Filip Emil Schjerven

Prediction models for hypertension using the HUNT Study data

June 2020

Master's thesis
2020
Filip Emil Schjerven

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of Science and Technology

NTNU
Norwegian University of Science and Technology

NTNU
Norwegian University of
Science and Technology

Prediction models for hypertension using the HUNT Study data

Filip Emil Schjerven

Mathematical Sciences
Submission date: June 2020
Supervisor: Ingelin Steinsland
Co-supervisor: Frank Lindseth

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

In this thesis we compare different model-families' ability to predict the 11 year binary hypertension status, using data from the Trøndelag Health Study, HUNT. The model-families used are that of logistic regression, random forest and neural networks. The goal of each prediction model was to predict the risk of hypertension at the time of HUNT-3 for otherwise healty people at HUNT-2, using measurements taken at HUNT-2.

First, a literature review was conducted to assess the current status of research on hypertension risk prediction models. It was not possible to determine that one model family should be better than the others based on the included literature.

With the relevant features identified from the literature study, a subset of relevant data was extracted from the available HUNT data. The final dataset consisted of $n = 18249$ participants and $p = 19$ features. An exploratory analysis of the dataset showed that 'Systolic BP.', 'Diastolic BP.' and 'Age' are the features most correlated with the hypertension status at HUNT-3. 'Cholesterol', 'Hypertension history in close family' and physical characteristics, like 'Waist-circumference', were also notable.

A repeated training and testing scheme was used to obtain performance distributions for the three model-families. Along with the performance distribution, the Framingham model was evaluated on the datasubset that matched the features used in the Framingham model. All models were evaluated by the area under the Receiver-Operator-Curve and the Precision-Recall-Curve, a modified Brier score and a score named Tjur's R^2.

We conclude that the variability in the dataset had a greater effect than the choice of model-family on the performance measures, as the differences between model-families was smaller than the difference within each model-family. The results suggests that if non-linear effects exists in the data at all, they have little additional predictive power compared to the linear effects. Further, a subset of particularly important features was identified by importance scores. Repeating the analysis using only these features for the logistic regression and random forest model-families produced scores that were equally good as using the full feature set for these model families.

The results for all models and feature sets used were comparable to those obtained by the Framingham model and to the relevant literature. Finally, taking into account model properties, a logistic regression model using the features 'Systolic BP.', 'Diastolic BP.', 'Age', 'Waist-circumference' and 'Hypertension history in close family', fitted with some regularization, but without balanced loss, is proposed as the optimal modelling setup for this problem. For future work, analysis of datasubsets where the models were highly wrong or disagreed across model-families is suggested, along with a bias assessment of the literature on hypertension risk models.

Sammendrag

I denne oppgaven sammenlignes forskjellige modellfamiliers evne til å predikere 11-års risikoen for binær hypertensjon status, ved bruk av data fra helseundersøkelsen i Trøndelag, HUNT. Modellfamiliene som ble valgt var logistisk regresjon, nevrale netverk og random forest. Målet for hver enkelt modell var å predikere risikoen for hypertensjon ved HUNT-3 studien for individer som var friske ved HUNT-2, ved bruk av målinger tatt i HUNT-2.

Til å begynne med ble det gjennomført et litteraturstudie for å få oversikt over forskningen på risiko modeller for hypertensjon. Det var ikke mulig å fastslå at en av modellfamiliene skulle være bedre enn de andre basert på litteraturstudiet.

Etter å ha identifisert de relevante attributtene i litteraturstudiet, ble et subsett av relevante data valgt ut fra det tilgjengelige HUNT datasettet. The endelige datasettet hadde $n = 18249$ individer og $p = 19$ attributter. En utforskende analyse av datasettet viste at 'Systolisk blodtrykk', 'Diastolisk blodtrykk' og 'Alder' var de attributtene som var mest korrelerte med hypertensjon-status ved HUNT-3. 'Kolesterol', 'Familiehistorie med hypertensjon' og fysiske attributter som 'Midjemål' var også verdt å nevne.

Et repetert trening og testing oppsett ble brukt for å produsere fordelinger av ytelsesmål for de tre modellfamiliene. I tillegg ble Framingham modellen evaluert på et subsett av data hvor attributtene passet og var tilgjengelig. Alle modellene ble evaluert med området under Receiver-Operator-kurven og Precision-Recall-kurven, samt et modifisert Brier mål og et mål kalt Tjur's R^2.

Vi konkluderer med at ytelsen var mer påvirket av variabiliteten i datasettet enn valget av modellfamilie, ettersom det var større forskjell innad i fordelingene enn mellom modellfamilier. Resultatene antyder at hvis det er noen ikke-lineære effekter, så har de lite ekstra prediktiv kraft sammenlignet med lineære. Videre ble et subsett av attributtene identifisert som særdeles viktige vha. viktighetsmål. En gjentagelse av analysen med dette subsettet i logistisk regresjon og random forest ga ytelsesmål som var like gode som ved bruk av alle attributtene for disse modellfamiliene.

Resultatene fra alle modellfamiliene og attributsettene brukt var sammenlignbare med det som Framingham modellen oppnådde og til den relevante litteraturen. Til slutt, ved å ta hensyn til egenskaper til modellene, så ble den logistiske modellfamilien som bruker 'Systolisk blodtrykk', 'Diastolisk blodtrykk', 'Alder', 'Midjemål' og 'Familiehistorie med hypertensjon' som attributter, tilpasset med regularisering, uten balansert tapsfunksjon, foreslått som det optimale modelloppsett for problemet. For videre arbeid ble det foreslått å analysere subsett av data hvor modellene predikerte store feil eller var uenige på tvers av modellfamilier, i tillegg til å gjennomføre en subjektivitets-analyse av litteraturen som omhandler hypertensjon risiko modeller.

Preface

This thesis was written under the supervision of Prof. Ingelin Steinsland and Prof. Frank Lindseth. I am grateful for their patience, the helpful advice and for them continuously challenging me.

Firstly, I would also like to thank the HUNT Cloud team for help with using the HUNT Cloud client, ensuring that the HUNT data was properly handled. Secondly, I would like to thank my colleague Emma Ingeström for her helpful advice and assistance with this thesis. I would like to thank Fride Nordstrand Nilsen for valuable discussions on statistics and thesis-structure.

I would like to thank my friends Magnus, Didrik, Eiolf, Morten, Peter and all the other members of the 'green couch' lunch group for the many fun mathematical discussions we have had.

Lastly, I would like to thank my family for their never-ending support.

Filip Schjerven,
June 2020

Contents

Chapter 1

Introduction

Hypertension is a disease that is estimated to affect more than 1.1 billion people all over the world [1]. It is estimated that hypertension is the cause of over 8 million deaths each year on a global scale, and increasing. This means that hypertension is one of the most prevalent cause of human deaths worldwide, at a staggering 14% of all deaths [2] [3]. The World Health Organization (WHO) classifies hypertension as the singlemost important risk-factor for being subject to an early death or serious disease [4]. Its impact is also reflected in economic terms: It is estimated that 10% percent of the worlds health expenditure can be linked to hypertension [5].
Hypertension in itself is a complex disease that rarely can be traced back to a single cause. Risk-factors include both genetic causes as well as lifestyle choices. In addition to its complexity, hypertension is often dubbed "the silent killer" due to its lack of clearly noticeable symptoms. However, the disease is considered treatable, either by changes in lifestyle-choices, medication or a combination. Accesibility and usage of health resources are therefore important for detecting and initializing treatment of hypertension [6].

The topic of this thesis was chosen in relation to a project called ”A Digital Twin For Essential Hypertension Management And Treatment - My Medical Digital Twin”[1], MyMDT for short. It is a multi-diciplinary research effort involving PhD students and researchers from Medicine, Mathematics, Biomechanics, Computer Science and Mechanical engineering at the Norwegian Univerisity of Science and Technology (NTNU). The research group is led by Prof. Ulrik Wisloeff. The research project is focused on developing a personalized medical digital tool that gives insight into an individual’s blood pressure. This digital tool will be produced by merging mathematical models derived from population-data with mathematical models derived from personal sensor data, collected by custom-made sensors. As an important source of population data, a

[1] https://www.ntnu.no/cerg/mymdt

large population study called HUNT is utilized.

The HUNT Study is a large population study used for medical and health-related research. HUNT is an acronym for the study's norwegian name: ***H****else****u****ndersøkelsen i* ***N****ord-****T****røndelag*. As the name suggests, the study population is dervied from the county of Troendelag in Norway. The study includes cohorts from the 1980's, starting with the health survey HUNT-1 (1984-86), covering over 125 000 participants. These cohorts were followed with health surveys, conducted every 11 years: HUNT-2 (1995-97) and HUNT-3 (2006-08) [7]. There is a more recent survey conducted, named HUNT-4, but neither data nor any results have been published or made available at the time of writing.
The motivation behind the HUNT study was primarily to address arterial hypertension, diabetes, chest X-ray screening of tuberculosis and quality of life. The scope of the study has expanded since then, becoming an important data source for the purposes of gaining knowledge on numerous effects, causes and associations in medical science. For the purpose of this thesis, it is data collected in health surveys HUNT-2 and HUNT-3 that is used.
The aim of this thesis is to construct and evaluate predictive models for hypertension, based on HUNT-2 and HUNT-3 study. In particular we want to compare prediction models based on logistic regression, random forest and neural networks. Further we want to explore if an ensembling method, regularization methods and class weight scaling improve the predictive performance of the resulting prediction models. Secondary aims include reviewing the current state of literature on predictive models for hypertension where similar model-families have been applied and comparing the results found in literature to those achieved in this thesis.

To construct prediction models, the model-families of logistic regression, random forests and neural networks were chosen for their differing model properties. Logistic regression models are simple, yet effective and easily understandable models. It is therefore a popular choice for modelling problems with binary outcomes. While a logistic regression model is capable of capturing the linear effects of its input, neural network and random forest models are capable of modelling the non-linear effects of its inputs. The motivation for using both neural network and random forest models are that by their construction, the non-linearities they capture can be quite dissimilar. In total, the three families were chosen to complement each other. This should allow the analysis to be able to capture a wide span of different patterns if they are present in the data.
Modelling hypertension risk and creating a well-performing prediction model for hypertension is a goal of this thesis as well as for the MyMDT project. While HUNT data has 11 years between health surveys and measurements, the MyMDT project will use real-time measurements obtained from sensors. A well-performing risk prediction

model derived from the HUNT data can inform the work that is done in MyMDT, by providing a benchmark and inform the priors of datamodels. In addition, a last contribution of this thesis is to validate an existing risk prediction model, the Framingham model [8], on the HUNT data. This model has not been validated on a Scandinavian population before.

Since the topic of this thesis touches upon different fields, it should be noted that some terms are used that have different names in different fields. One of these is 'feature', which is also called predictor, explanatory variable, independent variable, covariate, a risk factor, and more. Another is 'target', which is also called response, predicted variable, dependent variable, outcome, label, and more. Another important term is 'binary', which is called dichotomous in some fields. There are possibly more, but these are the most important ones.

In Chapter 2, the methods for constructing the prediction models are outlined. A literature review on hypertension risk models is given in Chapter 3. How relevant data was selected from the HUNT data and an exploratory analysis is detailed in Chapter 4, before the setup for the analysis is given in Chapter 5. The results from the analysis is presented in Chapter 6. A discussion follows in Chapter 7 before a conclusion and suggestions for future work is given in Chapter 8.

Chapter 2

Background

2.1 Data learning methods

2.1.1 Generalization error vs. training error

The error of a data-learning model can be divided into two categories: Its training error and its generalization error. The training error is the error of a model as measured on the data that is used to fit the model, while the generalization error is the performance of a model as measured on unseen data [9]. To get an unbiased estimate of the generalization error, it is common to divide the dataset into a training set and a test set. The training set is used to fit the model, while the test set *is only used* to measure the error of the fitted model, as a estimate of its generalization error.

Ideally, a model has enough flexibility to fit the general patterns of the data, without fitting to random noise patterns that may be present. If a model is not flexible enough, the model would be *underfitting* the data, not capturing the general patterns. Hence, both generalization and training error would be higher than necessary. If a model is too complex, it may be *overfitting*, fitting to random noise patterns. This is a simplistic view of the differences between generalization error and training error, but it suffices as a motivation for the usage of regularisation methods.

To mitigate the risk of overfitting the model to its training data, regularisation methods are applied. Regularisation in this context is defined as *"... is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error"* [9]. For each model-family, the regularisation method used with it is described.

2.1.2 Notation and terms

Assume n datapoints and p features. Each datapoint is described as a tuple $(\boldsymbol{x}_i, y_i)$ for $i \in \{1, \ldots, n\}$, where $\boldsymbol{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \ldots & x_{ip} \end{bmatrix}$ is a row-vector encoding the features

and y_i encodes the associated target value.

All datapoints can be described as a $n \times p$ feature matrix $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n\end{bmatrix}^T$ associated with a $n \times 1$ target vector $\boldsymbol{y} = \begin{bmatrix}y_1 & y_2 & \dots & y_n\end{bmatrix}$. Each $y_i \; \forall i$ is modelled to be a realization of a Bernoulli-distributed random variable Y_i, assumed independent of all other $Y_j, i \neq j$. The probability parameter π_i is conditional on y_i's associated features $\boldsymbol{x}_i$. A Bernoulli distribution has probability mass function:

$$f_{Y_i}(y_i|\pi_i) = \begin{cases} 1 - \pi_i, & y_i = 0 \\ \pi_i, & y_i = 1 \end{cases}$$

The probability parameter π_i is assumed to be a function of the features, i.e. $\pi_i = \mathcal{P}(y_i = 1|\boldsymbol{x}_i) = g(\boldsymbol{x}_i)$. If a model assumes a functional form for the function $g(.)$, the model is referred to as a *parametric* method. In that case, fitting a model entails fitting the parameters of the assumed function. If a model approximates the function $g(.) \approx \hat{g}(.)$ directly, the model is *non-parametric*. Regardless, the prediction of π_i is denoted $\hat{\mathcal{P}}(y_i = 1|\boldsymbol{x}_i) = \hat{y}_i$ by convention.

The model-families used to produce $\hat{y}_i$ is that of logistic regression, random forest and neural networks. For simplicity, each method is described for a single datapoint, the tuple $(\boldsymbol{x}, y)$ and the predicted probability parameter $\hat{y}$.

Note that the parameters for a model family that is set before modellfitting is referred to as that models *tuningparameters*.

2.1.3 Logistic regression

This section is based upon chapter 5.1 in [10]. To accomodate an intercept in the model, $\boldsymbol{x}$ is extended as $\boldsymbol{x}^* = \begin{bmatrix}1 & \boldsymbol{x}\end{bmatrix}$, i.e. with a 1 appended as the first vector value.

Logistic regression is a type of generalized linear model, useful for modelling relationships between features and binary target values. It is a parametric method, relating $\hat{y}$ with features $\boldsymbol{x}^*$ by a link function h. In the case of logistic regression, this link function is the logit function. Defining $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \dots & \beta_p\end{bmatrix}$ as a vector parameters, logistic regression models the relation between $\hat{y}$ and $\boldsymbol{x}^*$ as

$$\begin{aligned} g(\hat{y}) &= \boldsymbol{x}^*\boldsymbol{\beta}^T \\ \implies \text{logit}(\hat{y}) = \log(\frac{\hat{y}}{1-\hat{y}}) &= \boldsymbol{x}^*\boldsymbol{\beta}^T = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \end{aligned} \tag{2.1}$$

Th parameters β_j are called coefficients by convention. They assign weight to the different features $x_j \; j \in \{1, \dots, p\}$. The intercept, β_0, is a base effect applied regardless of the values of $\boldsymbol{x}^*$. To produce a prediction $\hat{y}$, the inverse of the logit is applied to $\boldsymbol{x}^*\boldsymbol{\beta}^T$:

$$\hat{y} = \texttt{Sigmoid}(\boldsymbol{x}^*\boldsymbol{\beta}^T) = \frac{e^{\boldsymbol{x}^*\boldsymbol{\beta}^T}}{1 + e^{\boldsymbol{x}^*\boldsymbol{\beta}^T}} \tag{2.2}$$

The Sigmoid function is shown in Figure 2.2.

An important reason for the popularity of logistic regression is seen by investigating the odds of event probabilities: $\frac{\hat{y}}{1-\hat{y}} = e^{\boldsymbol{x}^*\boldsymbol{\beta}^T} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} = e^{\beta_0}\prod_{j=1}^{p} e^{\beta_j x_j}$. Not only is this easy to do calculations on, but it also gives an intuitive interpretation of changes in features. Suppose an increase in feature x_j of 1, a positive or negative β_j would give an increasing or decreasing effect on the odds.

Fitting a logistic regression model entails estimating the coefficient vector $\boldsymbol{\beta}$, often done by using the maximum likelihood principle with respect to a set of datapoints $(\boldsymbol{X}, \boldsymbol{y})$. A commonly used loss-function that is optimized to find estimates of $\boldsymbol{\beta}$ is the binary cross-entropy loss function:

$$L(\hat{\boldsymbol{y}}) = -\sum_{i=1}^{n} y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i) \tag{2.3}$$

Although the solution $\hat{\boldsymbol{\beta}}$ is not available in closed form, the optimization problem is *convex*. This means that any locally best solution is the unique global solution [11]. Hence, the numerical procedure is guaranteed to asymptotically find the optimal solution of $\hat{\boldsymbol{\beta}}$ that minimizes equation 2.3 as the number of datapoints increases. This implies that a logistic regression model can produce the best possible prediction model for *linear* feature effects.

2.1.3.1 Regularisation: Lasso, Ridge and Elastic loss

This section is based upon chapter 6.2.2 in [11] as well as [9].

Lasso and Ridge regularisation are types of regularisation that modifies the loss-function that the model is fitted to minimize. The modification is to add the norm of parameters in a model to the loss. For Lasso, the norm is the L1 norm. For Ridge, the norm is the L2 norm. The norm is scaled by a tuningparameter λ to control the intensity of the regularisation. Denoting a set of parameters in a model as $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \dots & \beta_p\end{bmatrix}$, the penalty terms added to the loss function is:

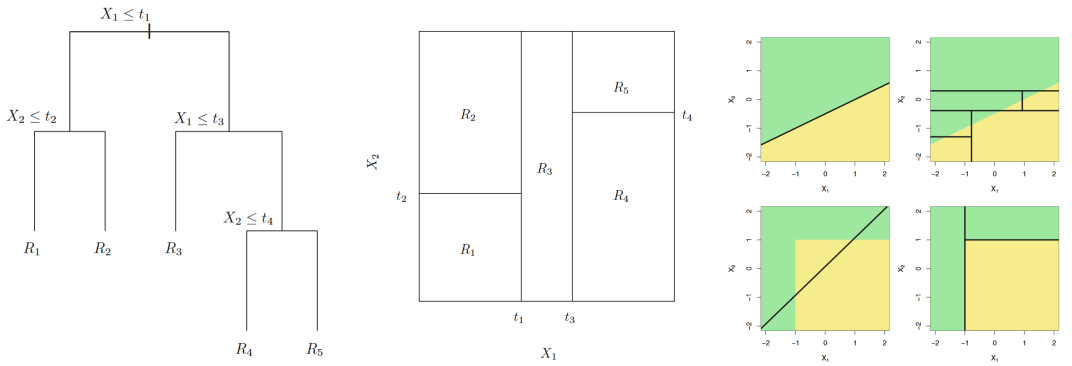
$$\text{Lasso penalty: } \lambda||\boldsymbol{\beta}||_1 = \lambda\sum_{i=0}^{p}|\beta_i|, \quad \text{Ridge penalty: } \lambda||\boldsymbol{\beta}||_2 = \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} = \lambda\sum_{i=0}^{p}\beta_i^2 \tag{2.4}$$

with λ as a tuningparameter determining the strength of regularization. Lasso loss has the property that it encourages sparse solutions, i.e. coefficients may be set to zero, while Ridge regression encourages parameters to be more equal in absolute size.

