simulated population *before* real-life implementation, such that we can evaluate different screening policies by comparing their costs and effectiveness, as well as assessing the risk of false positives and overdiagnosis (Loeve et al., 1999). This model is an adapted version of Habbema et al. (1985)'s MISCAN microsimulation model for the evaluation of screening.

The model simulates a large number of individual life histories in which several colorectal lesions can emerge, and consequently produces incidence and mortality rates in the simulated population with (or without) screening, using information on the epidemiology, natural history of the disease, and screening

[5]For an assessment of operative risk in CRC surgery, we refer to Fazio et al. (2004).

and demography characteristics as input. By comparing the simulated life histories with, and without screening, MISCAN-Colon can evaluate the costs and benefits of a specific screening strategy.

Figure 3 shows an exemplified version of the three parts of MISCAN-Colon, using a fictive individual named Robin. The upper line simulates the life of Robin without cancer, referred to as the demography part, who dies at 87 years old of other causes than CRC. The middle line simulates Robin's life *with* cancer, but without screening, which adds a natural history to the demography part. In this scenario, Robin dies at 72 due to CRC. The bottom line simulates Robin's life when screening is overlayed, with 15 life years gained as a result.

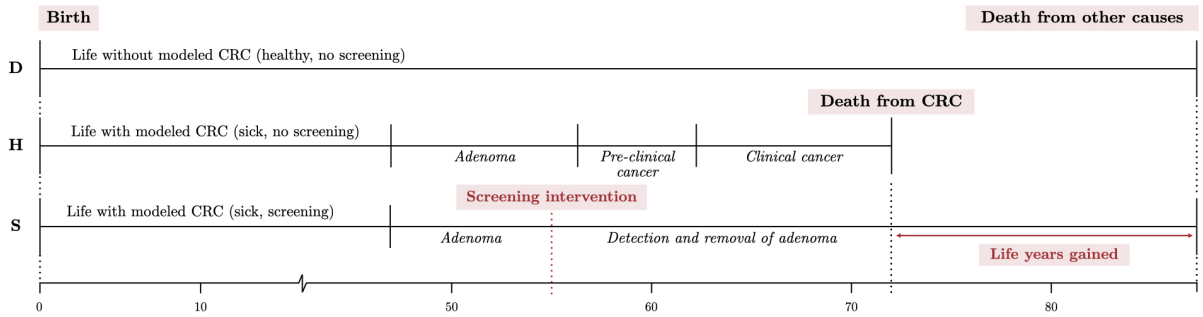

Figure 3: Simulations from the MISCAN-Colon model, where the upper bar shows the demography part (D), the middle bar adds the natural history (H) to D, and the lower bar adds both H and screening (S)

Three remarks on Figure 3. First, the survival of a lesion after diagnosis depends on the stage of the cancer (and other risk factors). Thus, screening does not ensure that an individual no longer dies from CRC. The possible prognostic consequences after a positive test result for cancer screening are: total cure, delay in moment of death, no change in moment of death, or premature death by complications of treatment. Second, the figure only shows examples of individuals with *one* lesion for simplicity, but the MISCAN-Colon model allows for the modelling of multiple lesions. New lesions that appear after clinical diagnosis of cancer are accounted for in the simulated survival of the clinically diagnosed cancer. Third, this figure only shows adenomas that progress to cancer, but it is also possible that an individual develops non-lethal adenomas which would never result in death of an individual, and it is possible that a lesion is invasive from the beginning, i.e., a cancer without a preceding polyp.

2.3 Methods

As mentioned previously, oftentimes healthcare data is longitudinal, with (possibly) repeated measurements over different intervals of time, which could cause correlations within patient data. One solution to this problem would be to employ 'regular' machine learning models while explicitly modeling the interpatient correlation through inclusion of momentum-specific variables (e.g., current number of test, previous Hb value, maximum Hb value). However, the nature of this data suggests that better estimation may be possible if the information of the repeated measurements would be included at the level of the algorithm itself. In this section, we provide an overview of the literature on machine learning – specifically artificial neural networks (ANNs) and support vector machines (SVRs) – in longitudinal health data.

2.3.1 Artificial neural networks

The trajectory of cancer is clearly non-linear, highly variable and dependent on a large variety of factors, most of which are not understood to this day. The flexibility of neural networks can be used to effectively address these problems. Also, Haghani et al. (2017) show that ANNs are suitable machine learning algorithms for the prediction of non-negative variables, which is corroborated by Sakthivel and Rajitha (2017). However, ANNs, just as SVMs, make the implicit assumption of iid data. While certain ANNs have been succesfully adjusted to account for temporal trends (e.g., recurrent neural networks in Choi et al. (2016)), longitudinal data could contain unequal time intervals between measurements, and an unequal number of observations per individual. To this end, several methods for using ANNs on longitudinal data have been proposed.

Xiong et al. (2019) propose a new type of neural network called the mixed effects neural network model, which adapts mixed effects within a deep neural network architecture for gaze estimation, based on eye images. This model is person-specific, and uses few calibration samples to eliminate the person-specific bias in longitudinal data. In the field of Alzheimers disease, Tandon et al. (2006) propose another mixed effects neural network to accurately model the nonlinear course of the disease. Their model generalizes a linear mixed effects model by incorporating a general non-linear function of the input variables. Their model is shown to be much more accurate and effective compared to standard ANNs and linear mixed effects models. Lastly, Mandel et al. (2021) propose a generalized neural network mixed model, which is structured as a GLMM, where the linear fixed effect is replaced by a feed-forward neural network and a random effect component is added. They use this approach to predict depression and anxiety levels of schizophrenic patients using longitudinal data.

2.3.2 Support vector machines

In an attempt to merge longitudinal data with machine learning, Luts et al. (2012) propose a mixed-effects least squares support vector machine (LS-SVM) classifier using regression modeling and a prediction step. This approach is computationally efficient as only a linear system needs to be solved. The research by Cheng et al. (2014) provides analytical expressions of confidence and prediction intervals of mixed-effects LS-SVM approaches such as this one. An alternative approach to modeling longitudinal data using SVM is proposed by Chen and DuBois Bowman (2011). They generalize the optimization problem of SVM by constructing a support vector classifier based on linear combinations of features from different cross-sectional time-points to make predictions, using an expectation-maximization algorithm.

Another branch of literature focuses on the generalisation of SVM to SVR. For example, the longitudinal SVM classifier by Chen and DuBois Bowman (2011) is extended by Du et al. (2015) to perform regression. Another SVR model suitable for longitudinal data is the semiparametric mixed-effects least squares support vector regression (LS-SVR) model by Seok et al. (2011). This model shows slightly improved performance and prediction over 'standard' LS-SVR using pharmacokinetic and pharmacodynamic data. Finally, Cho (2010) propose a mixed-effects LS-SVR where a random-effect term is added to the optimization function of LS-SVR to include random effects in the model.

In this study, we follow the analytic framework by Ngufor et al. (2019), which integrates the random-effects structure of GLMM in non-linear machine learning models, compatible with longitudinal data. While their paper only shows interpretable tree based mixed-effect machine learning models, the framework can easily be extended to other (common) machine learning models.

3 Data

All data for this research is obtained from the Dutch CRC Screening Program during the period 2014-2021. Four rounds of data are available from the biennial screening with the FIT test. Only persons who consistently responded to the invitations for screening and any follow-up examination with colonoscopy were included. Persons who did not respond to one of the invitations were excluded. Persons for whom the results of any follow-up examination was missing are also excluded. Table 1 shows the original variables included in the data set.

Table 1: Original variables in the data set provided by the Erasmus Medical Centre

Variable	**Description**	**Range**
`Age`	Age of respondent at time of screening	55 − 78
`Bloodtest result`	Indicator for result of screening bloodtest	0 (Favourable), 1 (Unfavourable)
`Haemoglobin current`	Hb value	0 − 306
`Haemoglobin threshold`	Threshold value used to determine bloodtest result	275, 88
`ID`	Personal identification	1 − 2,493,999
`Round`	Indicator for presence of individual per round	0 (Participated), 1 (Non-respondent, non-participant)
`Stage current`[1]	Stage of cancer at time of screening	1 (Healthy), 2 (Non-advanced adenoma), 3 (Advanced adenoma), 4 (Cancer stage 1), 5 (Cancer stage 2), 6 (Cancer stage 3), 7 (Cancer stage 4), 8 (Unknown)
`Sex`	Gender of respondent	0 (Male), 1 (Female)

Notes: [1]`Stage current` is one-hot encoded, such that the resulting dummy variables are equal to one for the current stage of cancer, and zero otherwise.