The Elastic loss is simply a mixed Lasso and Ridge loss, where the ratio is controlled by an another tuningparameter $\gamma \in [0, 1]$:

$$\text{Elastic loss: } \lambda\big(\gamma||\boldsymbol{\beta}||_1 + (1-\gamma)||\boldsymbol{\beta}||_2\big) \tag{2.5}$$

For the cases $\gamma = 0$ or $\gamma = 1$, the elastic loss reduces to the Ridge or Lasso loss respectively.



(a) Basic example of a decision tree using two predictors X_1, X_2.

(b) 2 dimensional representation of the predictor space for the tree in Figure 2.1a, with regions detailing the predicted decision.

(c) Linearly separable regions in top row, regions ideal to split by decision trees in bottom row. Example decision boundaries for logistic regression and decision trees are shown.

Figure 2.1: Decision trees, decision regions and comparison between optimal logistic regression and decision tree spaces. Adapted from [11].

2.1.4 Random Forest

This section is based upon [12], [13] and [11].

The Random Forest model-family is a non-parametric model-family, quite different from logistic regression and neural networks. Random forest models are applicable for both classification and regression tasks, but is only described for the classification-case. In essence, random forest models produce ensembles of decision trees using a combination of bagging and a stochastic method for fitting decorrelated decision trees.

The basic building block of a random forest is the decision tree. An example of a simple decision tree is shown in Figure 2.1a. In Figure 2.1b, the decision regions corresponding to the endnodes of Figure 2.1a is shown. In Figure 2.1c, a comparison of two target spaces is shown. The figures in the top row are ideal for separation using logistic regression on the axis values, while the figures in the bottom row are ideal for separation using decision trees. Notice that the methods allows for quite different separation boundaries.

A decision tree is read by starting at the top node, and moving down the branches to the left if the datapoints feature satisfy the criterion, and to the right if it does not. When a terminal node, also called a end-node or leaf, is reached, the decision is the label of that leaf.

As an example, given a datapoint $(X_1, X_2) = (t_1 - 1, t_2 + 1)$ and the tree in 2.1a, the decision path goes left at the top node, and right at the subsequent node, resulting in decision R_2. Single decision trees are highly interpretable, as the mechanics are easily explained.

A decision tree is built by considering two steps: 1) Dividing the feature space into non-overlapping regions, and 2) For every decision that reaches a specific region, associate the same category or predict the same value.

Decision trees are learned by a *top-down, greedy* approach. This means that we start building our decision trees from the "top" where all the training data is in one region. A split is chosen if it is optimal *at that node*, not considering what implications this split may have for later splits.

The measure used to quantify the usefulness of a split, is commonly the Gini index, G:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.6}$$

where $\hat{p}_{mk}$ is the proportion of training data in the mth region that are from the kth target category. Notice that this criteria corresponds to the variance of K independent Bernoulli distributed variables. The Gini index is often referred to as a measure of *impurity*, as values of $\hat{p}_{mk}$ close to 0 or 1 will give a low Gini index. A high Gini index suggests a split will produce a region with more datapoints of mixed target categories, hence "impurity".

Modelfitting stops when each terminal node has lower than a fixed number of training datapoints in its region. One could also stop training by the number of splits or stop when all possible splits yield a reduction in impurity lower than a predefined value. All of these choices involve tuningparameters.

In the case of a classification problem, a simple way of producing a prediction for datapoints that end up in a region, is to classify them the target catecory that the most training datapoints in that region belong to. An alternative used in this thesis is to predict target probabilities equal to the distribution of target categories in the training data belonging to that region.

Ensembles are in its simplest form a way of constructing a prediction model from an average of multiple prediction models:

$$\hat{y} = \hat{f}_{Ensemble}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(\boldsymbol{x}) \tag{2.7}$$

where each $\hat{f}^{(b)}(\boldsymbol{x})$ is a separate prediction model.

It has shown to be an effective way of constructing high performing prediction models, even though the individual prediction models in the ensemble are not [11].

However, given the same training data, fitting multiple decision trees would give identical trees and an ensemble would give no benefit.

Bagging, a conjunction of the words *bootstrap aggregating*, is used to introduce stochasticity in how the ensemble is created. To understand bagging, we first begin with the bootstrap method.

The bootstrap method is a resampling method where a sample from a dataset is generated by sampling from the empirical distribution on the dataset. Simply, samples are drawn by choosing randomly from the dataset with replacement, where each datapoint have equal probability of being chosen. Since we are sampling with replacement, the bootstrap sample may be of arbitrary size, but it is common for it to be of equal size to the original dataset.

A bootstrap sampled dataset may contain multiple examples of the same datapoint, as each datapoint may be sampled more than once. If we denote the size of the dataset as b, the probability of a datapoint not being included is asymptotically given as:

$$\text{For n independent picks: } P(\text{not selected})^b = (1 - \frac{1}{b})^b \rightarrow_{b\rightarrow\infty} e^{-1} \approx \frac{1}{3}. \tag{2.8}$$

A motivation for bagging in this context of ensembles is a reduction in variance compared to the individual prediction models: Consider a set of B independent prediction models, $Z_1, ..., Z_B$, each with variance σ^2. The mean $\bar{Z}$ has variance $\frac{\sigma^2}{B}$, i.e. the average of a set of independent prediction has lower variance than its individual models [11].

Performing bootstrap sampling B times on the same dataset, produces B bootstrapped datasets. Denoting each prediction model fitted on a bootstrap sample as a function $\hat{f}^{(b)}(\boldsymbol{x})$, bagging will produce a final model:

$$\hat{y} = \hat{f}_{bag}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{(b)}(\boldsymbol{x}) \tag{2.9}$$

Bagging in itself has yielded great results for many predictive models. However, there is one flaw in our argument: As the bootstrapped datasets are derived from the same dataset, the prediction models are not trained on completely independent datasets. Hence, the prediction models themselves will not be independent. Assuming $Cov(Z_i, Z_j) = \rho\sigma^2, \forall i \neq j$, the variance in the mean $\bar{Z}$ is really:

$$Var(Z_1 + ... + Z_B) = \sum_{j=1}^{n}\sum_{i=1}^{n} Cov(Z_i, Z_j) = \frac{\sigma^2}{B} + \rho\sigma^2 - \frac{1}{B}\rho\sigma^2 \rightarrow_{B\rightarrow\infty} = \rho\sigma^2 \tag{2.10}$$

This means that the reduction in variance for the aggregated prediction is dependent on the correlation between the trees.

In random forest models, the prediction models used are decision trees fitted on bootstrapped datasets. To further decorrelate the decision trees, each split is only considered for a subset of the features in the training data. The subset of features to split by is randomly chosen with equal probability for all at each split. The number of features to subsample is a tuningparameter and commonly set to $m = \sqrt{p}$. The number of decision trees B is not a tuningparameter in the typical sense, it just needs to be large enough for the ensemble to converge. This can be seen by monitoring the training performance as more trees are added the ensemble.

2.1.4.1 Regularisation: Cost-complexity pruning

Cost-complexity pruning is a post-hoc regularising method for decision tree models [12]. A motivation for applying the method is that it allows training of decision trees without imposing any size restrictions during modelfitting. Starting with a fitted decision tree of arbitrary size, cost-complexity pruning adds a penalty to the impurity measure used to generate the split. The cost-complexity G_α for a tree T is now calculated as:

$$G_\alpha(T) = G(T) + \alpha|T| \tag{2.11}$$

where $G(T)$ is the sum of impurity for the endnodes of T. See equation 2.6 for the impurity measure for one node. The tuningparameter α is used to determine the strength of pruning. The cost-complexity of a single endnode t is calculated as:

$$G_\alpha(t) = G(t) + \alpha \tag{2.12}$$

Let T_t be the tree using node t as a root node. In general, a node has more impurity than the sum of its terminal nodes. Setting these equal can yield a threshold for α, denoted α_t, where the penalty is high enough that the split at node t does not reduce the cost-complexity:

$$G_{\alpha_t}(t) = G_{\alpha_t}(T) \implies G(t) + \alpha_t = G(T) + \alpha_t|T| \implies \alpha_t = \frac{G(t) - G(T)}{T - 1} \tag{2.13}$$

Recording this value for all internal nodes in a tree, the node with the minimal $\alpha_t < \alpha$ is pruned. The process is then repeated until all nodes have $\alpha_t \geq \alpha$.

An advantage of cost-complexity pruning is that it removes the need for tuningparameters on tree-size. Although decision trees are fitted greadily, a seemingly ineffective split at one node may lead to effective splits at later nodes. Cost-complexity pruning takes this into account by only pruning a node if the average effect of all subsequent splits is lower than α.

2.1.5 Neural networks

This section is based to a large extent on [14] [15] [9].

The neural network model-family is a family of non-parametric models. In essence, neural networks are powerful function approximators used to approximate the function relating the features $\boldsymbol{x}$ to the target y, by compostions of differentiable functions. A relatively standard feed-forward neural network is described in this section.

Some basic terminology is needed to describe a neural network. The matrix $\boldsymbol{W}_{(l)}$ and vectors $\boldsymbol{b}_{(l)}$ are referred to as a weight matrix and intercept vector for layer l, respectively. The vectors $\boldsymbol{a}^{(l)}$ are called intermediate activations, while $f_{(l)}$ are called activation functions.

A standard feedforward neural network with q layers is mathematically described as:

$$\begin{aligned} \boldsymbol{a}^{(2)} &= f_{(1)}(\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{(1)}}^T + \boldsymbol{b}_{(1)}), \qquad l = 1 \\ \boldsymbol{a}^{(l+1)} &= f_{(l)}(\boldsymbol{a}^{(l)}\boldsymbol{W}_{\boldsymbol{(l)}}^T + \boldsymbol{b}_{(l)}), \qquad 1 < l < q \\ \hat{y} &= f_y(\boldsymbol{a}^{(q)}\boldsymbol{W}_{\boldsymbol{(q)}}^T + \boldsymbol{b}_{(q)}), \qquad l = q \end{aligned} \tag{2.14}$$

where the activation functions $f(l)$ are applied elementwise to its input.

For binary classification problems, f_y is often the Sigmoid function from equation 2.2 and Figure 2.2. Of the q layers described in equation 2.14, the layers $1 < l < q$ are referred to as 'hidden layers'. The dimensionality of $\boldsymbol{W}_{(l)}$, $\boldsymbol{b}_{(l)}$ determines the dimensionality of activations, allowing the intermediate activations $\boldsymbol{a}_{(l)}$ to have arbitrary length. The number of hidden layers and the length of $\boldsymbol{a}_{(l)}$ are often referred to as 'depth' and 'width' of a network, respectively. While $\boldsymbol{a}_{\boldsymbol{(l)}}$ can have varying length in each layer l, i.e. a network with varying width, it is not uncommon for all layers to have the same width in a feed-forward neural network.

As long as the activation function is not an algebraic polynomial, a version of the universal approximation theorem states that any continuous function defined on R^p may be approximated arbitrarily well by a neural network with at least one hidden layer, as the width of this layer goes to infinity [16]. Empirical results have shown that it is easier to fit well-performing neural networks with multiple layers rather than only using a single layer, for the same amount of parameters [9]. However, this gives no insights in how find the optimal values for $\boldsymbol{W}_{(1)}, \boldsymbol{b}_{(1)}, \ldots, \boldsymbol{W}_{(q)}, \boldsymbol{b}_{(q)}$.

For practical optimization of neural networks, gradient based optimization is used. Automatic differentation methods are used to compute the gradients for each parameter as a chain of partial derivatives. These are implemented in dedicated software libraries, like PyTorch [17]. This chain of partial derivatives also highlights the need for the neural network to be composed of *differentiable* activation functions, or the gradient will not be available. These parameter-gradients are calculated in order to minimize a loss function, often the binary cross-entropy described in equation 2.3.

There are some similarities between the neural networks and logistic regression model families. If $q = 1$ and $f_y(.)$ is the Sigmoid function as shown in equation 2.2, equation 2.14 describes a logistic regression model. Assuming $q > 1$, two key distinctions between logistic regression and neural networks can be made. Firstly, logistic regression models linear effects of features, and has a convex optimization problem. The solution is hence the best possible for linear effects. Neural networks are capable of modelling non-linear effects of features, but has a non-convex optimization problem. For neural networks an optimal solution is not guaranteed, so it is often sufficient in predictive problems to find a solution that performs well-enough.

Extending on the standard feed-forward neural network, a Bayesian ensembling method called *Multi-SWAG* is also utilized. Multi-SWAG is a Bayesian method for constructing a predictive model, extending on a method named *Stochastic Weight Averaging-Gaussian* (SWAG) [18][19]. SWAG is a method for constructing an ensemble of well-performing neural network models. This is done by approximating a high-likelihood region for parameters of a neural network by a Gaussian distribution. A high-likelihood region is defined to be the parameterspace surrounding the parameters of a converged neural network. Using samples from the approximate distribution as parameters, the method can produce a distinct and well-performing neural network for each sample.

However, it is common for neural networks to have multiple high-likelihood parameter solutions. Rather than fitting a single minima, Multi-SWAG expands on SWAG by applying the procedure on multiple high-likelihood regions, i.e. using multiple converged neural networks. The idea is that in the case of multiple high-likelihood solutions in the parameter-space, the converged networks would randomly distribute among these minimas due to the networks being randomly initialized.

The idea is that given some converged neural networks models, Multi-SWAG offers a computationally cheap way of generating many more distinct, well-performing neural network models. These are used in an ensemble as the final prediction model. In the following parts, the SWAG procedure is detailed.

To simplify notation, $\boldsymbol{W}_{(1)}, \boldsymbol{b}_{(1)}, \ldots, \boldsymbol{W}_{(q)}, \boldsymbol{b}_{(q)}$ are denoted as a joint weight vector $\boldsymbol{w}$.

SWAG is based on approximating fully Bayesian inference on $\boldsymbol{w}$. Consider the probability distribution of targets where the model parameters are marginalized out:

$$P(y|\boldsymbol{x}, \mathcal{D}) = \int P(y|\boldsymbol{x}, \boldsymbol{w}) P(\boldsymbol{w}|\mathcal{D}) \mathrm{d}\boldsymbol{w} = E_{\boldsymbol{w} \sim P(\boldsymbol{w}|\mathcal{D})}(P(y|\boldsymbol{x}, \boldsymbol{w})) \tag{2.15}$$

where $\mathcal{D}$ denotes the data-distribution. The expectation in 2.15 is approximated by Monte Carlo sampling, using R random draws from the posterior of model parameters

$P(\boldsymbol{w}|\mathcal{D})$:

$$\int P(y|\boldsymbol{x},\boldsymbol{w})P(\boldsymbol{w}|\mathcal{D})\mathrm{d}\boldsymbol{w} \approx \frac{1}{R}\sum_{r=1}^{R} P(y|x,\boldsymbol{w}_r), \quad \boldsymbol{w}_r \sim P(\boldsymbol{w}|\mathcal{D}) \tag{2.16}$$

The $P(y|\boldsymbol{x},\boldsymbol{w})$ is modelled by a neural network, meaning that equation 2.16 can be described as an ensemble of multiple neural networks, each with a randomly drawn set of model parameters. In the article introducing the method, they refer to it as a *Bayesian model average*. In the case of 'classical' training of neural networks, the weight-posterior is approximated as:

$$P(\boldsymbol{w}|\mathcal{D}) \approx \begin{cases} 1 & \boldsymbol{w} = \hat{\boldsymbol{w}} \\ 0 & \boldsymbol{w} \neq \hat{\boldsymbol{w}} \end{cases} \tag{2.17}$$

where $\hat{\boldsymbol{w}} = argmax_{\boldsymbol{w}} P(\boldsymbol{w}|\mathcal{D})$.

The SWAG method relies on approximating the posterior $P(\boldsymbol{w}|\mathcal{D})$ by a multivariate Gaussian distribution. The SWA of SWAG refers to the method used to fit the mean and variance. Starting from a converged neural network, training is continued using a relatively large gradient step. The idea is that the weight parameters will take multiple high-likelihood values close to the local high-likelihood solution the model originally converged to. These parameters are sampled at intervals of the gradient steps. After K number of parameter samples are acquired, the posterior mean and variance Σ are approximated as

$$\begin{aligned} \mathrm{E}(\boldsymbol{w}) &\approx \bar{\boldsymbol{w}} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{w}_k \\ \mathrm{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma} &\approx \hat{\boldsymbol{\Sigma}} = \frac{1}{K-1}\sum_{k=1}^{K}(\boldsymbol{w}_k - \bar{\boldsymbol{w}})(\boldsymbol{w}_k - \bar{\boldsymbol{w}})^T = \frac{\boldsymbol{D}\boldsymbol{D}^T}{K-1} \end{aligned} \tag{2.18}$$

where the columns of $\boldsymbol{D}$ is $\boldsymbol{D}_k = (\boldsymbol{w}_k - \bar{\boldsymbol{w}})$. As the number of parameters for a neural network is often quite large, the tuningparameter K is set to a lower value to enable efficient sampling of D. The resulting approximate posterior for model parameters is $\mathcal{N}(\bar{\boldsymbol{w}}, \hat{\boldsymbol{\Sigma}})$.

2.1.5.1 Regularisation: Dropout

Dropout is a regularisation technique specific for neural networks [20]. It is described for a neural network defined as in equation 2.14. On each training iteration, some elements of $\boldsymbol{a}^{(l)}\ \forall l$ are randomly set to zero. The action of zeroing out a single element is modelled as a Bernoulli distributed random variable with probability $1 - \pi_{Drop}$ of keeping the element. After modelfitting, the elements of $\boldsymbol{a}^{(l)}\ \forall l$ are multiplied by the

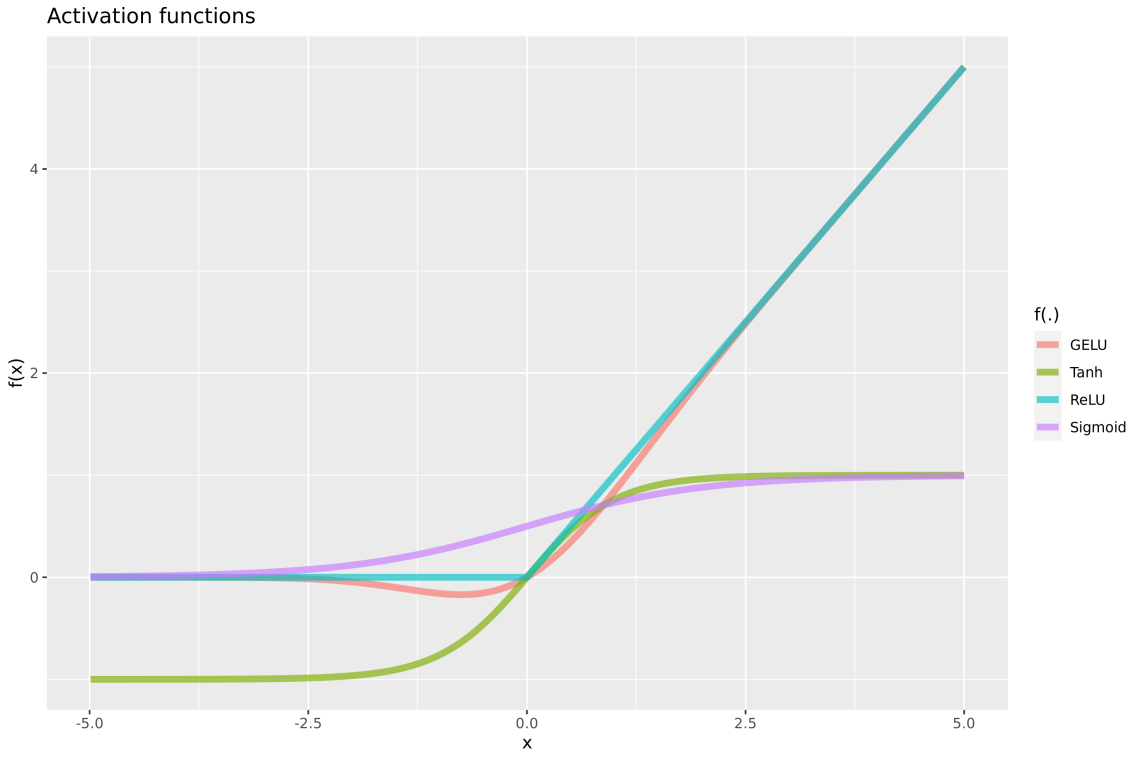


Figure 2.2: Examples of non-linear activation functions that can be used in neural networks.

mean value $1 - \pi_{Drop}$.
By zeroing out elements of $\boldsymbol{a}^{(l)}$, the resulting neural network is a subnetwork of the full network.

Hinton et al. claims Dropout is equivalent to taking the geometric mean of the probability distributions over targets predicted by all possible subnetworks. Assuming all of them do not make the same prediction, this mean is guaranteed to have higher log probability of the correct class compared to any of the subnetworks.[20]

2.1.5.2 Regularisation: Early stopping

Another regularisation method commonly applied to neural networks is early stopping [9]. The method consists simply of witholding some data from the training data and monitoring the performance of the model upon the withheld data during modelfitting. The modelfitting is stopped as the model exhibits worsening or stagnating performance on the withheld data over a predefined number of training iterations.

2.1.6 Class imbalance loss

As detailed in Table 9 in the appendix, it is common to have an imbalance on the ratio of normotensives to hypertensives in datasets used to fit hypertension risk models. This can be referred to as "Class imbalance" and may affect the optimization of some methods. In addition, it is often of importance to ensure good predictive power of the minority class. A common way to ensure that models learn to discriminate different classes sufficiently well is to apply a scaling factor for each class in the loss [21]. Doing so would alter the cross-entropy loss in equation 2.3 to

$$L(\hat{\boldsymbol{y}}) = -\sum_{i=1}^{n} \tau y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \tag{2.19}$$

and the Gini-index criteria for each split in equation 2.6 with $K = 2$ to

$$G = \tau \hat{p}_{m1}(1 - \hat{p}_{m1}) + \hat{p}_{m2}(1 - \hat{p}_{m2}) \tag{2.20}$$

where the tuningparameter $\tau > 0$ is a loss-scaling factor for class 1 relative to class 0. A value of $\tau < 1$ implies that the loss for datapoints of class 0 is minimized more, while $\tau > 1$ implies the same for class 1.

2.2 Performance measures

To evaluate the performance of data models, there exists a multitude of different measures. In this section, the measures used in this thesis are presented. The measures

are divided into two sections, those that are independent of a probability threshold and those that are not.

While some measures detail the discriminatory power of the model, other focus on calibration, or a combination of the two. Discrimination is a models ability to identify datapoints of different targets. Calibration is the agreement between predictions and target values. A common definition of perfect calibration is that *"we should observe p% outcomes among datapoints with a predicted risk of p%"* [22].

The notation as outlined in section 2.1.2 is utilized here.

2.2.1 Performance measures varying with threshold

The following measures varies with a threshold value c such that the predicted class of a datapoint i is $\hat{y}_i^* = I(\hat{y}_i > c) = \begin{cases} 1, & \hat{y}_i > c \\ 0, & \hat{y}_i < c \end{cases}$. The set of indices corresponding to $y_i = 1$ and $y_i = 0$ are denoted N_1 and N_0 respectively. The set of indices corresponding to $\hat{y}_i^* = 1$ and $\hat{y}_i^* = 0$ are denoted N_1^* and N_0^* respectively. Further denote $|N_1|, |N_0|, |N_1^*|, |N_0^*|$ as the number of indices in each set.

Tpr: True positive rate

$$P(\hat{y}^* = 1 | y = 1) \approx \frac{\sum_{i \in N_1} I(\hat{y}_i^* = 1)}{|N_1|} \tag{2.21}$$

Tpr is the probability that the prediction is 1 when the true target is 1. Note: This measure is also referred to as *recall* [23].

Fpr: False positive rate

$$P(\hat{y}^* = 1 | y = 0) \approx \frac{\sum_{i \in N_0} I(\hat{y}_i^* = 1)}{|N_0|} = 1 - \frac{\sum_{i \in N_0} I(\hat{y}_i^* = 0)}{|N_0|} \tag{2.22}$$

Fpr is the probability that the prediction is 1 when the true target is 0. It is equal to 1 - *True negative rate* [23].

Ppv: Positive prediction value

$$P(y = 1 | \hat{y}^* = 1) \approx \frac{\sum_{i \in N_1^*} I(y_i = 1)}{|N_1^*|} \tag{2.23}$$

Ppv is the probability that the target value is 1 when the predicted value is 1. Note: This measure is also referred to as *precision* [23].

Npv: Negative prediction value

$$P(y = 0|\hat{y}^* = 0) \approx \frac{\sum_{i \in N_0^*} I(y_i = 0)}{|N_0^*|} \tag{2.24}$$

Npv is the probability that that the target value is 0 when the predicted value is 0 [23].

Accuracy:

$$P(\hat{y}^* = y) \approx \frac{\sum_{i=1}^{n} I(\hat{y}_i^* = y_i)}{n} \tag{2.25}$$

Accuracy is the probability that the predicted value is equal to the target [23].

2.2.2 Performance measures not requiring a threshold

The measures in this section are invariant changes in threshold values. With the exception of the Hosmer-Lemeshow measure, these measures will be used in evaluating the final models. For comparison, a baseline model is defined. The model predicts the risk for all datapoints as the proportion of hypertensives in the dataset, which is 0.214. This model will be referred to as the "no-skill" model.

2.2.2.1 AUC: Area under the curve

Both of the area under the curve measures utilize some of the threshold-dependent measures. However, both measures integrate over all threshold values $c \in [0, 1]$. Both measures are used to describe a models overall ability to discriminate data observations. Note that neither gives any indication how well calibrated the model probabilities are. E.g. if all observations with target $y = 1$ are predicted $\hat{y} = 0.1$, while those with target $y = 0$ are predicted $\hat{y} = 0.09$, both AUC measures would give a perfect score of 1.

AUC_{ROC}: Area under the Receiver-Operator-Curve AUC_{ROC} is a commonly used measure of the discriminative power of a model with binary target values. It can be seen graphically as the area under the receiver-operator-curve (ROC). A ROC curve can be seen in Figure 2.3, and is the plot of a models true positive rate vs. its false positive rate at varying threshold levels. Mathematically, it can be interpreted as the probability of observations with target value 1 being predicted to a higher value than observations with target value 0:

$$AUC_{ROC} = P(\hat{y}_i > \hat{y}_j)_{i \in N_1, j \in N_0} \approx \sum_{\Delta T} \sum_{\Delta T'} I(T' > T)\, Tpr(T')\, Fpr(T)\, \Delta T' \Delta T$$

A score of 1 is entails perfect discrimination of the classes [24][23]. The no-skill model will have $AUC_{ROC} = 0.5$, although the interpretation fails since all predictions are equal.

AUC_{PR}: Area under the Precision-Recall-curve In a similar manner to AUC_{ROC}, the AUC_{PR} can be seen graphically as the area under the Precision-Recall-Curve (PR). A PR curve can be seen in Figure 2.3. It is a the plot of a models precision vs. its recall, i.e. the positive prediction value vs. the true positive rate, at varying threshold levels:

$$AUC_{PR} \approx \sum_{\Delta T} \sum_{\Delta T'} I(T' > T)\ Tpr(T')\ Ppv(T)\ \Delta T' \Delta T$$

This measure distinguishes itself from AUC_{ROC} by only involving measures obtained from one class. This may be of benefit when there is an imbalance in the number of targets for each class. Suppose there is an imbalance of targets in the dataset. The AUC_{ROC} can then be artifically high due to overpredicting examples of one class. The negative effect of overpredicting the minority class for the Fpr is mitigated by the higher number of examples of the majority class. There is a decreasing penalty to the AUC_{ROC} measure for overpredicting the minority class as the imbalance increases [23][25].

Class imbalance, as seen in Table 9 in the appendix, is a common attribute of datasets used for making hypertension risk models.

A score of 1 would mean perfect discrimination on the class it is measured on or perfect for both classes in binary classification. The AUC_{PR} for a class would be equal to the proportion of the dataset with that class as target in the case of a no-skill model. For the no-skill model, this is coincidentally 0.214 for the hypertensive class.

The Hosmer Lemeshow (HL) statistic: The HL statistic is commonly used in many articles as an measure of the calibration of the model. It orders and stratifies by value a set of predictions into G groups, and calculates

$$HL = \sum_{g=1}^{G} \Big(\frac{(O_{1,g} - E_{1,g})^2}{E_{1,g}} + \frac{(O_{0,g} - E_{0,g})^2}{E_{0,g}} \Big) \tag{2.26}$$

where $E_{1,g}, O_{1,g}$ is expected and observed number of datapoints with target 1, $E_{0,g}, O_{0,g}$ are expected and observed number of datapoints with target 0, for group g. This quantity follows asymptotically a χ^2_{G-2} distribution with increasing number of datapoints in each group [26].

Although no longer recommended to use for its intended purpose, it is nevertheless reported for numerous models for hypertension modelling, and therefore included here [27].

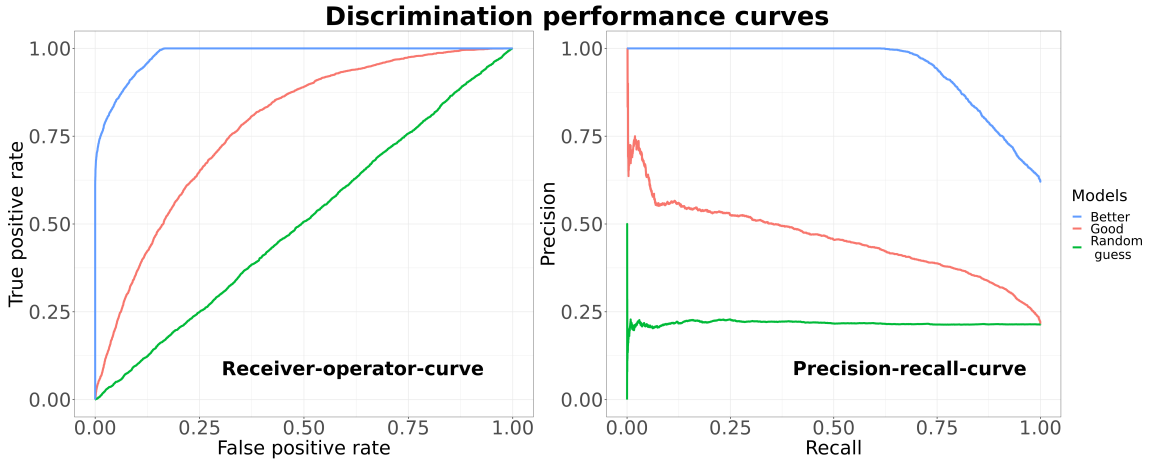


Figure 2.3: Receiver-operator and precision-recall curves for a random guess model, a good model and a better model. The random guess predicts random values in $[0, 1]$ for all datapoints. Each point on the curves correspond to a pair of measures obtained using a fixed thresholdvalue c. The curves are constructed by calculating measures as c takes values in $[0, 1]$.

The Brier Score: The Brier score is a performance measure calculated directly on the predictions. For this thesis an altered version is used [28]. It is defined as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2.27}$$

A Brier score of 1 indicates that the predictions $\hat{y}_i = y_i \forall i$, i.e. the model produces perfect predictions on all datapoints. Note that this measure gives an indication of the produced probabilities as well as discrimination, as the predicted value is used directly without thresholding. The "no-skill" model will have an Brier score of 0.1684.

Tjur's R^2: Tjur's R^2 is a coefficient of determination included for its simplicity and applicability across methods. Denoting the set of indices corresponding to $y_i = 1$ and $y_i = 0$ as N_1 and N_0 respectively, with $|N_1|, |N_0|$ their respective sizes, it is defined as

$$R^2_{Tjur} = \frac{\sum_{i \in N_1} \hat{y}_i}{|N_1|} - \frac{\sum_{j \in N_0} \hat{y}_j}{|N_0|} \tag{2.28}$$

In the context of this thesis, Tjur's R^2 is the difference between the average prediction of hypertensives and the average prediction for normotensives. It can be interpreted as a measure of the models overall confidence, as high confidence would correspond to hypertensives being predicted to higher probabilities than normotensives [29]. For the no-skill model, this measure is 0.

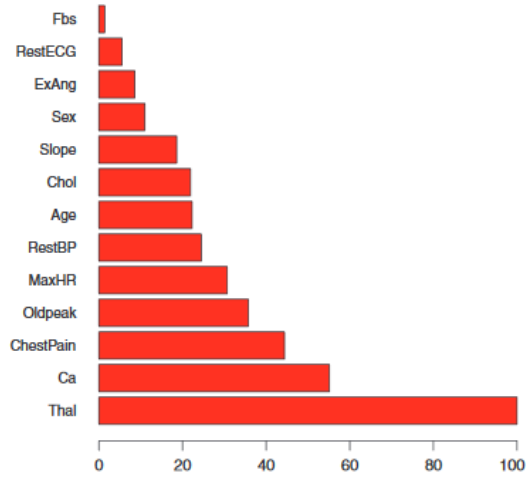


Figure 2.4: Example of variable importance plot for some features. Values are percentage relative to the highest importance score. Adapted from [11].

2.3 Feature importance measures

This section lists methods that have been utilized to score how important each feature is for the predictions of a model.

2.3.1 Variable importance

A measure of feature importance specific for random forest models is variable importance, proposed in [12]. The importance measure is calculated as the sum of impurity reduction for nodes using that feature to split by, divided by the number of trees. With B trees in a random forest model, denote s_t as the impurity decrease for a split in node t, $p(t)$ the proportion of training data that reached node t and $v(s_t)$ as the feature used to split at node t. The importance of a feature j, VI_j is then calculated as:

$$VI_j = \frac{1}{B} \sum_{b=1}^{B} \sum_{t \in f_j^{(b)}} p(t) s_t \tag{2.29}$$

where $f_j^{(b)}$ denotes the set of nodes in tree $f^{(b)}$ that splits using feature j. An example of features ranked by their variable importance scores, adapted from [11], can be seen in Figure 2.4.

2.3.2 Permutation importance

As a post-hoc method of assessing how vital features are for the predictive performance, simple permutation importance is measured [13]. This method may be applied to any prediction model, regardless of model-family. The permutation importance of a feature is measured as the change in a performance score on a dataset when that feature is randomly permuted among the datapoints in the dataset.

Assume higher performance score is better. Denoting the full dataset as in section 2.1.2, each column in $\boldsymbol{X}$ encodes the values of a single feature while each each row is a datapoint. Denote $\boldsymbol{X}_{*j}^{(t)}$ as $\boldsymbol{X}$ with column j randomly reordered for iteration t. Denote permutation importance for feature j as PI_j, the performance measure as a function $M : R^n \rightarrow R$ mapping predictions to performance scores. Then PI_j for a prediction model $f(.)$ is calculated as:

$$PI_j = M(\hat{\boldsymbol{y}}) - \frac{\sum_{t=1}^{T} M(\hat{\boldsymbol{y}}_{*j}^{(t)})}{T}, \qquad \hat{\boldsymbol{y}}_{*j}^{(t)} = f(\boldsymbol{X}_{*j}^{(t)}), \quad \hat{\boldsymbol{y}} = f(\boldsymbol{X}) \tag{2.30}$$

where T is the number of permutations used to control for the stochasticity introduced by the permutations. Further, $f(\boldsymbol{X})$ gives a vector of predictions by row-wise application of $f(.)$ on $\boldsymbol{X}$.

2.3.3 Logistic regression with Lasso loss

Lasso loss, as described in section 2.1.3.1, have a sparsifying effect in logistic regression models. Suppose that we are fitting logistic regression models with Lasso loss, increasing the penalty tuningparameter incrementally. The sparsifying effect means that coefficients that are less useful in minimizing the loss function are zeroed out earlier than those that are useful. As an importance measure for features, the penalty at which the coefficients are zeroed out can be recorded as the score for their associated features. The higher the importance of a feature, the higher the penalty needed to be before its coefficient was zeroed out.

2.4 Methods for choosing tuningparameters

2.4.1 K-fold cross-validation

K-folds cross-validation is a method for obtaining an estimate of the generalization performance a model. The method is based on distributing the training data into K equally-sized subsets, called folds. Iteratively, for $i \in \{1, \cdots, K\}$, a model is fitted using folds $K_{-i} = [1, 2, \cdots, i-1, i+1, \cdots, K]$ and calculating performance scores upon the i'th fold. This will generate K performance scores than can be summarize to approximate the models true performance. As models are fitted with partially overlapping

folds, there will be some correlation between measures. The tuningparameter K can be used to control for this. A low number of K generates high-bias, low-variance estimates, while a high K gives low-bias, high variance estimates [11]. A limitation is that K-fold cross-validation requires fitting a model K times, which is prohibitive for large K or expensive function evaluations.

2.4.2 Bayesian search using Gaussian processes

An alternative for tuningparameter search is to use a Bayesian search strategy [30].

In the Bayesian tuningparameter search presented, the number of possibly costly model fittings can be controlled for. At the same time, it may provide better tuningparameters compared to a random search.

At its core, the performance scores are modelled as realizations of a Gaussian process on the tuningparameters. A Gaussian process is completely specified by its mean function and its covariance function, called the *kernel* function. For simplicity, the mean function is set to 0, the performance values are assumed noise-free and the number of tuningparameters the process is modelled on is 1. Let λ denote a tuningparameter, with $M(\lambda)$ denoting the performance score of a prediction model fitted using the tuningparameter λ.
Suppose we have t number of tuningparameters $\boldsymbol{\lambda}_{1:t} = \begin{bmatrix}\lambda_1 & \dots & \lambda_t\end{bmatrix}$ along with their known performance scores $M(\boldsymbol{\lambda}_{1:t}) = \begin{bmatrix}M(\lambda_1) & \dots & M(\lambda_t)\end{bmatrix}$. The kernel function, $k(\lambda_i, \lambda_j)$ details the influence between different tuningparameter settings λ_i and λ_j. The joint distribution of an arbitrary point $M(\lambda_{t+1})$ on the process and $M(\boldsymbol{\lambda}_{i:t})$ is Gaussian:

$$\begin{bmatrix}M(\boldsymbol{\lambda}_{1:t})^T \\ M(\lambda_{t+1})\end{bmatrix} \sim \mathcal{N}\Big(0, \begin{bmatrix}\boldsymbol{K} & k(\boldsymbol{\lambda}_{1:t}, \lambda_{t+1}) \\ k(\boldsymbol{\lambda}_{1:t}, \lambda_{t+1})^T & k(\lambda_{t+1}, \lambda_{t+1})\end{bmatrix}\Big) \tag{2.31}$$

With

$$\boldsymbol{K} = \begin{bmatrix}k(\lambda_1, \lambda_1) & \dots & k(\lambda_1, \lambda_{t+1}) \\ \vdots & \ddots & \vdots \\ k(\lambda_{t+1}, \lambda_1) & \cdots & k(\lambda_{t+1}, \lambda_{t+1})\end{bmatrix} \tag{2.32}$$

and

$$k(\boldsymbol{\lambda}_{1:t}, \lambda_{t+1})^T = \begin{bmatrix}k(\lambda_1, \lambda_{t+1}) & k(\lambda_2, \lambda_{t+1}) & \dots & k(\lambda_t, \lambda_{t+1}))\end{bmatrix} \tag{2.33}$$

An important result for Gaussian processes, is that the posterior distribution of any point on the process is Gaussian, with mean and variance available in an analytical form. The posterior distribution of $P(M(\lambda_{t+1})|M(\boldsymbol{\lambda}_{i:t}), \boldsymbol{\lambda}_{i:t}, \lambda_{t+1})$ is Gaussian with mean μ_{t+1} and variance σ^2_{t+1}:

$$\begin{aligned}\mu(\lambda_{t+1}) &= k(\boldsymbol{\lambda}_{1:t}, \lambda_{t+1})^T \boldsymbol{K}^{-1} M(\boldsymbol{\lambda}_{1:t}) \\ \sigma^2(\lambda_{t+1}) &= k(\lambda_{t+1}, \lambda_{t+1}) - k(\boldsymbol{\lambda}_{1:t}, \lambda_{t+1})^T \boldsymbol{K}^{-1} k(\boldsymbol{\lambda}_{1:t}, \lambda_{t+1})\end{aligned} \tag{2.34}$$

Using these equations, a search can be done for the next candidate tuningparameter that is optimal with respect to an *acquisition function*. There are numerous suitable options to use as acquisition functions. The acquisition function $acq(\lambda_{t+1})$ used in this thesis is the *confidence bound*, defined as:

$$acq(\lambda_{t+1}) = \mu(\lambda_{t+1}) + \kappa\sigma^2(\lambda_{t+1}) \tag{2.35}$$

with $\kappa > 0$ as a separate tuningparameter. A large value of κ will lead the search towards points that have a large predicted uncertainty, i.e. exploration, and a low value towards the areas that displayed a good performance, i.e. exploitation. Having found a suitable λ_{t+1}, the model is be fitted using it as the tuningparameter. This process is repeated for an arbitrary choice of iterations.