Also, besides using a MEml framework to account for dependencies within the data, I am planning on introducing additional variables: some version of a lagged dependent variable (previous FIT values, also because Grobbee et al. (2017) find that an undetectable Hb concentration two years ago decreases the current risk of having CRC), and the difference between the current and previous test to allow for 'directional trends' so to speak (increase since previous test, or decrease since previous test).

Given that mixed effects models are virtually unexplored for NN and SVM, we must make an active effort to minimally violate the iid assumption. To this end, we introduce the additional variables described in Table 2, to allow for as much individual variation as possible.

Table 2: Variables which are added to the original data set

Variable	Description	Range
`FIT number`[1]	Indicator for sequence number of the FIT test	$1-4$
`Haemoglobin difference`	Difference between current and previously obtained Hb value at time of screening	-306 − 306
`Haemoglobin max`	Maximum obtained Hb value over all tests at time of screening	$0-306$
`Haemoglobin previous`	Previously obtained Hb value at time of screening	$0-306$

Notes: [1]`Fit number` is one-hot encoded, such that the resulting dummy variables are equal to one for the current FIT test, and zero otherwise. [2]`Stage previous` is one-hot encoded in the same way as `Stage current` (see Table 1).

4 Methodology

4.1 Machine learning

4.1.1 Artificial neural networks

Artificial neural networks (ANNs), developed by Lippmann (1987), are inspired by the human brain, mimicking the way that biological neurons signal to one another. ANNs are comprised of (1) an input layer, (2) possibly one or more hidden layers, and (3) an output layer. The input variables are related to the output variable(s) through a network of interconnected nodes, with associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. The optimal values for these weights are estimated when the ANN is fitted, such that a predetermined loss function is minimized – the squared error loss function in our case. The input layer of the ANN consists of p nodes, where p is equal to the number of explanatory variables. In our setting, the output node $\hat{f}(x)$ represents the predicted Hb concentration.

To advance from one layer to another, the ANN uses activation functions $h(\cdot)$, with the sum of the weights and the intercept, referred to as the bias, as input. We compare four commonly used activation functions: the identity function $h(x) = x$, the logistic sigmoid function $h(x) = \frac{1}{1+e^{-x}}$, the hyperbolic tangent function $h(x) = \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$, and the rectified linear unit (ReLu) function $(x) = \max(0, x)$[6].

Hornik et al. (1989) show in their universal approximation theorem that an ANN with at least one hidden layer, and a large enough number of neurons, can approximate any finite-dimensional Borel measurable function up to any arbitrary accuracy. In other words, an ANN with zero hidden layers can only represent linear functions, whereas we can approximate *any* function with a continuous mapping with finite spaces using an ANN with one hidden layer. In practice, however, a network with multiple hidden layers can be more efficient. Therefore, I consider ANNs with both one, and two hidden layers. In case of an ANN with two hidden layers, with H nodes in the first layer and L nodes in the second,

[6]This comparison will take place in a later stage of the research, as the (optimal) activation function depends on the data, and the non-zero nature of the outcome variable should also be taken into account

the values at each node are calculated as follows:

$$z_h^1 = g\left(\sum_{j=1}^{p} w_{hj}^1 x_j\right) \qquad \forall h \in \{1,\ldots,H\},$$
$$z_l^2 = g\left(\sum_{h=1}^{H} w_{lh}^2 z_h^1\right) \qquad \forall l \in \{1,\ldots,L\},$$
$$\hat{f}(x) = g\left(\sum_{l=1}^{L} w_l^3 z_l^2\right),$$

where x_j represents each of the input regressors, z_i^j represents the i^{th} node of the j^{th} hidden layer, and w_{ik}^j is the weight of node k on node i in hidden layer j. Figure 4 shows an example of an ANN with two hidden layers. We use 8-fold[7] cross-validation to determine the number of layers, and nodes in each layer.

One of the risks of Neural Networks is that it tends to overfit on the training data. To mitigate overfitting in the ANNs, we use the efficient early stopping regularization (Prechelt, 1998). We will also explore other regularization terms (Lasso or Ridge) and dropout (Srivastava et al., 2014) as options to minimize overfitting in each ANN.