Chapter 3

Literature review

3.1 A briew overview on blood pressure and hypertension

The World Health organization (WHO) defines hypertension as a condition of lasting, elevated blood pressure (BP) during a resting state of an individual [4]. BP is defined as the force of circulating blood on the walls of the arteries. The blood pressure naturally varies with heart beats, crudely described as increasing rapidly after the heart beats, and slowly sinking until the next heartbeat.

Due to this natural variability, blood pressure is characterized by two values, taken during the same reading: The *systolic* and *diastolic* BP. Systolic BP is the BP at its maximum, i.e. right after the heart beats, whilst diastolic BP is the BP at its minimum, i.e. between heart beats. A BP measurement is given in units of *millimeters of mercury* (mmHg) and commonly written on the form 'sys. BP/dia. BP', e.g. 120/75.

There is some discrepancy in at what levels hypertension is diagnosed for otherwise healthy adults. There is consensus among major guidelines that having systolic BP at more than 140 mmHg or diastolic BP at or above 90 mmHg defines having hypertension. As for differences, the Eight Join National Commitee (JNC 8) defines hypertension as starting at a sys. BP. levels above 130 mmHg or dia. BP. levels above 80 mmHg. This difference translate into some differences in how treatment is prescribed, mainly through the use of non-pharmacological therapy in this interval [31] [32] [33]. In any case, one should note that the threshold of when one is diagnosed with hypertension is arbitrary. The distinction is useful however, as a tool for patient assessment and treatment.

There are different types of hypertension that a person can be diagnosed by. A clear majority (95%) of hypertension cases are of the type called *essential*. It is also called *primary* or *idiopathic* hypertension. Essential hypertension is defined as suffering from hypertension when secondary causes are not present. There are further

subdivisions of hypertension, depending on the cause and exact BP values [6].

Hypertension is often characterized as a largely symptomless disease, a "silent killer". Except for the case of extreme BP levels (>180/120), it is not common to experience symptoms of hypertension [33]. Despite this, being inflicted by hypertension over time may lead to numerous negative health consequences, including premature death. It is estimated that more than 1.1 billion persons are afflicted with hypertension, with high economic costs and loss of life as expected consequences [2] [4] [3] [5].

Blood pressure levels are known to naturally change during a persons lifetime. As a person ages, both the systolic and diastolic BP levels increase up until roughly 50 years of age. After that, systolic BP continue to rise, while the diastolic BP stagnates or lowers slightly [34]. Although the exact mechanics behind hypertension is not perfectly understood, it has been discovered that genetic variations, diet and other factors like stress, aging and obesity are influentual in causing high BP levels. An example of how genetics can be seen to affect BP levels is studies indicating that hypertension risk is associated with having hypertension in near family. However, it is quite difficult to assess the exact genetic causes, as the two phenotypes that determine BP levels, cardiac output and total periphertal resistance, are further controlled by numerous intermediary phenotypes [6].

Despite not knowing the exact cause, hypertension is generally perceived as a treatable disease, i.e. that it is possible to lower BP levels to a healthier level through treatment. Recommended treatments include medication, lifestyle changes and other nonpharmacological therapy. The different treatments are recommended relative to the BP levels measured. As an example, treatment with medication for otherwise healthy individuals is recommended if their BP levels surpass 140/90. Below this threshold, the European Society of Cardiology/European Society of Hypertension (ESC/ESH) recommends lifestyle changes to address elevated BP levels, whereas the JNC guideline suggests nonpharamacological treatment in addition to lifestyle changes when the levels are above 130/80 [33]. Despite the rigorness of guidelines, extensive research conducted, the economic cost and awareness on hypertension, the bottom line is that 2 out of 3 still fail to lower their BP levels to acceptable levels when diagnosed with pre-hypertension (>130/80) or hypertension (>140/90) [35].

3.2 Literature review on predictive models for hypertension

The literature on hypertension risk modelling is rich and diverse as there are good reasons for focusing research efforts onto hypertension risk modelling. Among the motivation stated in literature, one can find articles seeking to investigate associations of hypertension and features, constructing models for usage in a clinical setting

or constructing simplified risk models for areas where medical resources are scarce. For this thesis, a collection of articles have been included to assess the current state of research on hypertension risk modelling. The articles selected are a subset of the available literature on the topic, to ensure relevance of the included literature. The chosen articles were filtered based on what method they used, the purpose and the reporting standards of the study. One exception was to include the Framingham model [8] and articles validating it. This was motivated by the well-known standing the Framingham model has in the literature, and is used as a benchmark.
Mainly, articles were selected if the following search criterias all applied:

Purpose: A hypertension risk model using a binary outcome of hypertension status as endpoint was constructed,

Application: The models were applied or validated on a cohort and at least 1 performance measure was reported,

Method: At least one of methods used were either from the logistic regression family, decision-tree based or neural network model families **OR** the article validated or mimicked the Framingham model.

If an article validates another model from literature it is meant that the performance of the model to be validated is reported for a new, unseen dataset. Note that this is done without any modellfitting. If a model is said to 'mimick' a model, it is meant that the model-setup and features are identical to the mimicked model. In this case, the model parameters are fitted to the data available in that study.

A total of 31 articles were found to fulfill the search criterias. Details on the studies and their reported models in tabular form can be found in the appendix. Table 9 details the main properties of the study and cohort, Table 10 details the models reported performance whilst Table 11 details input features used in each paper.

Relevant review articles are those by Echouffo-Tcheugui et al. [36] and Sun et al. [37] for statistical methods, while Krittanawong et al. [38] details the usage of some machine learning methods on the topic.

3.2.1 The main findings

The construction of a predictive model for hypertension risk assessment has been performed in a large number of scientific articles. Reviewing the literature on the topic, there is a high level of heterogeneity in terms of data sources, methodology and more. However, one should not highlight one particular study setup for constructing a risk model as the correct one. This is arguably a function of study purpose, available data and other, possibly uncontrollable, factors. In the case of medical data such a factor could be the high standards and costs of careful and ethically proper handling of data.

The following paragraphs are based upon the findings detailed in Table 9 in the appendix.

Country Accounting only for the country datapoints were collected from, multiple countries were only represented once. This is a common point emphasized in many articles, as the generalizability of the results may be affected. A model derived from one study population may yield different performance on other populations or even other etnicities [39] [40] [41].

This fact encourages the construction of a risk model when data from new populations become available. It is possibly more important to validate existing models upon different populations to evaluate their usability across populations. Even still, in the included literature there is only one risk model, the Framingham risk model, that have been externally validated or mimicked on more than two populations.

Age ranges Blood pressure is known to increase with age and is therefore important for hypertension risk modelling [34]. Details on the age distribution at the time of measurement is included, with a subdivision by endpoint status if reported. The mean or median age in most articles seems to be around 50, with a relatively high standard deviation, indicating an elevated age of participants. Two of the articles reporting a mean age below 30 used the same dataset. Perhaps more interesting is the fact that many articles have chosen to limit their study population by only including participants above a certain age. This may contribute to skew the results, possibly by slighly obscuring the fact that normotensives tend to be younger.

Exclusion criterias Many articles failed to report any exclusion criterias at all. Among the reported, current or a history of cardio-vascular disease, pregnancy and diabetes were among the most reported. All hypertensives were excluded from the initial population in the prospective studies.

Number of participants The number of participants included in the different studies varied greatly, from being in the hundreds up to the millions.

Most of the studies utilized medical cohort data that were collected for the explicit purpose of general medical study, like the HUNT data that is analyzed in this thesis. The largest studies, those with more than 25 000 participants, used either electronic health records (EHR) or in one case data collected by telephone interview. The difference between these types of data collection methods are noteworthy and one should not compare without this in mind. The medical cohort data used are more specific and complete, i.e. fewer features and less missing data, compared to EHR data used.

Hypertension definition Most definitions were consistent with the ESC/ESH guidelines [33], i.e. hypertension was defined as having resting systolic BP above 140 mmHg or resting diastolic BP above 90 mmHg. Some used usage of BP medication or having been diagnosed as hypertensive at any point in time, in addition. It was not clearly stated in most cases if the BP measurements was taken at two different points in time. The exceptions used either: other level thresholds, did not report their definition, or used "yes/no" to an interview question.

Study type The articles used primarily cross-sectional and prospective study types. For this thesis, cross-sectional studies are defined to be studies that analyze data collected at a specific point in time, i.e. with no time difference between measurements and the hypertension status being recorded. Prospective studies are studies that have a time difference between the measurements being taken, and the outcome recorded. Studies in the included articles have a roughly even distribution between the two types.

The mean or median follow-up times for prospective studies have a range of 1 to 23 years. Most articles were on the lower end and only 2 articles are based on studies with follow-up time longer than 10 years.

Methods The most popular method in the included literature was logistic regression, used in $\frac{16}{31}$ articles. This was followed by different decision trees based methods and neural networks, with $\frac{10}{31}$ and $\frac{9}{31}$ articles using these respectively.

Articles using logistic regression or survival analysis methods like Weibull regression were more consistent in methodology and had more similar approaches compared to those utilizing neural networks or tree-based methods. This seemed to be a result of a more rigorous treatment of input features and wanting to explore input-output associations rather than just predict hypertension risk. Using logistic or Weibull regression models with linear feature effects are an established and accepted way of accomplishing this.

Models from the random forest and neural network model families are less established within the medical literature. Articles using these models were more focused on predictive performance, i.e. investigating the feasibility of using the models for this purpose.

An example of this variation is found in how models are reported per article. If an article reported performance measures for multiple logistic regression models, the intention in all cases was to display the effect of adding or removing features to the model. If an article reported performance measures for multiple neural network or random forest models, the intention was to display the effect of varying the tuningparameters in the model.

The following paragraphs are based upon the findings detailed in Table 10 in the appendix.

Validation 'Validation' details the method the authors used to validate the models they constructed in their articles, if any method was used.

For a majority of articles, some form of validation procedure was used. The most common method, used in $\frac{16}{31}$ articles, is a single split of the dataset into a training set and a test set. In some cases a validation set used to choose tuning parameters was used. Another relatively common procdure used was K-fold cross-validation as described in section 2.4.1, found in $\frac{6}{31}$ articles. Two articles used repeated splitting and one reported using bootstrap simulations. A minority of $\frac{8}{31}$ articles reported no validation procedure.

A consequence of not using a validation procedure or using only a single split of the dataset is that the performance scores may be biased compared to the true generalization performance.

Number of features and performance scores In the cases where multiple models with negligible differences in model setup were reported, the best model was included in this review. The best model was either specified in the article or determined by the reported performance scores.

Across all articles, the number of features used is fairly low, apart from a few specific articles. The only models that have more than 20 features used EHR data or involved a large amount of genetic information.

The performance measures are described in section 2.2. As there is a large variation in how performance was reported, it is difficult to properly decide if a model performed better than another. This is not accounting for the variations in data, dataset size and the model input. These can all influence the performance scores and making a proper comparison even more difficult.

Perhaps the most relevant measure for comparison is AUC_{ROC}, as it is the most reported measure and do not rely on a threshold value c. In a majority of articles that reported measures using a threshold, the threshold value was not reported.

The AUC_{ROC} ranges from almost barely better than random, 0.537, to almost perfect discrimination, 0.93. However, a clear majority of AUC_{ROC} scores are found in, and close to, the range $[0.7, 0.85]$. This seems to be the expected score range for a risk model. Zheng [42] reports that an AUC_{ROC} score above 0.8 should be considered as "very good". Wang et al. [43] states "Senior physicians suggest that 30.0% is an acceptable error rate for the diagnosis of hypertension", which is stated from an earlier source.

The following paragraphs are based upon the findings detailed in Table 11 in the

appendix.

Feature selection For selecting the features to use in their risk models, most articles relied on available literature to do so. In Table 11, the column "Feature selection method" detail any mathematical method that have been used to select the features of their final model.

Among parametric methods the usage of significance level to select features was common, found in $\frac{11}{31}$. Common p-values thresholds were 0.05 and 0.1. A few specified explicitly that they utilized a stepwise-procedure, adding features sequentially by the feature that improved the performance the most or removing from the full set those that decreased it the least. The best model was then selected by likelihood or measures like *Bayes Information Criterion* (BIC). In many of the articles, simplified models with few features were also reported. In many cases this was motivated by medical customs, wanting easier models to apply in a clinical setting or for displaying a predictive effect of a small subset of features.

Other methods included correlation filtering, information measures, variable importance as described in section 2.3.1, or simply mimicking the Framingham model by using the same features it does.

Features used in literature Reviewing the features included in the different models that were presented in the literature, there are many that are used frequently. Some common features were: Age, BMI, systolic and diastolic blood pressure, sex, cholesterol levels, smoking and some detail on family history of hypertension. These features were seen just as often in articles that used feature selection methods as in those only relying on literature.

Four of the included articles used genetic information as input to their models. Of these, three of them, Fava et al., Lu et al. and Niiranen et al., fitted models that only differed by inclusion of carefully selected genetic information [44][45][46]. Niiranen et al. fitted models for both cross-sectional study data and prospective study data. In all models, the number of genes was relatively low, from 19 to 38 genes. Based on the AUC performance measures, the genetic information made little difference in performance for all models reported. In contrast with the other three, Alzubi et al. constructed a model with a large amount of genetic information, using 417 523 genes as inputs [47]. However, they did not report any model without genetic information and did not report an AUC_{ROC} score, making it hard to compare to the other models. Other measures suggested that the model of Alzubi et al. performed well.

There are numerous articles that presented models with one or a few features, that still achieved fairly high performance measures.
Paynter et al. saw little improvement in predictive performance after adding systolic, diastolic BP, age, BMI and race as input features.

Muntner et al. proposed two models using only categorized version of systolic and diastolic blood pressure respectively [39]. The one using systolic blood pressure achieved an AUC_{ROC} of 0.768, while a model using the same features as the Framingham model achieved only an increase in AUC_{ROC} to 0.788.
Chien et al. presented a model with age, sex, BMI, systolic and diastolic blood pressure that achieved an AUC_{ROC} of 0.74 [48]. Expanding this model with biochemical markers decreased (!) the predictive performance slightly.
Lim et al. achieved an AUC_{ROC} of 0.707 with only the single input of prehypertension status [49]. Validating the Framingham model gave a performance AUC_{ROC} of 0.791, suggesting that prehypertension alone is not sufficient.
Carson et al. presented two models derived only on women that only used prehypertension status in one and an age times diastolic blood pressure in the other [50]. The one using prehypertension achieved an AUC_{ROC} of 0.71, while the one using only the interaction term achieved a score of 0.81. They compared it to a mimick model of the Framingham model, which achieved an AUC_{ROC} of 0.84. This might suggest that age and diastolic blood pressure are high-value features for predictive performance.
Sathish et al. presented a model intended for being easy to use in areas with limited medical resources [51]. In this model, age above 35, smoking, prehypertension and a measure of obesity is the only features. This model achieved an AUC_{ROC} of 0.802. The other measures suggests the model is much better at identifying normotensives than hypertensives.
Lu et al. presented multiple models, sequentially increaseing the number of included features[45]. While there is little improvement with adding the genetic risk score, the addition of smoking, drinking, pulse and education increased the discriminatory power somewhat. The addition of systolic and diastolic blood pressure vastly improved the AUC_{ROC} from 0.687 to 0.777.
Fitriyani et al. presented a model with only age and five markers describing the physical proportions of ones body [52]. This model achieved an AUC_{ROC} score of 0.87 for the male population, and 0.76 for the female population. However, the dataset used only contained 325 datapoints.

Only a few articles included any non-linear transformations of features as inputs to the model, barring indicator-functions. Those seen in the included literature are 'Age $\times$ dia. BP', 'Age^2', 'Sex $\times$ Age', and 'Age $\times$ Waist-circumference'.

In summary, there do exists some features that are used frequently when constructing hypertension prediction models. These are: Age, BMI, systolic and diastolic blood pressure, sex, cholesterol levels, smoking, some detail on family history of hypertension. Further, the model results seems to indicate that genetic information contains little discriminative power for prediting hypertension risk if only a relatively low number of genes are included. As a contrast, smaller models seems to be able to perform surprisingly well, with multiple examples found in the literature.

Chapter 4

Data and exploratory analysis

4.1 Available dataset

The dataset made available for this thesis originally collected as part of the HUNT Study. The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology NTNU), Nord-Trøndelag County Council, Central Norway Regional Health Authority, and the Norwegian Institute of Public Health.

Not all of the data made available was relevant for the analysis. As a part of the thesis aim, models are constructed for the intention of predicting hypertension incident risk between HUNT2 and HUNT3 for individuals that were healthy at HUNT2. The dataset made available for this analysis included data on participants that participated in studies HUNT1, HUNT2 and HUNT3. To select the largest possible relevant subset of this data, some inclusion criterias were defined. These criterias were constructed with the help of Emma Ingeström, a PhD student in the MyMDT-project, and by reviewing relevant literature on the topic. Participants that fulfilled all criterias were included in the subset used in the analysis, while those who did not are excluded. The criterias for including a single participant were:

- the participant had to be present at both HUNT2 and HUNT3.
- the participant does not have any missing values in blood pressure measurements at HUNT2 and HUNT3
- the participant does not have any missing info on current usage of blood pressure medication at HUNT 2 and HUNT3
- the participant does not have any missing info on diabetes, history of cardiovascular disease at HUNT2

- the participant was healthy at the time of HUNT2, i.e. normotensive, without diabetes and has no history of cardiovascular disease

- the participant has no missing values in the selected features

Hypertension was defined as having systolic BP above 140 mmHg or diastolic BP above 90 mmHg. In the case that an participant used blood pressure medication, 15 mmHg was added to the systolic BP measurement and 10 mmHg to the diastolic BP measurement, as done and suggested in literature [46]. Applying these criterias to the full dataset made available for this thesis, 18249 participants were found to satisfy them.

The selection of features was guided by relevant literature on the topic, see section 3.2, and the advice of Emma Ingeström, a PhD student working with the same dataset in her research. In addition, there was a desire for obtaining as many possibly useful features as possible to increase the probability that any non-linear pattern the models can learn was present in the dataset. These two considerations were weighted against the number of missing values each feature had, as participants with any missing features were not included in the dataset.

Details on the resulting choices of features and target are listed in the appendix, see Tables1, 3 and 2. More detailed information on the variables available in the full dataset is available at the HUNT databank[1]. A notable exclusion from the selected features was information on alcohol consumption, which was dropped due to a high missing rate among participants.

4.2 Data exploration

4.2.1 Feature and target summary statistics

Summary statistics on numerical properties of the 21 features and target split by categorical and continuous types can be found in Tables 4.1 and 4.2.

4.2.2 Variable associations

To assess any possible associations between features and target in the dataset, a correlelogram has been calculated and is presented in Figure 1 in the appendix. For categorical features that had more levels than 2 and were not ordered, correlation was calculated for each level separately.
The correllelogram indicate that some features exihibit high levels of correlation. This is especially true for the physical trait measures, like 'Weight', 'Height', 'BMI' and 'Waist-circumference'. 'BMI' has no empirical correlation with 'Height' in this dataset,

[1]url: https://hunt-db.medisin.ntnu.no/hunt-db/#/

Table 4.1: Summary stats. on categorical variables in the final dataset. Some levels are encoded in plot margins, with the encoding given here.

Variable	Ordered levels	(Encoding) Level: Counts	
Sex	No	Male: 7096	Female: 11153
GFREstStag	No	Stage 1: 6197	Stage 2: 12052
CarInfFam1	No	Yes: 6933	No: 11316
FamHypEv	No	Yes: 6628	No: 11621
PAIlevel	Yes	Low: 6231 High: 7203	Medium: 4815
SmoStat	No	(1) Never: 8382 (3) Daily: 5055	(2) Form. daily: 4812
InvMuniciGeo	No	(1) Inland: 2328 (3) Coastal: 3012	(2) Fjord: 12909
LoveStat	No	(1) Partner: 12034 (3) Separated: 1309	(2) No partner: 4473 (4) Widow(er): 433
Educ	No	(1) Sec. school: 7139 (3) High-school: 2045 (5) Higher, > 4 y: 1964	(2) Upper sec. school: 4018 (4) Higher, < 4 y: 3083
Target value: Hyp.HUNT3	No	Yes: 3905	No: 14344

Table 4.2: Summary stats. on continuous variables in the final dataset.

Variable	Mean	Std. dev.	Min.	25th pctl.	50th pctl.	75th pctl.	Max.
Features:							
Hei	171.00	8.78	136.00	164.00	170.00	177.00	206.00
Wei	74.20	13.02	35.00	64.50	73.00	82.50	150.00
Bmi	25.40	3.63	15.00	22.90	25.00	27.30	52.80
WaistCirc	82.60	10.73	54.00	74.00	82.00	90.00	160.00
BPSystMn23	123.00	10.16	81.00	116.00	124.00	132.00	140.00
BPDiasMn23	74.20	7.88	39.00	69.00	75.00	80.00	90.00
SeCreaCorr	67.1	12.86	8.00	58.00	66.00	75.00	193.00
PartAg	42.81	11.70	19.20	34.10	42.40	50.50	86.30
SeChol	5.57	1.13	1.90	4.80	5.50	6.30	13.30
SeHDLChol	1.42	0.38	0.50	1.10	1.40	1.60	4.10
SeTrig	1.50	0.95	0.21	0.89	1.26	1.82	17.25
SeGluNonFast	5.11	0.84	2.00	4.60	5.00	5.50	10.90

due to being adjusted for this during datacollection. The high correlation values of the physical traits might suggest that only one or two of these features should be included in models, as they might contain redundant information. Other features that have high correlation is the blood pressure measurements. This is somewhat expected, as they describe closely related properties of blood-circulation.
It seems like there is some association in the dataset across variable types. 'Sex' has a high degree of association with several variables, suggesting some redundancy in the dataset. 'Age' has some association with different features. For the endpoint 'Hyp.HUNT3', there is some association with the blood pressure measurements and 'Age', with a small association with 'BMI', 'Waist-Circumference' and 'Serum Cholesterol'.

A boxplot detailing the distributions of continuous variables, split by the endpoint hypertension status, is presented in Figure 4.1. All continuous features exhibit little difference between their splitted distributions. By visual inspection of the boxplots, it seems as there is a small, but equal difference in median for 'Weight', 'BMI', 'Waist-Circumference'. The equal difference is no surprise considering that they exhibit high correlation. Another variable that had some difference was 'Serum Cholesterol'. For three variables, the median for hypertensives is higher than the 75th percentile for normotensives. This is the case for 'Systolic BP.', 'Diastolic BP' and 'Age'. This may suggest that these are the most effective single covariates to have in a model. Apart from these, most variables display little difference in their splitted distributions.

Finally, to further investigate associations between discrete variables, an information theoretic measure called Theil's uncertainty coefficient has been calculated, shown in Figure 2 in the appendix. Theil's coefficient for a random variable X given the random variable Y, $U(X|Y)$, is defined:

$$U(X|Y) = \frac{H(X) - H(X|Y)}{H(X)} = \frac{H(Y) - H(Y|X)}{H(X)} \tag{4.1}$$

where $H(X) = E_X(log(P_X)), H(X|Y) = E_{X,Y}(log(P_{X|Y}))$ is the entropy of X and conditional entropy of $X|Y$ respectively. Theil's coefficient takes values in $[0, 1]$, where 1 indicate a perfect association and 0 indicate no association. Due to the denominator, it is an asymmetric measure, loosely interpreted as estimating the percentage of information in X that may be explained by Y. Reviewing the figure, the coefficient is fairly low for all variables. The endpoint 'Hyp.HUNT3' was best predicted by 'Love Status'. However, the measure values are quite small and hence indicate that none of the categorical variables have a strong association with the target.

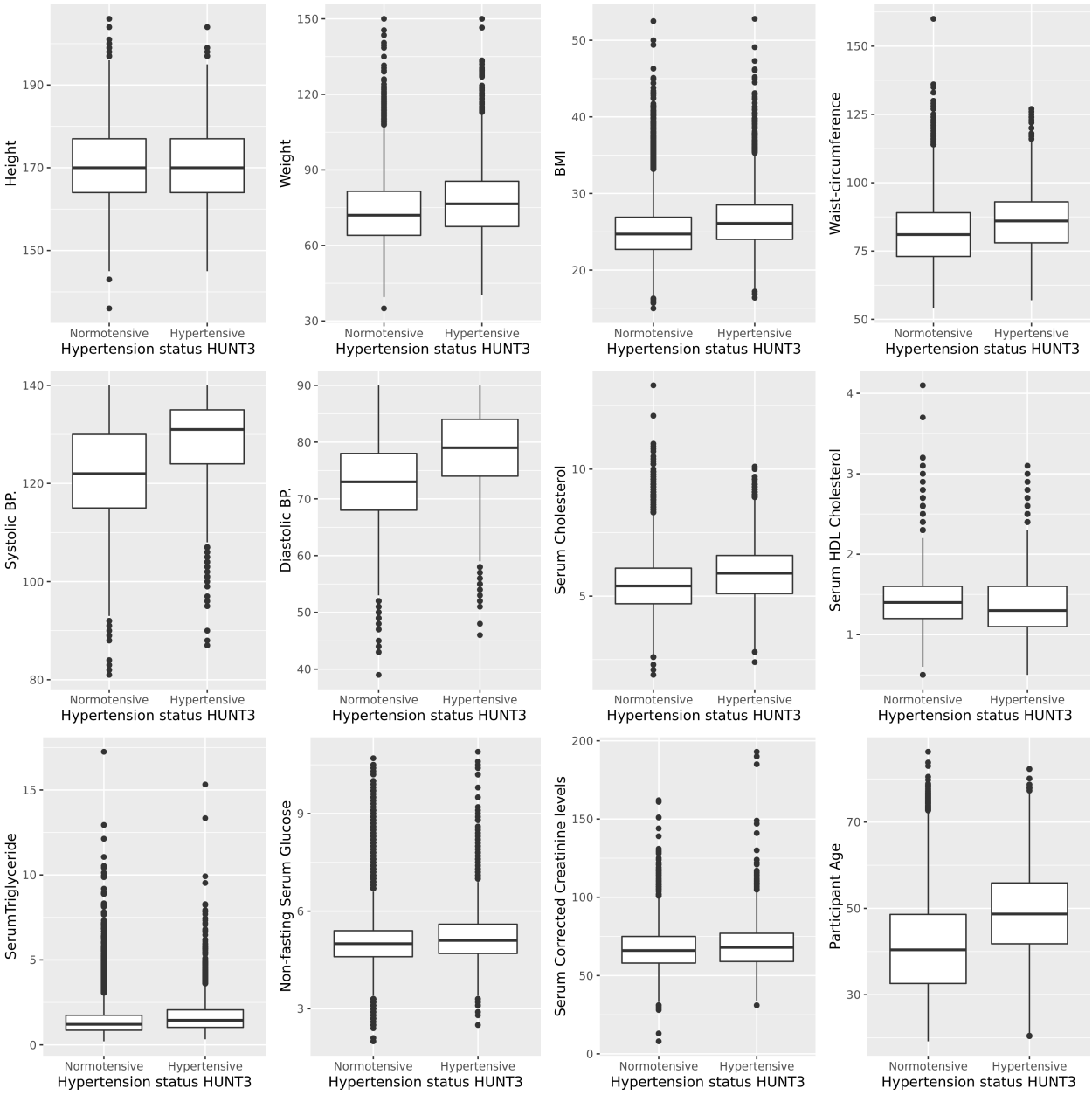


Figure 4.1: Continuous variable distributions if split by endpoint status.

Chapter 5

Methodology

The analysis is performed by fitting models from the three model families of logistic regression, random forest and neural network model families. These are subject to the exact same model fitting and evaluation routine for measuring the predictive performances of each model-family. Further, each model is applied with model-specific regularisation methods and tuningparameter searches. Using the fitted models, feature importance measures are calculated for each model family to determine key features for hypertension risk modelling.

5.1 Analysis setup common for all model-families

This section details the components of the analysis setup that was common for models of all model-families.

5.1.1 Training and testing regime

This describes the full training and testing regime.

5.1.1.1 Routine for a single pair of training and test set

To obtain a measure of the generalization error for models, the dataset $\boldsymbol{D}$ is divided into derivation set $\boldsymbol{D}_{Fit} = (\boldsymbol{X}_{Fit}, \boldsymbol{y}_{Fit})$ and one test set $\boldsymbol{D}_{Test} = (\boldsymbol{X}_{Test}, \boldsymbol{y}_{Test})$, with $\boldsymbol{D}_{Fit} \cap \boldsymbol{D}_{Test} = \emptyset$. The training set is fixed to be 4 times larger than the test set.

To find suitable tuningparameters for each model-family, a tuning parameter search specific for each model family is performed. Each tuning parameter configuration in the search is evaluated by using 4-fold cross-validation as described in section 2.4.1. The configuration that achieves the highest mean AUC_{ROC} on the out-of-fold evaluations is chosen. The final model using this configuration is then fitted on the whole

training set. A final evaluation is made on the test set as a measure of the models generalization performance.

The choice of AUC_{ROC} was due to this measure being the most frequently reported measure for models in the relevant literature, which makes comparison of results with those in literature more relevant.

5.1.1.2 Repeated training and testing routine

As an attempt to distinguish between the variability in the data and the variability of the models from the different model families, the training and testing procedure described in section 5.1.1.1 is repeated independently 20 times. This will in turn produce 20 performance scores which detail the variability in performance scores occuring from different splits in the dataset.

5.1.2 Class imbalance loss

The training and evaluation regime as described in sections 5.1.1.1, 5.1.1.2 is repeated once, with the only difference being that an adjusted loss function is used. The loss function is adjusted as described in section 2.1.6. The scaling factor τ is set to R for the hypertensive class, where R is the ratio of normotensives to hypertensives in the appropriate training set that is used.

5.1.3 Standardization

All models are paired with a standardization routine where the continuous features are univariately standardized to have mean 0 and variance 1. This serves multiple purposes. Standardization is a common approach for achieving faster convergence when gradient-based optimization methods are used [9]. These methods are used to fit model from the logistic regression and neural networks model families. In addition, the coefficient sizes in logistic regression are easier to compare against each other as the coefficient sizes are relative to identical data distributions. The features mean and variance was estimated only from the appropriate training set in use, to avoid introducing any bias.

5.1.4 Permutation importance

After a final model has been selected, permutation importance as described in section 2.3.2 is calculated for all features using the appropriate training set. The number of permutations for each feature in each model is fixed at 20. Some features displaying a high level of correlation were permuted jointly, to assess a grouped permutation importance. The resulting permutation importances are aggregated for each model-family. The choice of calculating this measure on the training set was due to using the

importance scores to identify feature subsets that could be used in fitting new models. In this way, we avoid being biased by calculating the scores on the test set.

5.1.5 Feature inputs

All models are fitted using only linear inputs of the features, i.e. no transformation except standardization is performed on the features. The motivation behind this is that the logistic regression family will only fit linear effects of the features, while the random forest and neural network model families are still capable to fit to potentially non-linear feature effects. This provides a nice contrast when results are compared.

As the features 'BMI', 'Weight' and 'Waist-circumference' were highly correlated, only the latter was used of the three.

5.2 Analysis setup specific for each model-family

The following section detail parts of the analysis setup that was specific for models in each model-family.

5.2.1 Logistic regression family

The logistic regression models are implemented with regularization as described in section 2.1.3.1. The models are fitted to minize the binary cross-entropy loss function plus the regularisation penalty. The tuning parameter search strategy for these models is chosen to be a grid-search, i.e. that all possible combinations are trialled. The axis defining the grid is displayed in Table 5.1. Note that a regularization intensity of 0 corresponds to unregularized modellfitting.

Table 5.1: Hyperparameter grid values used in tuningparameter search for logistic regression.

Parameter	Grid values
- Regularization intensity	$\{0, 10^{-4}, 10^{-3.9} \ldots, 10^{3.9}, 10^{4}\}$.
- Elastic loss ratio	$\{0, 0.1, \ldots, 0.9, 1\}$

5.2.2 Random forest family

Random forest models are implemented using cost-complexity pruning as described in section 2.1.4.1. Since pruning is used, the individual decision trees are fitted without restrictions on size. The number of features considered for each split is chosen to be 5. Each forest is constructed from 100 decision trees. The tuningparameter α used

Table 5.2: Grid values defined for the neural network tuningparameters. Bayesian tuningparameter-search is applied on this grid to find well-performing tuningparameters.

Tuningparameter	Grid values
Depth	$\{1, 2, 3, 4, 5\}$
Width	$\{32, 64, 128, 256\}$
Activation function	{'ReLU', 'GELU', 'Tanh', 'Sigmoid'}
Batch-size	$\{16, 32, 64, 128\}$
Learning-rate, α	$\alpha \in [10^{-3}, 10^{-1}]$
Learning-rate decay	$\{0\%, 1\%, 2.5\%, 5\%\}$ per epoch.
Dropout rate π_{Drop}	$\{0, 0.1, \ldots, 0.9\}$

in pruning is selected by an exhaustive search among the values $\{0, 0.001, \ldots, 0.02\}$. Note that $\alpha = 0$ corresponds to no regularisation.

5.2.3 Neural network family

Modelfitting for neural network models is most often a lot more computationally expensive compared to logistic regression and random forest models. In addition, the number of tuningparameters is higher. This makes grid-searches time-consuming. To search for tuningparameters, a Bayesian search as described in section 2.4.2 is used, with a large value of κ to ensure the search-space is explored thoroughly. The axis of the tuningparameter search grid is displayed in Table 5.2. Note that having 0 as the value for π_{Drop} in Dropout means not using Dropout. The number of search iterations is set to 50, with the first 15 points being random samples from the grid.

Using the selected tuningparameter configuration found in the search, a Bayesian method named Multi-SWAG is applied [18]. This method introduces several more tuningparameters, but these are fixed for simplicity to values suggested in the article describing the method and well-performing values. The number of converged models SWAG is applied to is 20 models, each using the selected tuningparameter configuration. 30 parameter samples are taken per model to estimate the models parameter posterior distribution. To construct predictions, 10 samples is drawn from the posterior distribution, yielding 10 models per the converged model SWAG is applied to. This produces a total of 200 predictions to use in the Bayesian model averaging. A predictions is made by a simple mean.

The non-linear activation functions mentioned in Table 5.2 are shown in Figure 2.2.

5.3 The Framingham model

The Framingham model [8] is one of the more known risk prediction models. To compare to the results achieved in this analysis, the continuous version of the Framingham risk prediction model was implemented to validate the model on HUNT data. Using the continuous model, the 4 year risk of incident hypertension is calculated as:

$$
\begin{aligned}
4 \text{ year risk } &= 1 - e^{\tilde{z}}, \quad \tilde{z} = -4e^{\tilde{\boldsymbol{x}}\boldsymbol{\beta}/0.8769} \\
\text{where } \tilde{\boldsymbol{x}}\boldsymbol{\beta} &= 22.9495 - 0.1564 \times \text{'Age'} - 0.2029 \times \text{'Woman'} - 0.0593 \times \text{'systolic BP'} \\
&\quad - 0.1285 \times \text{'diastolic BP.'} - 0.1907 \times \text{'Smoking'} - 0.1661 \times \text{'Parental hypertension'} \\
&\quad - 0.0339 \times \text{'BMI'} + 0.0016 \times \text{'Age'} \times \text{'diastolic BP'}
\end{aligned}
$$

Note the double exponential. The features used in the Framingham model are not exactly the same as the ones used in the analysis. The features 'Parental hypertension' differs by being the number of parents suffering or having suffered from hypertension that an individual has. The feature 'Smoking' differ by only encoding if an individual smokes or not. A subset fulfilling the inclusion criterias defined in section 4.1, but with the features used in the Framingham model is extracted from the HUNT data. This yielded 15776 participants.

The dataset containing these are referred to as the Framingham data. It should be noted that the differing features are more restrictive definitions of the those defined and used for the analysis, i.e. all participants in the Framingham data are also used in the analysis.

In addition, the risk model only modelled 4 year risk while the HUNT data was collected with 11 years between measurements and endpoint. To accommodate this, 7 years are added to the age of the participants in an adjusted evaluation. Each evaluation of the Framingham model is calculated only once on the whole Framingham data, since the model was fitted using an external dataset.

5.4 Implementation

All code for implementing models have been written by the thesis author, except a script for calculating the feature Physical Activity Indicator. That script was supplied as a `R`-script written by Emma Ingeström. The logistic regression and random forest models as well as the tuningparameter search methods were implemented using the `Python` library `Scikit-learn` [53]. Models from the neural network family were implemented using the `Python` library `PyTorch` [17]. To speed up computation, modelfitting was implemented in parallel using the `Multiprocessing` library [54]. Preprocessing, visualization and post-processing of results was done in `R` [55] using

the integrated development environment `RStudio` [56]. `R` packages used for these purposes include `ggplot2` [57], `reticulate` [58] and `Tidyverse` [59].

All data was hosted and analysis performed over remote-connection to HUNT cloud, an external computing infrastructure hosted for research on HUNT data[1]. HUNT cloud is hosted as part of the HUNT data center, owned by the Director of HUNT Research Centre and affiliated with HUNT Research Centre, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, NTNU, Norway. Due to privacy reasons, the data used in this thesis is not publically available and not included in this thesis[2]

[1] https://www.ntnu.edu/mh/huntcloud
[2] http://www.ntnu.edu/hunt/data

Chapter 6

Results

This chapter contains the main results obtained from the analysis performed as described in Chapter 5. The results are derived from the evaluation of models on the test sets, which was repeated 20 times to produce 20 evaluations per performance and importance measure for each model-family. The results are presented first for logistic regression, then random forest and neural networks before the results are collected in a single table. Along with the results for each model, the no-skill model that gives 0.214 as its prediction for all datapoints is included as a reference.

6.1 Logistic regression models

6.1.1 Performance measure scores

The logistic regression models performed well. Excluding Brier score in balanced models, the models outperformed the no-skill model on all measures. There seems to be little variability in the results, indicating that the fitted models are able to generalize somewhat equally well. Comparing the models to their counterparts using balanced loss, there is no difference in the discriminative performance measures. However, the Brier and Tjur measures increase using balanced loss.