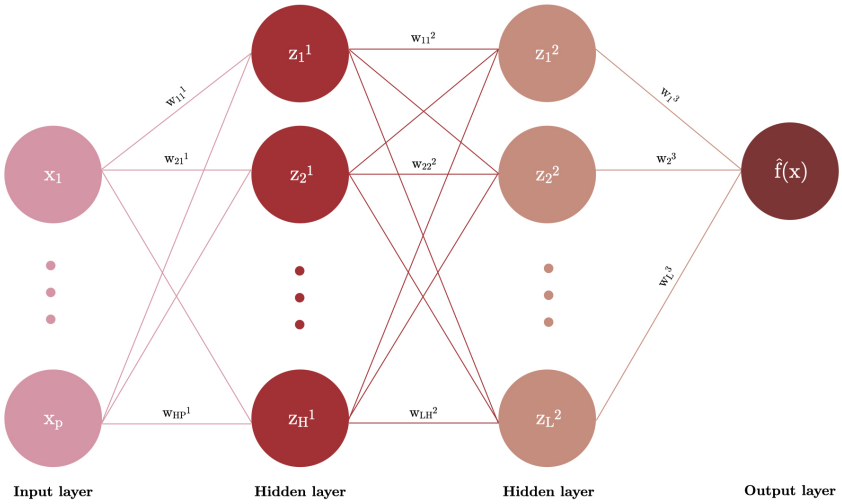

Figure 4: Example of an artificial neural network with two hidden layers and one output node

4.1.2 Support vector regression

The second algorithm is based on Support Vector Machines (SVMs). SVMs separate binary classified data using a hyperplane as decision boundary such that the margin between the classes is maximised (Cortes and Vapnik, 1995). To this end, the input variables x_j are transformed into an m-dimensional feature space using a non-linear mapping, after which the SVM algorithm searches for the optimal separating hyperplane represented by a set of support vectors – the data points on either side of the hyperplane that are closest to the hyperplane.

The use of a kernel implicitly maps the input vector to higher dimensional feature spaces, where the

[7]Eight folds are chosen to efficiently parallelize across four CPU's.

problem becomes a linear surface that fits the data, which allows for SVMs to handle highly non-linear data by using kernels. SVMs model non-perfectly separable data through the introduction of soft margins, where some slack is allowed for observations to be on the wrong side of the margin.

The sparse solution and good generalization of the SVM lend themselves to adaptation to regression problems: Support Vector Regression (SVR) (Awad and Khanna, 2015). Support Vector Regression uses the same principle as the SVMs, but predicts discrete values. The basic idea behind SVR is to find the best fit line within a threshold value ε. To this end, we introduce the ε-tube – an ε-insensitive region around the function – which reformulates the optimization problem to find the tube that best approximates the continuous-valued function, while balancing model complexity and prediction error. More specifically, SVR looks for the flattest tube that contains most of the training instances.

The fit time complexity of SVR is more than quadratic with the number of samples, thus for large data sets, Linear SVR is preferred – which provides a faster implementation than SVR, as it only considers the linear kernel. Consequently, this is the employed method for this research.

4.2 Mixed-effects machine learning

In a general GLMM framework, the model assumes that the responses y_{it} for a single subject i, conditional on an (assumed iid normal) subject-specific risk factor γ_i, are independent and follow a distribution from the exponential family with mean: $E(y_{it}|\gamma_i) = \mu_{it} = h(\eta_{it})$, where $\eta_{it} = \beta' x_{it} + \gamma_i$, where $g(\cdot) = h^{-1}(\cdot)$ represents the link function and β represents the vector of population fixed-effect coefficients. The GLMM assumes a parametric distribution and imposes restrictive linear relationships between the link function $g(\cdot)$ and the covariates. Machine learning algorithms do not make *a priori* assumptions on the distribution, but they do often implicitly make the iid assumption.