Table 6.1: Results for logistic regression models using the full feature set. LR = Logistic Regression.

Model	AUC_{ROC}	AUC_{PRC}	Brier	Tjur
LR	0.787 ± 0.007	$\mathbf{0.469} \pm 0.014$	$\mathbf{0.139} \pm 0.003$	0.178 ± 0.006
LR, balanced	$\mathbf{0.788} \pm 0.007$	0.468 ± 0.014	0.192 ± 0.003	$\mathbf{0.251} \pm 0.006$
No-skill model	0.5	0.214	0.1684	0

Note that a balanced loss weights performance on each class equally. However, the

performance measures are calculated on populations where the class distributions are not equally large.

6.1.2 Feature importance scores

The importance of features used in the logistic regression models was assessed using Lasso loss with increasing penalty and permutation importance. The logistic regression coefficients using Lasso loss vs. increasing penalty is shown in Figure 6.1a. The permutation importance is shown in Figure 6.3. The results for the two methods are more or less consistent with each other. 'Age', 'systolic BP.' and 'diastolic BP' are the most important features, while 'Waist circumference' and 'Hypertension history in close family' as two notable features. Permuting the features 'Age', 'systolic BP.' and 'diastolic BP' jointly dropped the AUC_{ROC} by so much that model performs only slightly better than random. Notice that this corresponds nicely with the Lasso importance plot, as the mean out-of-fold AUC_{ROC} is barely affected as the penalty increases, dropping only slighly until 'Age', 'systolic BP.' and 'diastolic BP' are zeroed out along with the others.

6.1.3 Results using feature subsets

Motivated by the feature importances scores found for all models, see Figures 6.1 and 6.3, two subsets of features were defined. The modellfitting and evaluation routine were repeated using these subsets of features for logistic regression. The two subsets were:

Reduced: 'Age', 'systolic BP.', 'diastolic BP', 'Waist-circumference' and 'History of hypertension in close family'

Minimal: 'Age', 'systolic BP.' and 'diastolic BP'

These are referred to as the *reduced* and the *minimal* feature sets, respectively. The performance scores obtained for models fitted using these feature subsets, along with those obtained using the full feature set can be seen in Table 6.2.

The obtained performance scores using the reduced and minimal features sets are quite good considering that $> 70\%$ of the included features in the full set are not utilized. For the reduced feature set, the performance was more or less identical to using the full feature set, although a slightly lower mean was observed on all measures. Using the minimal feature set, the performance scores were slightly worse than using the two other feature sets. In addition, the same pattern for models using balanced loss fitted to the full feature set was observed for the models using the feature subsets.

Table 6.2: Results for logistic regression models using different feature subsets. LR = Logistic Regression.

Model	AUC_{ROC}	AUC_{PRC}	Brier	Tjur
Full feature set:				
LR	0.787 ± 0.007	$\mathbf{0.469} \pm 0.014$	0.139 ± 0.003	0.178 ± 0.006
LR, balanced	$\mathbf{0.788} \pm 0.007$	0.468 ± 0.014	0.192 ± 0.003	$\mathbf{0.251} \pm 0.006$
Reduced feature set:				
LR	0.785 ± 0.006	0.462 ± 0.015	$\mathbf{0.139} \pm 0.003$	0.174 ± 0.006
LR, balanced	0.785 ± 0.006	0.461 ± 0.014	0.193 ± 0.003	0.246 ± 0.006
Minimal feature set:				
LR	0.778 ± 0.007	0.449 ± 0.015	0.141 ± 0.003	0.162 ± 0.005
LR, balanced	0.778 ± 0.007	0.448 ± 0.015	0.196 ± 0.003	0.230 ± 0.008
No-skill model	0.5	0.214	0.1684	0

6.1.4 Coefficient sizes

The sampling distribution for coefficients in the logistic regression models are shown in Tables 4 and 5 split by models fitted using different feature sets or balanced loss. A subtable for the most important features 'Age', 'systolic BP.', 'diastolic BP', 'Waist-circumference' and 'History of hypertension in close family' is displayed in Table 6.3.

Table 6.3: Summary statistics for the coefficient sizes of the 20 fitted logistic regression models fitted with and without balanced loss. Subtable of Tables 4 and 5 in the appendix.

Feature	Full	Reduced	Minimal
Systolic BP. (Sys)	$\mathbf{0.5914} \pm 0.0171$	0.5413 ± 0.0175	0.5528 ± 0.0303
Diastolic BP. (Dia)	0.4276 ± 0.0145	0.4446 ± 0.0135	0.4705 ± 0.0214
Age	0.5042 ± 0.0145	$\mathbf{0.5602} \pm 0.011$	$\mathbf{0.5558} \pm 0.0317$
Hyp. history in fam.	0.3409 ± 0.0222	0.3593 ± 0.0209	
Waist circumference	0.1986 ± 0.0117	0.203 ± 0.0072	
Using balanced loss:			
Systolic BP. (Sys)	$\mathbf{0.5964} \pm 0.0159$	0.5443 ± 0.0174	0.5515 ± 0.0228
Diastolic BP. (Dia)	0.4322 ± 0.0137	0.4511 ± 0.0126	0.474 ± 0.0178
Age	0.5455 ± 0.0162	$\mathbf{0.6156} \pm 0.0103$	$\mathbf{0.6058} \pm 0.0269$
Hyp. history in fam.	0.3683 ± 0.0229	0.3753 ± 0.0297	
Waist circumference	0.2097 ± 0.012	0.2139 ± 0.0079	

The major difference in coefficient sizes for models with and without balanced loss

is that the coefficient corresponding to 'Age' is higher for balance loss. The coefficient mean for other features are within a standard deviation of the balanced loss models feature mean. Note also that the coefficient sizes change as we go from the full to the reduced and minimial feaure sets.

6.1.5 Selected tuningparameters

The tuningparameters that were found in the grid-search to be best-performing are listed in Table 6 in the appendix. A common trait was that a regularised logistic regression model with a penalty $\lambda \approx 10$ was found to perform the best at all iterations when the full feature set was used. For the reduced and minimal feature sets more unregularized models were chosen, but without any obvious patterns emerging. The was found for the models using balanced loss.

6.2 Random forest models

6.2.1 Performance measure scores

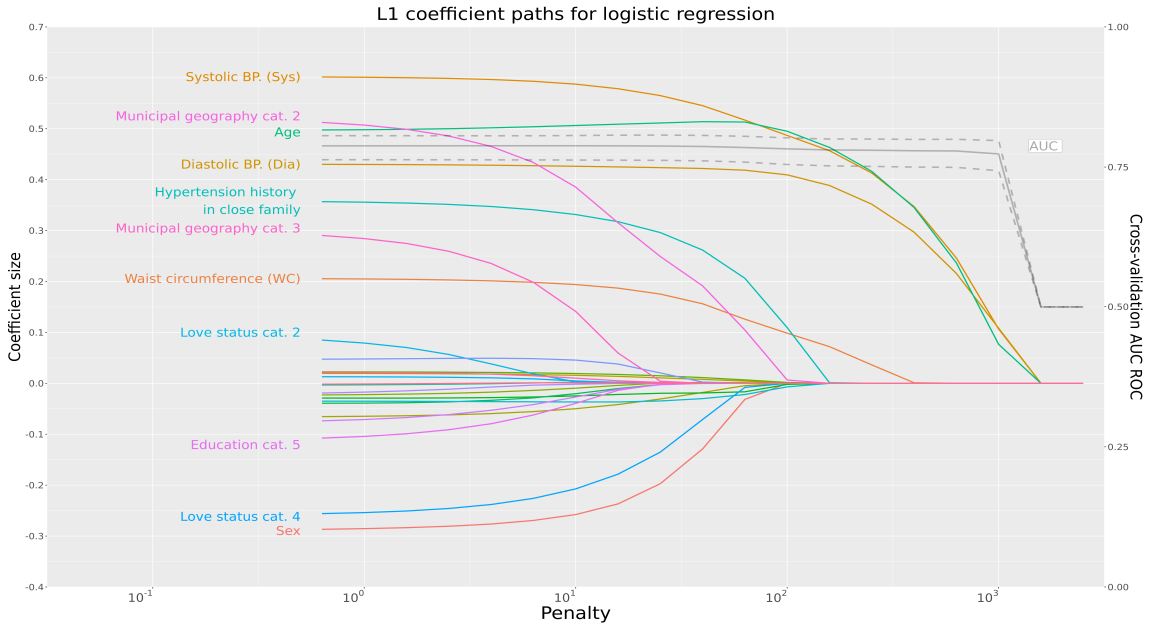
The random forest model performs well, outperforming the no-skill model on all measures, except the balanced model on Brier score. The same pattern for models using balanced loss as for logistic regression models was seen.

Table 6.4: Results for random forest models. RF = Random Forest.

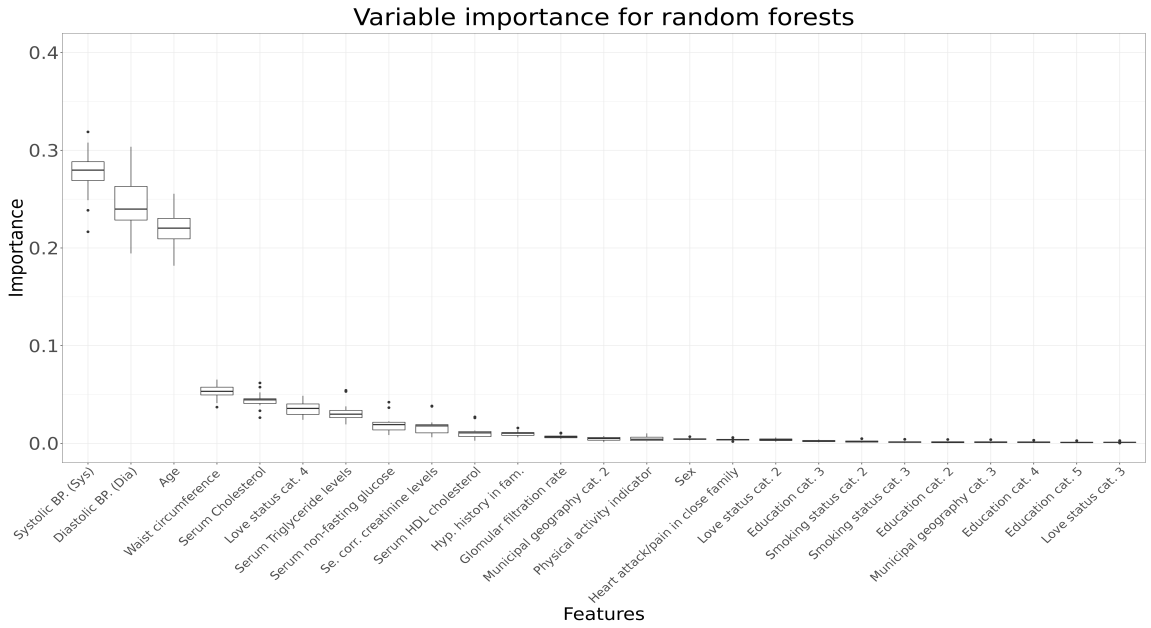
Model	AUC_{ROC}	AUC_{PRC}	Brier	Tjur
RF	**0.781** $\pm$ 0.006	**0.460** $\pm$ 0.013	**0.141** $\pm$ 0.002	0.139 $\pm$ 0.005
RF, balanced	0.781 $\pm$ 0.005	0.459 $\pm$ 0.011	0.186 $\pm$ 0.004	**0.214** $\pm$ 0.007
No-skill model	0.5	0.214	0.1684	0

6.2.2 Feature importance scores

The importance of features used in the random forest models was assessed using variable importance and permutation importance. The variable importance is shown in Figure 6.1. The permutation importance is shown in Figure 6.3. The features regarded as important in logistic regression are important for random forest models as well. In addition, the variable importance plot suggests that the features 'Love status', 'Serum cholesterol' and 'Serum Triglyceride' may be of some importance. However, this does not seem to be supported by their permutation importance.



(a) Mean coefficient size vs L1 penalty for logistic regression. For penalties $0 < \lambda < 10^{-1}$, coefficient sizes are more or less constant at their max absolute value. The solid grey line indicate the mean out-of-fold AUC_{ROC} during tuning-parameter search vs. increasing penalty, with dashed maximum and minimum lines.



(b) Variable importance measured by mean decrease in impurity in random forest models

Figure 6.1: Feature importance scores for logistic regression and random forest

6.2.3 Models using reduced feature sets

As decribed in section 6.1.3, random forest models were fitted using two subsets of features. Table 6.5 show the obtained performance using these subsets in addition to the performance of random forest models when the full feature set is used.

Table 6.5: Results for Random forest models using different feature subsets. RF = Random Forest.

Model	AUC_{ROC}	AUC_{PRC}	Brier	Tjur
Full feature set:				
RF	0.781 ± 0.006	0.460 ± 0.013	0.141 ± 0.002	0.139 ± 0.005
RF, balanced	0.781 ± 0.005	0.459 ± 0.011	0.186 ± 0.004	0.214 ± 0.007
Reduced feature set:				
RF	$\mathbf{0.782} \pm 0.006$	$\mathbf{0.462} \pm 0.015$	$\mathbf{0.140} \pm 0.002$	0.157 ± 0.005
RF, balanced	$\mathbf{0.782} \pm 0.005$	0.459 ± 0.014	0.191 ± 0.003	$\mathbf{0.230} \pm 0.005$
Minimal feature set:				
RF	0.778 ± 0.006	0.456 ± 0.015	0.141 ± 0.002	0.150 ± 0.006
RF, balanced	0.778 ± 0.006	0.454 ± 0.013	0.194 ± 0.003	0.220 ± 0.005
No-skill model	0.5	0.214	0.1684	0

The random forest models using the reduced feature set obtained better performance score means than the models using the full feature set. This is seen as the Tjur score is clearly higher while the other measures are somewhat equal. This is true for the model fitted using the minimal feature set as well. The effect of using a balanced loss is the same regardless of the feature set used. The random forest models using the minimal set performed more or less equally well as the two others.

6.2.4 Selected tuningparameters

The selected pruning intensity for the random forest models was very low. Table 7 in the appendix displays all tuningparameter chosen, split by model iteration and by usage of balanced loss. Along with the pruning intensity, the mean and standard deviation on the number of end nodes per decision tree in the random forest models is reported. All pruning intensities were low, in the range $[3, 11] \times 10^{-4}$. In the random forests fitted using the reduced and minimal features sets, the average number of endnodes in each decision tree seems lower. No discernable difference could be spotted between models fitted with or without balanced loss. In Figure 6.2, the mean out-of-fold AUC_{ROC} score is plotted vs. the pruning intensity used in the tuning-parameter search. In short, some regularization is useful, but this was no surprise considering that the individual trees were overfitted by intention before pruning them.

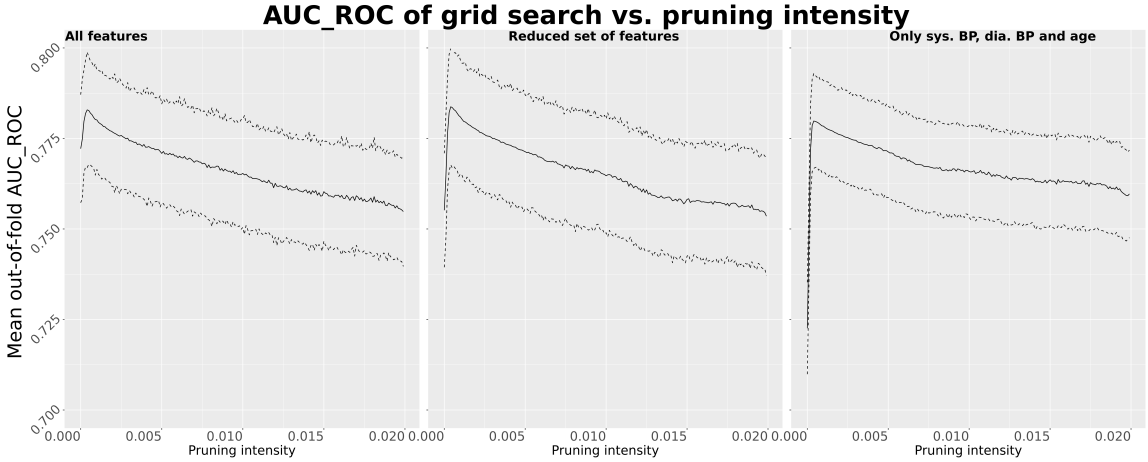


Figure 6.2: Out-of-fold AUC_{ROC} score obtained during tuning-parameter search vs. pruning intensity for Random forest models. Solid line is mean with dashed lines 2 standard deviations.

6.3 Neural network models

6.3.1 Performance measure scores

Both the single neural network models and the ensembling method Multi-SWAG produced well-performing models. The usage of Multi-SWAG improved scores on all measures. However, a question can be asked as to how effective the method is, as the Multi-SWAG ensemble is comprised of 200 predictions made from 20 neural networks. The increase in computation is high, while the mean performance scores for a neural network and its Multi-SWAG ensemble are within 2 standard deviation of each other.

Table 6.6: Results for neural network models. NN = Neural network.

Model	AUC_{ROC}	AUC_{PRC}	Brier	Tjur
NN	0.777 ± 0.007	0.449 ± 0.014	0.144 ± 0.005	0.159 ± 0.027
NN, balanced	0.779 ± 0.007	0.451 ± 0.017	0.199 ± 0.012	$\mathbf{0.240} \pm 0.017$
Multi-SWAG	0.785 ± 0.007	0.461 ± 0.016	$\mathbf{0.140} \pm 0.002$	0.160 ± 0.010
Multi-SWAG, balanced	$\mathbf{0.786} \pm 0.006$	$\mathbf{0.465} \pm 0.015$	0.194 ± 0.003	0.234 ± 0.012
No-skill model	0.5	0.214	0.1684	0

6.3.2 Feature importance scores

The permutation measure calculated for neural networks are shown in Figure 6.3. The scores are coinciding with those calculated for logistic regression and random forest models. There was a slightly larger variation in the importance estimates of neural networks, but this is likely due to the larger variation in models from the neural network family compared to the others.

6.3.3 Selected tuningparameters

The tuningparameters that were found in the grid-search to be best-performing are listed in Table 8 in the appendix. There seems to be no obvious pattern, except that the maximum size of intermediate activations, the width of the network, was 128 across all models. In addition, it should be noted that models with a large number of parameters as well as few were chosen. E.g. in Table 8, on the 9th run in the analysis, a model with depth 5 and width 128 and thus having more than 60 000 parameters was selected for the models without balanced loss. On the 19th run in the analysis, a model using fewer than 500 parameters was selected for those with balanced loss. Judging by the performance scores, both models were quite similar in performance.

6.4 Framingham model

The performance scores for the Framingham model is shown in Table 6.7. Note that the results was obtained by a single evaluation, since none of the applicable data had been used in fitting the model.

Table 6.7: Performance of the Framingham 4-year hypertension risk model on the Framingham data. For the adjusted evaluation, 'Age' is added 7 years.

Model	AUC_{ROC}	AUC_{PRC}	Brier	Tjur
Framingham	**0.79**	0.458	**0.139**	**0.118**
Framingham, unadj.	**0.79**	**0.46**	0.143	0.088

6.5 Comparison across model families

A figure showingcasing the results for models fitted on the full feature set are given in the appendix, see Figure 3. A table containing the performance measures of models from all model families, split by usage of balanced loss and which feature set fitted on, is given in Table 6.8. It can be seen that all models exhibit the same pattern when

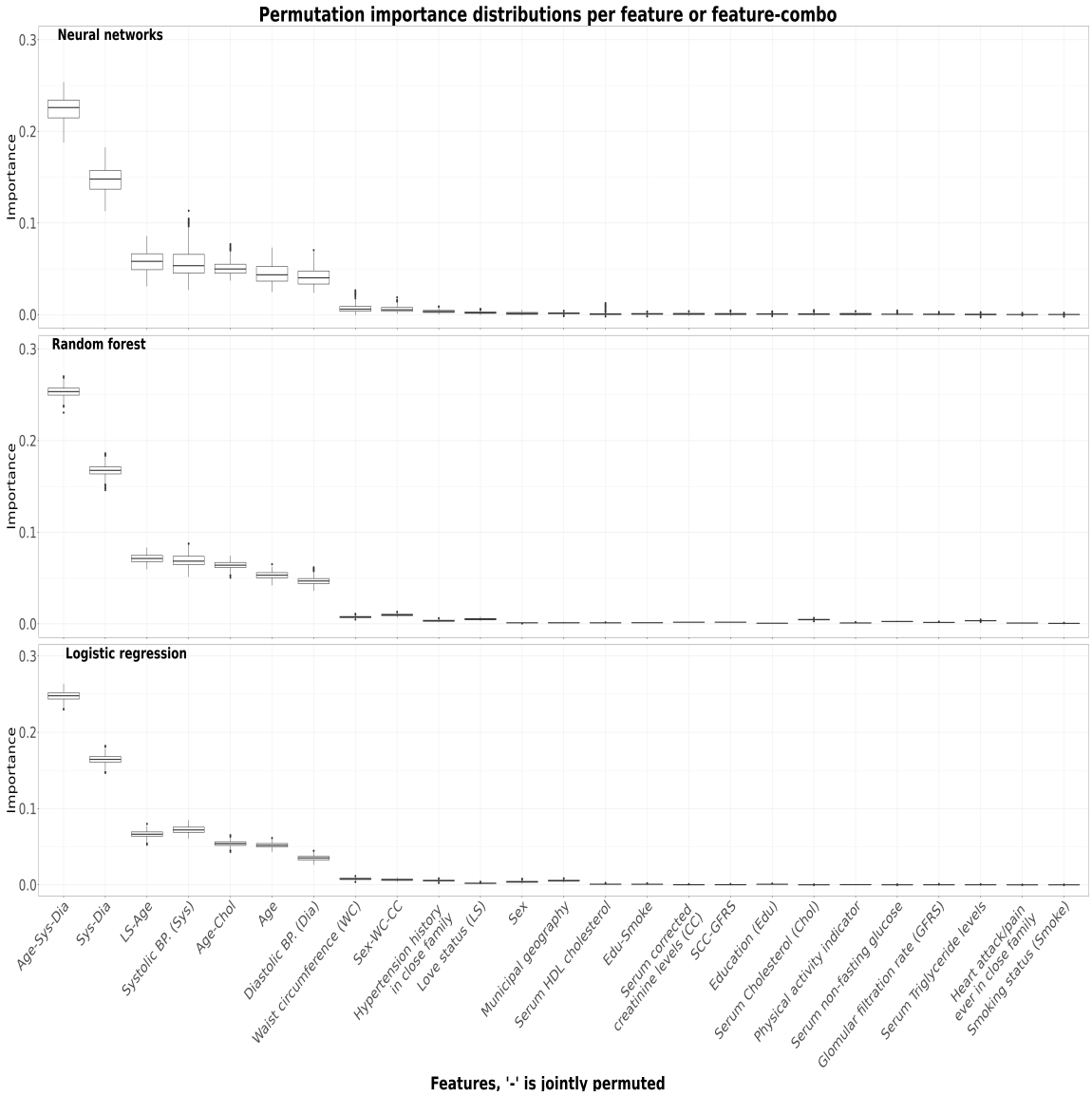


Figure 6.3: Permutation importance measured as decrease in AUC_{ROC} for neural network, logistic regression and random forest model. Sorted by their median importance in neural networks.

Table 6.8: Model result distributions by model family and feature set that have been used in fitting the model. LR = Logistic regression, RF = Random Forest, NN = Neural Network.

Model	AUC_{ROC}	AUC_{PRC}	Brier	Tjur
Full feature set:				
LR	0.787 ± 0.007	$\mathbf{0.469} \pm 0.014$	$\mathbf{0.139} \pm 0.003$	0.178 ± 0.006
LR, balanced	0.788 ± 0.007	0.468 ± 0.014	0.192 ± 0.003	$\mathbf{0.251} \pm 0.006$
RF	0.781 ± 0.006	0.460 ± 0.013	0.141 ± 0.002	0.139 ± 0.005
RF, balanced	0.781 ± 0.005	0.459 ± 0.011	0.186 ± 0.004	0.214 ± 0.007
NN	0.777 ± 0.007	0.449 ± 0.014	0.144 ± 0.005	0.159 ± 0.027
NN, balanced	0.779 ± 0.007	0.451 ± 0.017	0.199 ± 0.012	0.240 ± 0.017
Multi-SWAG	0.785 ± 0.007	0.461 ± 0.016	0.140 ± 0.002	0.160 ± 0.010
Multi-SWAG, balanced	0.786 ± 0.006	0.465 ± 0.015	0.194 ± 0.003	0.234 ± 0.012
Reduced feature set:				
LR	0.785 ± 0.006	0.462 ± 0.015	$\mathbf{0.139} \pm 0.003$	0.174 ± 0.006
LR, balanced	0.785 ± 0.006	0.461 ± 0.014	0.193 ± 0.003	0.246 ± 0.006
RF	0.782 ± 0.006	0.462 ± 0.015	0.140 ± 0.002	0.157 ± 0.005
RF, balanced	0.782 ± 0.005	0.459 ± 0.014	0.191 ± 0.003	0.230 ± 0.005
Minimal feature set:				
LR	0.778 ± 0.007	0.449 ± 0.015	0.141 ± 0.003	0.162 ± 0.005
LR, balanced	0.778 ± 0.007	0.448 ± 0.015	0.196 ± 0.003	0.230 ± 0.008
RF	0.778 ± 0.006	0.456 ± 0.015	0.141 ± 0.002	0.150 ± 0.006
RF, balanced	0.778 ± 0.006	0.454 ± 0.013	0.194 ± 0.003	0.220 ± 0.005
Framingham feature set:				
Framingham	**0.79**	0.458	**0.139**	0.118
Framingham, unadj.	**0.79**	0.46	0.143	0.088
No-skill model	0.5	0.214	0.1684	0

fitted with balanced loss. In addition, the performance scores are quite similar across model-families and feature sets.

Chapter 7

Discussion

In this chapter, we discuss the performance measures obtained in the analysis, which features that were important, what the model families may indicate about feature effects and how they compare against the models found in the literature.

7.1 Models using the full feature set

Reviewing the results listed in Table 6.8, the performance scores obtained for the different model families indicate that all model families can be used to construct fairly well-performing prediction models. This is evidenced by the fact that all models achieve performance scores that are substantially better than what a no-skill discriminator would achieve for each performance measure, with an exception for models using balanced loss, on Brier score. However, as the performance scores are quite similar across model families, it is not appropriate to distinguish one model family as superior based solely on performance measures. The standard deviation of performance measures was larger than the difference in mean between model families. This indicates that the variability in the data have a greater effect on a models performance compared to the particular choice of model family. Judging by the mean and median, the logistic regression model family scored slighly better at all measures, regardless of using balanced loss or not, excluding the Framingham model.

The fact that logistic regression using only linear feature effects performs as well or better than models from the two other model families, suggests that the problem is adequately solved without non-linear feature effects. While models from the two other model families are highly capable of learning non-linear feature effects, none succeeded in producing a single model that outperformed all logistic regression models. This is shown in Figure 3. In the case that there are any non-linear feature effects that these models have learned, the results seems to indicate that there is

little to none additional predictive power associated with them compared to linear effects.

By investigating the feature importance scores, it is evident that the most important features for all model families were 'Age', 'Systolic' and 'Diastolic' blood pressure. In addition, other features such as 'Waist-Circumference', 'Cholesterol' and 'Hypertension history in close family' were also emphasized, but with less importance. The combination of 'Age', 'Systolic' and 'Diastolic' blood pressure lead to a median decrease in the range of $[0.23, 0.25]$ on AUC_{ROC} scores for models fitted without balanced loss. A reductin in that range for AUC_{ROC} would mean the models performed only a bit better than the no-skill model. In addition, the combination of the three features has a greater reduction in AUC_{ROC} than the sum of the individual reductions. In total, all model families seems to be particularly dependent on these three for good predictive performance.
The model-setups that were found in the tuningparameter searches seems to indicate that there are multiple model-setups for neural networks that work satisfactory. There is no apparent pattern in the tuningparameters chosen for each best performing neural network, apart from the dropout-rate being relatively low and that the maximal width was not used. However, the neural networks were of wildly different sizes, yet still achieved somewhat similar scores. The lack of a pattern in tuningparameters for the neural networks might be expected, since the tuningparameter search space used to find models was large and not all combinations were tried out. As for the logistic regression and random forest models, some regularisation were needed to obtain the best performing models. Judging by the tuningparameters chosen for the logistic regression and random forest models, a relatively low amount of regularization is what gave the best performing models in general.

7.2 Results for models using feature subset

Fitting logistic regression and random forest models using only linear effects of 'Age', 'Systolic' and 'Diastolic' blood pressure, 'Waist-Circumference', and 'Hypertension history in close family' achieved good performance scores despite reducing the number of features by more than 70%. Judging by the mean performance value on the reduced set, logistic regression seems to have little loss in performance, while random forest slightly improves its Brier score and Tjur score compared to using the full set. This was somewhat surprising, but may be an indication that the effect of adding the less-important features was introducing more noise in the feature-space.
In Table 6.3, looking at the coefficient sizes as the features set changes, it is interesting to see that the coefficient size of 'systolic BP', 'diastolic BP' and 'Age' changes relatively much as smaller feature sets are used. Considering that the performance

measures barely changed between the full and reduced feature set, it could indicate that any predictive information in the removed features are already present in those of the reduced set.

7.3 The usage of balanced loss in modellfitting

Comparing the models to their counterparts using balanced loss, there is no difference in the discriminative performance measures. However, the Brier and Tjur measures increase using balanced loss. An interpretation is that the models become overconfident as the higher Tjur measure show a larger gap in average confidence for each class. Combined with the higher Brier score, the two measures indicates that the models are more confident on datapoints it predicts wrongly.

It is reasonable to expect that the predictions made by a model should have its mean close to the percentage of hypertensives in the dataset. However, using balanced loss, the scaling in the loss function can be seen as a way of mimicking an equal distribution of hypertensives and normotensives in the dataset. Hence, model predictions should have a larger mean at around 0.5, but not necessarily change how the model would rank datapoints. This might explain the results obtained for balanced models, as the discriminatory scores are unaffected while the predicted probabilities become more erronous and more confident at the same time.

7.4 A candidate for preferred model setup

Although no model produced clearly better results than others, model properties could also be considered when deciding upon an optimal model. Arguments for preferring the logistic regression model with fewer features are properties like ease of use, interpretability and cheap computational requirements. Arguably, neither random forest models or neural networks could be said to be interpretable. In addition, they are both more computationally demanding.

Lastly, the reduced set of features used was equally effective performance-wise as using the full set and slightly better than using the minimal set. In addition, the availability of the features in the reduced model is better than that of the full feature set as no biomarkers needs to be measured. It is only slightly more troublesome to collect than that of the minimal feature set, as an individual would need to know about the hypertension history in the individual's family. As the balanced loss lead to a "overconfidence" effect with equal discriminatory performance, it is not preferred.

Hence, the logistic regression model family fitted without balanced loss and using the reduced feature set could be a likely candidate as the preferred model setup for predicting 11-year hypertension risk using HUNT data.

7.5 Results compared to the literature

As discussed briefly in the literature review in Section 3.2, comparisons with the models found in the literature should be done with some care due to the high variability in modelling choices, study setup and more. Acknowledging this, it is arguably still only possible to properly compare the AUC_{ROC} measure among those reported, as this is the only performance measure reported more or less consistently and with enough detail in the literature.

Separate from this, the Framingham model was implemented and evaluated on an appropriate subset of the data used to fit and evaluate the other models in this thesis. The models performance was good for both the adjusted and unadjusted evaluation. The adjusted model had performance scores comparable with the best model families found in this analysis, suggesting that it is a valid tool to use on the population that the HUNT data is sampled from. However, some variability may have been introduced by using a subset of the total data. Another factor that hamper comparison is using the continuouos version of the risk prediction model rather than the discrete, although they were reported in the original article to be almost identical in their predictions [8].

An AUC_{ROC} score of close to 0.8 for the logistic regression models is arguably reasonable when comparing to other models found in the literature. Most logistic regression models achieve a score in the range $[0.7, 0.85]$, hence the results obtained here are in the upper half. Considering that the outcome-participant ratio in the HUNT data is moderately unbalanced, i.e. 1 hypertensive to 4 normotensives, and more unbalanced than most articles included, the AUC_{ROC} score may be somewhat optimistic compared to other models in the literature. An explanation for this notion is given in section 2.2.2. Four logistic regression and Weibull regression models fitted with equally or more unbalanced data displayed both higher and lower scores.

The AUC_{ROC} scores achieved in this analysis using random forests, multi-SWAG or neural networks models are harder to compare against the literature due to the larger variability in scores and performance reported. For neural network-related models and decision tree-based models, AUC_{ROC} scores were reported in the ranges $[0.7, 0.9]$ and $[0.68, 0.93]$ respectively. In both cases, a clear majority was in the lower end, i.e. < 0.8. The larger variability in results is not surprising as the neural network and decision tree-based model families are two large groups of models. This means models within each family are more likely to be dissimilar in setup, possibly contributing to the variability in performance scores. In addition, there are many articles using models from these families that do not report AUC_{ROC} scores at all, making the comparison more difficult.

To more closely examine the differences in results obtained by using different model families, one could look to the scores found in articles that utilized multiple different model families. Only including those that reported AUC_{ROC} scores as well, this totaled 5 articles. In these articles, the model performances varies more or less

only by model family, since they are applied upon the same data in the same study setup. These are [60], [43], [61], [62] and [63]. Judging by their results, there are some indications that it is possible to achieve higher scores with decision tree methods and neural networks compared to logistic regression. However, this difference may be small and was only shown for a single test set evaluation in these articles.

The performance obtained using the reduced and minimal subsets of features are comparable to some of the results seen in literature. There are multiple smaller models that have achieved decent AUC_{ROC} scores, seeing little improvement compared to models using larger feature-sets in the same article. Some of these were specifically emphasized in the literature review. Several of these models utilized the same features that were found to be highly important in this analysis: Age, systolic BP and diastolic BP.

7.6 Results in light of the dataset used

Some questions may be addressed towards the quality of the data that is used for this analysis. These questions are primarily motivated by the results, i.e. that most of the full feature set could be removed without affecting, and even improving, the performance scores for two model families.

An argument could be that the 11 year time-frame between measurements and the endpoint is affecting the utility of these features. Looking to the literature, a clear majority of prospective studies have less time between baseline and endpoint. However, the results show a trend that is not improving as the time-frame is reduced. A interesting article is that of [42], where the Framingham model is validated on two cohorts with 2 and 4 years as the time-frame. Results improved on the 4-year cohort compared to the 2 year cohort. Another more interesting article is that of [64], where the same model-setup was fitted to 3, 6, 9 year cohorts separately. The performance improved with longer time-frame. In addition, the performance score achieved by the Framingham model using the Framingham data was good, despite the model being fitted to data with 4 years between measurements and endpoint. Although not conclusive, this might indicate that the 11 year time-frame of the HUNT dataset is not inhibiting the performance of the models.

Although neural networks have obtained success in many domains, the model family is notorious for requiring large datasets in doing so. A hypothesis could be that the HUNT dataset is not large enough to properly fit the neural network models. Reviewing the literature, most articles have fewer datapoints than what is available in the dataset. Only a few articles used neural network models and reported AUC_{ROC}. For those that did, there is no clear connection between datasets size and performance scores. In [60], a neural net scored $AUC_{ROC} = 0.9$ on 3000 datapoints, while [61] achieved relatively low scores with roughly 23000 datapoints. In [43], more than 300

000 was used, but the neural network model only performed slightly better than a logistic regression.

Chapter 8

Conclusion and future work

8.1 Conclusion

The aim of this thesis was to construct and evaluate predictive models for hypertension based on data from the HUNT-2 and HUNT-3 studies. In particular we compared prediction models based on logistic regression, random forest and neural network model families. Secondary aims included reviewing the current state of literature on predictive models for hypertension where similar model-families were applied, comparing the results found in the literature to those achieved in this thesis. To fulfill these aims, a setup for choosing well-performing models from the model families of logistic regression, neural networks and random forest was implemented. The analysis was done using a repeated training and testing scheme to obtain empirical distributions of the predictive performance of models from the different model families. Based on feature importance scores, models using two subsets of the full feature set were evaluated. Ultimately, these performance distributions were compared to results found in the literature.

The performances are quite similar across model families, with higher variation in performance due to datavariability than which model-family was used. The results obtained are reasonable compared to that found in literature and to the performance measure of the Framingham model on HUNT data. We conclude that the prediction models from different model families are capable of predicting the hypertension risk more or less equally well. Furthermore, all models were consistent in identifying 'systolic BP.', 'diastolic BP.' and 'Age' as the most important features, along with 'Waist-circumference' and 'Hypertension history in close family' as notable features, judged by the importance scores estimated for each model-family. However, there is no indication that there is non-linear feature effects that are valuable for constructing the prediction models. Taking into account model properties, a logistic regression model using the features 'systolic BP.', 'diastolic BP.', 'Age', 'Waist-circumference' and

'Hypertension history in close family' and fitted with some regularization, but without balanced loss, is proposed as the preffered modelling setup to model hypertension risk using the HUNT data.

8.2 Future work

We suggest multiple possible directions for future work. Two subsets of the data could be identified and further analysed: A subset of data where the models had high-confidence, but erronous predictions, and a subset of data where the different model-families had high disagreement in predictions. Both could possibly yield insight into whether or not it is possible to improve the performance achieved in this thesis. Another avenue would be to construct other features than those used in this thesis, e.g. by utilizing a different measure of exercise than PAI. The literature review revealed that there is a high variation in how research on this topic is reported. A bias assessment could be performed to gain insights into how trustworthy the reported models and their performance are. This would allow for easier comparison of results obtained in different articles. Lastly, it would be interesting to investigate how the importance of features are affected by the time between measurements and endpoint. Considering the analysis in this thesis uses data with 11 years between measurements and endpoint, new studies could be conducted to study this. In the extreme case, this is partly what the MyMDT project is about by using features collected in real-time.

Bibliography

[1] B. Zhou, J. Bentham, M. D. Cesare, *et al.*, «Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants», *The Lancet*, vol. 389, no. 10064, pp. 37–55, Jan. 2017. DOI: 10.1016/s0140-6736(16)31919-5.

[2] S. S. Lim, T. Vos, A. D. Flaxman, *et al.*, «A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010», *The Lancet*, vol. 380, no. 9859, pp. 2224–2260, Dec. 2012. DOI: 10.1016/s0140-6736(12)61766-8.

[3] M. H. Forouzanfar, A. Afshin, L. T. Alexander, *et al.*, «Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015», *The Lancet*, vol. 388, no. 10053, pp. 1659–1724, Oct. 2016. DOI: 10.1016/s0140-6736(16)31679-8.

[4] W. H. Organization, *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*. WORLD HEALTH ORGN, Mar. 1, 2010, 62 pp., ISBN: 9241563877. [Online]. Available: https://apps.who.int/bookorders/anglais/detart1.jsp?sesslan=1&codlan=1&codcol=15&codcch=00772.

[5] T. A. Gaziano, A. Bitton, S. Anand, and M. C. Weinstein, «The global cost of nonoptimal blood pressure», *Journal of Hypertension*, vol. 27, no. 7, pp. 1472–1477, Jul. 2009. DOI: 10.1097/hjh.0b013e32832a9ba3.