Ngufor et al. (2019) propose a MEml framework, which estimates the fixed-effects component ($\beta' x_{it}$) using machine learning algorithms. Thus η_{it} is now defined as

$$\eta_{it} = f(x_i) + \gamma_i, \tag{1}$$

with estimated dependent variable

$$y_i = f(x_i) + \gamma_i + \varepsilon_i, \tag{2}$$

where the function $f(\cdot)$ is unknown, and must be estimated. While Ngufor et al. (2019) use only tree based algorithms to estimate $f(\cdot)$, they state that any supervised learning algorithm can be used. In turn, this research contributes to the existing literature by using both ANNs and SVRs in this MEml framework. The proposed MEml models are estimated using the expectation-maximization approach, in which Equation 1 and 2 are alternatively estimated. In essence, we first initialize the random effects $\hat{\gamma}_i = 0$, and use this $\hat{\gamma}_i$ to compute $y^*_{it} = y_{it} - \hat{\gamma}_i$. We then train our machine learning model to estimate $\hat{f}(x_{it})$ in Equation 2 using y^*_{it}. Finally, we estimate γ_i in Equation 1 using $\hat{f}(x_{it})$. This process repeats until convergence[8].

[8]For more details on the estimation procedure, we refer to Ngufor et al. (2019).

4.3 Tuning

As with most machine learning methods, the performance of both ANNs and SVRs are dependent on proper tuning. Due to the large dimensionality of the parameter grid, we cross-validate the hyperparameters of the different models using a Bayesian Randomized search (Bergstra et al., 2013). This method first explores the parameter space and then performs a guided search in (seemingly) promising subspaces in terms of cross-validated accuracies. The Hyperopt method can be seen as an exploration/exploitation strategy, that starts by exploring the performance across the candidate hyperparameter space, and subsequently randomly exploits the most promising subspace of hyperparameters. For the same number of iterations, this method can lead to better hyperparameter settings than the ones of random search.

For computational efficiency, Putatunda and Rama (2018) introduce Randomized Hyperopt. This method first randomly samples a predetermined fraction $\rho \in [0, 1]$ from the validation train fold without replacement, and then performs a Hyperopt iteration on this sampled fold. In their application, they show that the loss in performance is limited, while drastically decreasing computation time, allowing for more Hyperopt iterations. We employ Randomized Hyperopt with eight folds.

For the ANNs, we tune the number of hidden layers, dropout rate, early stopping, number of neurons, batch size, and the learning rate. We do not consider weight decay since we already account for overfitting with early stopping and dropout. For the SVRs we tune the kernel, degree of non-linearity, regularization parameter, and ε.

In addition, normalization might be necessary as Jayalakshmi and Santhakumaran (2011) show that the performance of NN is contingent on normalization of the explanatory variables. For SVMs, Herbrich and Graepel (2000) show that normalisation of the feature vectors leads to increased performance as well. We consider four distinct normalization schemes: no normalization, min-max normalization, standardization, and robust standardization using the median and $25\% - 75\%$ interquantile range.

4.4 Performance measures

The root mean squared error (RMSE), mean absolute error (MAE), and median absolute error (MedAE) are used to assess individual predictions. We use the Diebold-Mariano (DM) test to determine if model A generates significantly better predictions than model B.

4.5 Class imbalance

Furthermore, since the data is zero-inflated, and therefore highly unbalanced, we explore state-of-the-art rebalancing techniques, either using cost functions (assuming expert knowledge is available), or using a combination of SMOTE-NC (Chawla et al., 2002), along with either ENN (Wilson, 1972), Tomek Links (Tomek, 1976), or NearMiss, depending on computational feasibility within time constraints.

References

Awad, M. and Khanna, R. (2015). Support Vector Regression. In *Efficient Learning Machines*, pages 67–80. Springer.

Barton, M. B., Moore, S., Polk, S., Shtatland, E., Elmore, J. G., and Fletcher, S. W. (2001). Increased patient concern after false-positive mammograms. *Journal of General Internal Medicine*, 16(3):150–156.

Bergstra, J., Yamins, D., and Cox, D. D. (2013). Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. In *Proceedings of the 12th Python in Science Conference*, volume 13, page 20. Citeseer.

van den Berg, D. (2021). Simulation of haemoglobin concentrations in MISCAN-Colon using a mixed-effect machine learning model. Master's thesis, Erasmus University Rotterdam.

Botteri, E., Iodice, S., Bagnardi, V., Raimondi, S., Lowenfels, A. B., and Maisonneuve, P. (2008). Smoking and Colorectal Cancer: A Meta-analysis. *Journal of the American Medical Association*, 300(23):2765–2778.

Brasso, K., Ladelund, S., Frederiksen, B. L., and Jørgensen, T. (2010). Psychological distress following fecal occult blood test in colorectal cancer screening–a population-based study. *Scandinavian Journal of Gastroenterology*, 45(10):1211–1216.

Brenner, H., Stock, C., and Hoffmeister, M. (2014). Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. *British Medical Journal*, 348.

Brodersen, J. and Siersma, V. D. (2013). Long-Term Psychosocial Consequences of False-Positive Screening Mammography. *The Annals of Family Medicine*, 11(2):106–115.

Bronner, M. P. and Haggitt, R. C. (1993). The Polyp-Cancer Sequence: Do All Colorectal Cancers Arise from Benign Adenomas? *Gastrointestinal Endoscopy Clinics of North America*, 3(4):611–622.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, S. and DuBois Bowman, F. (2011). A Novel Support Vector Classifier for Longitudinal High-dimensional Data and its Application to Neuroimaging Data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(6):604–611.

Cheng, Q., Tezcan, J., and Cheng, J. (2014). Confidence and prediction intervals for semiparametric mixed-effect least squares support vector machine. *Pattern Recognition Letters*, 40:88–95.

Cho, D.-H. (2010). Mixed-effects LS-SVR for longitudinal data. *Journal of the Korean Data and Information Science Society*, 21(2):363–369.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Machine Learning for Healthcare Conference*, pages 301–318. Proceedings of Machine Learning Research.

Compton, C. C. and Greene, F. L. (2004). The Staging of Colorectal Cancer: 2004 and Beyond. *CA: A Cancer Journal for Clinicians*, 54(6):295–308.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Ding, H., Lin, J., Xu, Z., Chen, X., Wang, H. H., Huang, L., Huang, J., Zheng, Z., and Wong, M. C. (2022). A Global Evaluation of the Performance Indicators of Colorectal Cancer Screening with Fecal Immunochemical Tests and Colonoscopy: A Systematic Review and Meta-Analysis. *Cancers*, 14(4):1073.

Du, W., Cheung, H., Johnson, C. A., Goldberg, I., Thambisetty, M., and Becker, K. (2015). A Longitudinal Support Vector Regression for Prediction of ALS Score. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1586–1590. IEEE.

Fazio, V. W., Tekkis, P. P., Remzi, F., and Lavery, I. C. (2004). Assessment of operative risk in colorectal cancer surgery: the Cleveland Clinic Foundation colorectal cancer model. *Diseases of the Colon & Rectum*, 47(12):2015–2024.

Grobbee, E. J., Schreuders, E. H., Hansen, B. E., Bruno, M. J., Lansdorp-Vogelaar, I., Spaander, M. C., and Kuipers, E. J. (2017). Association Between Concentrations of Hemoglobin Determined by Fecal Immunochemical Tests and Long-term Development of Advanced Colorectal Neoplasia. *Gastroenterology*, 153(5):1251–1259.

Habbema, J., van Oortmarssen, G., Lubbe, J. T. N., and van der Maas, P. (1985). The MISCAN simulation program for the evaluation of screening for disease. *Computer Methods and Programs in Biomedicine*, 20(1):79–93.

Haghani, S., Sedehi, M., and Kheiri, S. (2017). Artificial Neural Network to Modeling Zero-inflated Count Data: Application to Predicting Number of Return to Blood Donation. *Journal of Research in Health Sciences*, 17(3):392.

Herbrich, R. and Graepel, T. (2000). A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Hewitson, P., Glasziou, P., Watson, E., Towler, B., and Irwig, L. (2008). Cochrane Systematic Review of Colorectal Cancer Screening Using the Fecal Occult Blood Test (Hemoccult): An Update. *Journal of the American College of Gastroenterology*, 103(6):1541–1549.

Holme, Ø., Bretthauer, M., Fretheim, A., Odgaard-Jensen, J., and Hoff, G. (2013). Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals (Review). *Cochrane Database of Systematic Reviews*.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer Feedforward Networks Are Universal Approximators. *Neural networks*, 2(5):359–366.

Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering*, 3(1):1793–8201.

Jenniskens, K., De Groot, J. A., Reitsma, J. B., Moons, K. G., Hooft, L., and Naaktgeboren, C. A. (2017). Overdiagnosis across medical disciplines: a scoping review. *BMJ Open*, 7(12):e018448.

Jiang, Y., Yuan, H., Li, Z., Ji, X., Shen, Q., Tuo, J., Bi, J., Li, H., and Xiang, Y. (2022). Global pattern and trends of colorectal cancer survival: a systematic review of population-based registration data. *Cancer Biology & Medicine*, 19(2):175.

Levin, B., Lieberman, D. A., McFarland, B., Andrews, K. S., Brooks, D., Bond, J., Dash, C., Giardiello, F. M., Glick, S., Johnson, D., et al. (2008). Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline From the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology*, 134(5):1570–1595.

Lippmann, R. (1987). An Introduction to Computing with Neural Nets. *IEEE ASSP magazine*, 4(2):4–22.

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., and Habbema, J. D. F. (1999). The MISCAN-COLON Simulation Model for the Evaluation of Colorectal Cancer Screening. *Computers and Biomedical Research*, 32(1):13–33.

Luts, J., Molenberghs, G., Verbeke, G., van Huffel, S., and Suykens, J. A. (2012). A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis*, 56(3):611–628.

Mandel, F., Ghosh, R. P., and Barnett, I. (2021). Neural networks for clustered and longitudinal data using mixed effects models. *Biometrics: A Journal of the International Biometric Society*.

Morson, B. (1974). The polyp-cancer sequence in the large bowel. *Journal of the Royal Society of Medicine*, 67:451–457.

Mousavinezhad, M., Majdzadeh, R., Sari, A. A., Delavari, A., and Mohtasham, F. (2016). The effectiveness of FOBT vs. FIT: A meta-analysis on colorectal cancer screening test. *Medical Journal of the Islamic Republic of Iran*, 30:366.

Ngufor, C., van Houten, H., Caffo, B. S., Shah, N. D., and McCoy, R. G. (2019). Mixed Effect Machine Learning: A framework for predicting longitudinal change in hemoglobin A1c. *Journal of Biomedical Informatics*, 89:56–67.

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767.

Putatunda, S. and Rama, K. (2018). A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. In *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, pages 6–10.

Sakthivel, K. and Rajitha, C. (2017). A Comparative Study of Zero-inflated, Hurdle Models with Artificial Neural Network in Claim Count Modeling. *International Journal of Statistics and Systems*, 12(2):265–276.

Seok, K. H., Shim, J., Cho, D., Noh, G.-J., and Hwang, C. (2011). Semiparametric mixed-effect least squares support vector machine for analyzing pharmacokinetic and pharmacodynamic data. *Neurocomputing*, 74(17):3412–3419.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

van der Steeg, A., Keyzer-Dekker, C., De Vries, J., and Roukema, J. (2011). Effect of abnormal screening mammogram on quality of life. *Journal of British Surgery*, 98(4):537–542.

Strum, W. B. (2016). Colorectal Adenomas. *New England Journal of Medicine*, 374(11):1065–1075.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Tandon, R., Adak, S., and Kaye, J. A. (2006). Neural networks for longitudinal studies in Alzheimer's disease. *Artificial Intelligence in Medicine*, 36(3):245–255.

Thanikachalam, K. and Khan, G. (2019). Colorectal Cancer and Nutrition. *Nutrients*, 11(1):164.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions Systems, Man and Cybernetics*, 6:769–772.

Toribara, N. W. and Sleisenger, M. H. (1995). Screening for Colorectal Cancer. *New England Journal of Medicine*, 332(13):861–867.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global Cancer Statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108.

Wardle, J., Williamson, S., Sutton, S., Biran, A., McCaffery, K., Cuzick, J., and Atkin, W. (2003). Psychological Impact of Colorectal Cancer Screening. *Health Psychology*, 22(1):54.

Welch, H. G. and Black, W. C. (2010). Overdiagnosis in cancer. *Journal of the National Cancer Institute*, 102(9):605–613.

Whitlock, E. P., Lin, J. S., Liles, E., Beil, T. L., and Fu, R. (2012). Screening for Colorectal Cancer: A Targeted, Updated Systematic Review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 157(2):120–134.

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421.

Xiong, Y., Kim, H. J., and Singh, V. (2019). Mixed Effects Neural Networks (MeNets) With Applications to Gaze Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zorzi, M., Fedeli, U., Schievano, E., Bovo, E., Guzzinati, S., Baracco, S., Fedato, C., Saugo, M., and Dei Tos, A. P. (2015). Impact on colorectal cancer mortality of screening programmes based on the faecal immunochemical test. *Gut*, 64(5):784–790.