[6] O. A. Carretero and S. Oparil, «Essential Hypertension», *Circulation*, vol. 101, no. 3, pp. 329–335, Jan. 2000. DOI: 10.1161/01.cir.101.3.329.

[7] S. Krokstad, A. Langhammer, K. Hveem, T. Holmen, K. Midthjell, T. Stene, G. Bratberg, J. Heggland, and J. Holmen, «Cohort Profile: The HUNT Study, Norway», *International Journal of Epidemiology*, vol. 42, no. 4, pp. 968–977, Aug. 2012. DOI: 10.1093/ije/dys095.

[8] N. I. Parikh, M. J. Pencina, T. J. Wang, E. J. Benjamin, K. J. Lanier, D. Levy, R. B. D'Agostino, W. B. Kannel, and R. S. Vasan, «A Risk Score for Predicting Near-Term Incidence of Hypertension: The Framingham Heart Study», *Annals of Internal Medicine*, vol. 148, no. 2, p. 102, Jan. 2008. DOI: 10.7326/0003-4819-148-2-200801150-00005.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[10] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression: Models, Methods and Applications*. Springer, 2013, ISBN: 978-3-642-34333-9.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics Book 103)*. Springer, 2013, ISBN: 978-1-4614-7138-7.

[12] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[13] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/a:1010933404324.

[14] Y. LeCun, Y. Bengio, and G. Hinton, «Deep learning», *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. DOI: 10.1038/nature14539.

[15] F. Chollet, *Deep Learning with Python*. Manning, Nov. 30, 2017, ISBN: 1617294438. [Online]. Available: https://www.ebook.de/de/product/28930398/francois_chollet_deep_learning_with_python.html.

[16] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, «Multilayer feedforward networks with a nonpolynomial activation function can approximate any function», *Neural Networks*, vol. 6, no. 6, pp. 861–867, Jan. 1993. DOI: 10.1016/s0893-6080(05)80131-5.

[17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, «PyTorch: An imperative style, high-performance deep learning library», in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[18] A. G. Wilson and P. Izmailov, «Bayesian Deep Learning and a Probabilistic Perspective of Generalization», Feb. 20, 2020. arXiv: 2002.08791v3 [cs.LG].

[19] W. Maddox, T. Garipov, P. Izmailov, D. Vetrov, and A. G. Wilson, «A Simple Baseline for Bayesian Uncertainty in Deep Learning», Feb. 7, 2019. arXiv: 1902.02476v2 [cs.LG].

[20] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, «Improving neural networks by preventing co-adaptation of feature detectors», Jul. 3, 2012. arXiv: 1207.0580v1 [cs.NE].

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, «Focal Loss for Dense Object Detection», Aug. 7, 2017. arXiv: 1708.02002v2 [cs.CV].

[22] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, «Assessing the Performance of Prediction Models», *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010. DOI: 10.1097/ede.0b013e3181c30fb2.

[23] H. He and E. Garcia, «Learning from Imbalanced Data», *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009. DOI: 10.1109/tkde.2008.239.

[24] J. A. Hanley and B. J. McNeil, «The meaning and use of the area under a receiver operating characteristic (ROC) curve.», *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982. DOI: 10.1148/radiology.143.1.7063747.

[25] T. Saito and M. Rehmsmeier, «The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets», *PLOS ONE*, vol. 10, no. 3, G. Brock, Ed., e0118432, Mar. 2015. DOI: 10.1371/journal.pone.0118432.

[26] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, Inc., Sep. 2000. DOI: 10.1002/0471722146.

[27] B. V. Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, *Calibration: the Achilles heel of predictive analytics*, Dec. 2019. DOI: 10.1186/s12916-019-1466-7.

[28] H. Hersbach, «Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems», *Weather and Forecasting*, vol. 15, no. 5, pp. 559–570, Oct. 2000. DOI: 10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2.

[29] T. Tjur, «Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination», *The American Statistician*, vol. 63, no. 4, pp. 366–372, Nov. 2009. DOI: 10.1198/tast.2009.08210.

[30] E. Brochu, V. M. Cora, and N. de Freitas, «A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning», Dec. 12, 2010. arXiv: 1012.2599v1 [cs.LG].

[31] P. A. James, S. Oparil, B. L. Carter, W. C. Cushman, C. Dennison-Himmelfarb, J. Handler, D. T. Lackland, M. L. LeFevre, T. D. MacKenzie, O. Ogedegbe, S. C. Smith, L. P. Svetkey, S. J. Taler, R. R. Townsend, J. T. Wright, A. S. Narva, and E. Ortiz, «2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults», *JAMA*, vol. 311, no. 5, p. 507, Feb. 2014. DOI: 10.1001/jama.2013.284427.

[32] B. Williams, G. Mancia, W. Spiering, *et al.*, «2018 ESC/ESH Guidelines for the management of arterial hypertension», *European Heart Journal*, vol. 39, no. 33, pp. 3021–3104, Aug. 2018. DOI: 10.1093/eurheartj/ehy339.

[33] P. K. Whelton and B. Williams, «The 2018 European Society of Cardiology/European Society of Hypertension and 2017 American College of Cardiology/American Heart Association Blood Pressure Guidelines», *JAMA*, vol. 320, no. 17, p. 1749, Nov. 2018. DOI: 10.1001/jama.2018.16755.

[34] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C. Cushman, L. A. Green, J. L. Izzo, D. W. Jones, B. J. Materson, S. Oparil, J. T. Wright, and E. J. R. and, «Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure», *Hypertension*, vol. 42, no. 6, pp. 1206–1252, Dec. 2003. DOI: 10.1161/01.hyp.0000107251.49515.c2.

[35] C. K. Chow, «Prevalence, Awareness, Treatment, and Control of Hypertension in Rural and Urban Communities in High-, Middle-, and Low-Income Countries», *JAMA*, vol. 310, no. 9, p. 959, Sep. 2013. DOI: 10.1001/jama.2013.184182.

[36] J. B. Echouffo-Tcheugui, G. D. Batty, M. Kivimäki, and A. P. Kengne, «Risk Models to Predict Hypertension: A Systematic Review», *PLoS ONE*, vol. 8, no. 7, A. V. Hernandez, Ed., e67370, Jul. 2013. DOI: 10.1371/journal.pone.0067370.

[37] D. Sun, J. Liu, L. Xiao, Y. Liu, Z. Wang, C. Li, Y. Jin, Q. Zhao, and S. Wen, «Recent development of risk-prediction models for incident hypertension: An updated systematic review», *PLOS ONE*, vol. 12, no. 10, T. Shimosawa, Ed., e0187240, Oct. 2017. DOI: 10.1371/journal.pone.0187240.

[38] C. Krittanawong, A. S. Bomback, U. Baber, S. Bangalore, F. H. Messerli, and W. H. W. Tang, «Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension», *Current Hypertension Reports*, vol. 20, no. 9, Jul. 2018. DOI: 10.1007/s11906-018-0875-x.

[39] P. Muntner, M. Woodward, D. M. Mann, D. Shimbo, E. D. Michos, R. S. Blumenthal, A. P. Carson, H. Chen, and D. K. Arnett, «Comparison of the Framingham Heart Study Hypertension Model With Blood Pressure Alone in the Prediction of Risk of Hypertension», *Hypertension*, vol. 55, no. 6, pp. 1339–1345, Jun. 2010. DOI: 10.1161/hypertensionaha.109.149609.

[40] A. P. Carson, G. Howard, G. L. Burke, S. Shea, E. B. Levitan, and P. Muntner, «Ethnic Differences in Hypertension Incidence Among Middle-Aged and Older Adults», *Hypertension*, vol. 57, no. 6, pp. 1101–1107, Jun. 2011. DOI: 10.1161/hypertensionaha.110.168005.

[41] A. Ramezankhani, A. Kabir, O. Pournik, F. Azizi, and F. Hadaegh, «Classification-based data mining for identification of risk patterns associated with hypertension in Middle Eastern population: A 12-year longitudinal study», *Medicine*, vol. 95, no. 35, 2016.

[42] L. Zheng, Z. Sun, X. Zhang, J. Li, D. Hu, J. Chen, and Y. Sun, «Predictive Value for the Rural Chinese Population of the Framingham Hypertension Risk Model: Results From Liaoning Province», *American Journal of Hypertension*, vol. 27, no. 3, pp. 409–414, Dec. 2013. DOI: 10.1093/ajh/hpt229.

[43] A. Wang, N. An, G. Chen, L. Li, and G. Alterovitz, «Predicting hypertension without measurement: A non-invasive, questionnaire-based approach», *Expert Systems with Applications*, vol. 42, no. 21, pp. 7601–7609, Nov. 2015. DOI: 10.1016/j.eswa.2015.06.012.

[44] C. Fava, M. Sjögren, M. Montagnana, E. Danese, P. Almgren, G. Engström, P. Nilsson, B. Hedblad, G. C. Guidi, P. Minuz, and O. Melander, «Prediction of Blood Pressure Changes Over Time and Incidence of Hypertension by a Genetic Risk Score in Swedes», *Hypertension*, vol. 61, no. 2, pp. 319–326, Feb. 2013. DOI: 10.1161/hypertensionaha.112.202655.

[45] X. Lu, J. Huang, L. Wang, S. Chen, X. Yang, J. Li, J. Cao, J. Chen, Y. Li, L. Zhao, H. Li, F. Liu, C. Huang, C. Shen, J. Shen, L. Yu, L. Xu, J. Mu, X. Wu, X. Ji, D. Guo, Z. Zhou, Z. Yang, R. Wang, J. Yang, W. Yan, and D. Gu, «Genetic Predisposition to Higher Blood Pressure Increases Risk of Incident Hypertension and Cardiovascular Diseases in Chinese», *Hypertension*, vol. 66, no. 4, pp. 786–792, Oct. 2015. DOI: 10.1161/hypertensionaha.115.05961.

[46] T. J. Niiranen, A. S. Havulinna, V. L. Langén, V. Salomaa, and A. M. Jula, «Prediction of Blood Pressure and Blood Pressure Change With a Genetic Risk Score», *The Journal of Clinical Hypertension*, vol. 18, no. 3, pp. 181–186, Oct. 2015. DOI: 10.1111/jch.12702.

[47] R. Alzubi, N. Ramzan, H. Alzoubi, and S. Katsigiannis, «SNPs-based Hypertension Disease Detection via Machine Learning Techniques», in *2018 24th International Conference on Automation and Computing (ICAC)*, IEEE, Sep. 2018. DOI: 10.23919/iconac.2018.8748972.

[48] K.-L. Chien, H.-C. Hsu, T.-C. Su, W.-T. Chang, F.-C. Sung, M.-F. Chen, and Y.-T. Lee, «Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan», *Journal of Human Hypertension*, vol. 25, no. 5, pp. 294–303, Jul. 2010. DOI: 10.1038/jhh.2010.63.

[49] N.-K. Lim, K.-H. Son, K.-S. Lee, H.-Y. Park, and M.-C. Cho, «Predicting the Risk of Incident Hypertension in a Korean Middle-Aged Population: Korean Genome and Epidemiology Study», *The Journal of Clinical Hypertension*, vol. 15, no. 5, pp. 344–349, Mar. 2013. DOI: 10.1111/jch.12080.

[50] A. P. Carson, C. E. Lewis, D. R. Jacobs, C. A. Peralta, L. M. Steffen, J. K. Bower, S. D. Person, and P. Muntner, «Evaluating the Framingham Hypertension Risk Prediction Model in Young Adults», *Hypertension*, vol. 62, no. 6, pp. 1015–1020, Dec. 2013. DOI: 10.1161/hypertensionaha.113.01539.

[51] T. Sathish, S. Kannan, P. S. Sarma, O. Razum, A. G. Thrift, and K. R. Thankappan, «A Risk Score to Predict Hypertension in Primary Care Settings in Rural India», *Asia Pacific Journal of Public Health*, vol. 28, no. 1_suppl, 26S–31S, Sep. 2015. DOI: 10.1177/1010539515604701.

[52] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, «Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension», *IEEE Access*, vol. 7, pp. 144 777–144 789, 2019. DOI: 10.1109/access.2019.2945129.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, «Scikit-learn: Machine learning in Python», *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[54] M. M. McKerns, L. Strand, T. Sullivan, A. Fang, and M. A. G. Aivazis, «Building a Framework for Predictive Science», Feb. 6, 2012. arXiv: 1202.1056v1 [cs.MS].

[55] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/.

[56] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA, 2015. [Online]. Available: http://www.rstudio.com/.

[57] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN: 978-3-319-24277-4. [Online]. Available: https://ggplot2.tidyverse.org.

[58] J. Allaire, K. Ushey, Y. Tang, and D. Eddelbuettel, *reticulate: R Interface to Python*, 2017. [Online]. Available: https://github.com/rstudio/reticulate.

[59] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, «Welcome to the tidyverse», *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019. DOI: 10.21105/joss.01686.

[60] S. Huang, Y. Xu, L. Yue, S. Wei, L. Liu, X. Gan, S. Zhou, and S. Nie, «Evaluating the risk of hypertension using an artificial neural network method in rural residents over the age of 35 years in a Chinese area», *Hypertension Research*, vol. 33, no. 7, pp. 722–726, May 2010. DOI: 10.1038/hr.2010.73.

[61] S. Sakr, R. Elshawi, A. Ahmed, W. T. Qureshi, C. Brawner, S. Keteyian, M. J. Blaha, and M. H. Al-Mallah, «Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project», *PLOS ONE*, vol. 13, no. 4, X. Li, Ed., e0195344, Apr. 2018. DOI: 10.1371/journal.pone.0195344.

[62] H. Kanegae, K. Suzuki, K. Fukatani, T. Ito, N. Harada, and K. Kario, «Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques», *The Journal of Clinical Hypertension*, Dec. 2019. DOI: 10.1111/jch.13759.

[63] F. Xu, J. Zhu, N. Sun, L. Wang, C. Xie, Q. Tang, X. Mao, X. Fu, A. Brickell, Y. Hao, *et al.*, «Development and validation of prediction models for hypertension risks in rural Chinese populations», *Journal of Global Health*, vol. 9, no. 2, 2019.

[64] A. V. Kshirsagar, Y.-l. Chiu, A. S. Bomback, P. A. August, A. J. Viera, R. E. Colindres, and H. Bang, «A Hypertension Risk Score for Middle-Aged and Older Adults», *The Journal of Clinical Hypertension*, vol. 12, no. 10, pp. 800–808, Jul. 2010. DOI: 10.1111/j.1751-7176.2010.00343.x.

[65] Y. M. Chae, S. H. Ho, K. W. Cho, D. H. Lee, and S. H. Ji, «Data mining approach to policy analysis in a health insurance domain», *International Journal of Medical Informatics*, vol. 62, no. 2-3, pp. 103–111, Jul. 2001. DOI: 10.1016/s1386-5056(01)00154-x.

[66] M. Ture, I. KURT, A. TURHANKURUM, and K. OZDAMAR, «Comparing classification techniques for predicting essential hypertension», *Expert Systems with Applications*, vol. 29, no. 3, pp. 583–588, Oct. 2005. DOI: 10.1016/j.eswa.2005.04.014.

[67] B. Akdag, S. Fenkci, S. Degirmencioglu, S. Rota, Y. Sermez, and H. Camdeviren, «Determination of risk factors for hypertension through the classification tree method», *Advances in therapy*, vol. 23, no. 6, pp. 885–892, 2006.

[68] M. Kivimäki, G. D. Batty, A. Singh-Manoux, J. E. Ferrie, A. G. Tabak, M. Jokela, M. G. Marmot, G. D. Smith, and M. J. Shipley, «Validating the Framingham Hypertension Risk Score», *Hypertension*, vol. 54, no. 3, pp. 496–501, Sep. 2009. DOI: 10.1161/hypertensionaha.109.132373.

[69] N. P. Paynter, N. R. Cook, B. M. Everett, H. D. Sesso, J. E. Buring, and P. M. Ridker, «Prediction of Incident Hypertension Risk in Women with Currently Normal Blood Pressure», *The American Journal of Medicine*, vol. 122, no. 5, pp. 464–471, May 2009. DOI: 10.1016/j.amjmed.2008.10.034.

[70] R. Samant and S. Rao, «Evaluation of artificial neural networks in prediction of essential hypertension», *International Journal of Computer Applications*, vol. 81, no. 12, 2013.

[71] H. F. Golino, L. S. de Brito Amaral, S. F. P. Duarte, C. M. A. Gomes, T. de Jesus Soares, L. A. dos Reis, and J. Santos, «Predicting Increased Blood Pressure Using Machine Learning», *Journal of Obesity*, vol. 2014, pp. 1–12, 2014. DOI: 10.1155/2014/637635.

[72] D. LaFreniere, F. Zulkernine, D. Barber, and K. Martin, «Using machine learning to predict hypertension from a clinical dataset», in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Dec. 2016. DOI: 10.1109/ssci.2016.7849886.

[73] C. Ye, T. Fu, S. Hao, Y. Zhang, O. Wang, B. Jin, M. Xia, M. Liu, X. Zhou, Q. Wu, Y. Guo, C. Zhu, Y.-M. Li, D. S. Culver, S. T. Alfreds, F. Stearns, K. G. Sylvester, E. Widen, D. McElhinney, and X. Ling, «Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning», *Journal of Medical Internet Research*, vol. 20, no. 1, e22, Jan. 2018. DOI: 10.2196/jmir.9268.

[74] M. Du, S. Yin, P. Wang, X. Wang, J. Wu, M. Xue, H. Zheng, Y. Zhang, D. Liang, R. Wang, D. Liu, W. Shu, X. Xu, R. Hao, and S. Li, «Self-reported hypertension in Northern China: a cross-sectional study of a risk prediction model and age trends», *BMC Health Services Research*, vol. 18, no. 1, Jun. 2018. DOI: 10.1186/s12913-018-3279-3.

[75] R. Patnaik, M. Chandran, S.-C. Lee, A. Gupta, C. Kim, and C. Kim, «Predicting the occurrence of essential hypertension using annual health records», in *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, IEEE, Feb. 2018. DOI: 10.1109/icaecc.2018.8479458.

[76] Y. Kadomatsu, M. Tsukamoto, T. Sasakabe, S. Kawai, M. Naito, Y. Kubo, R. Okada, T. Tamura, A. Hishida, A. Mori, N. Hamajima, K. Yokoi, and K. Wakai, «A risk score predicting new incidence of hypertension in Japan», *Journal of Human Hypertension*, vol. 33, no. 10, pp. 748–755, Aug. 2019. DOI: 10.1038/s41371-019-0226-7.

Appendix

Table 1: Details on target value

Name	Variable	Levels	Comment
Hypertension status at HUNT3	hyp_NT3	No. Yes.	'No' means both systolic BP. < 140 mmHg, diastolic BP. < 90 mmHg in HUNT3, else 'Yes'.

Table 2: Details on continuous variables in the final dataset. All obtained from HUNT2.

Name	Variable	Unit	Comment
Features:			
Height	Hei	cm	
Weight	Wei	kg	
Body Mass Index	Bmi	kg/m^2.	Calculated as BMI = Wei / Hei^2.
Waist-circumference	WaistCirc	cm	Rounded to nearest cm.
Systolic blood pressure	BPSystMn23	mmHg	Mean of 2. and 3. measurements.
Diastolic blood pressure	BPDiasMn23	mmHg	Mean of 2. and 3. measurements.
Age at participation	PartAg	Years.	
Serum Creatinine	SeCreaCorr	μmol/L	Corrected to enzymatic method.
Serum Cholesterol	SeChol	mmol/L	
Serum HDL Cholesterol	SeHDLChol	mmol/L	
Serum Triglyceride	SeTrig	mmol/L	
Non-fasting serum Glucose	SeGluNonFast	mmol/L	

Table 3: Details on categorical variables in the final dataset.

Name	Variable	Levels	Comment
Gender of participant	Sex	Male. Female.	
Estimated glomerular filtration rate stage	GFREstStag	Stage 1. Stage 2.	- Stage 1: Rate < 90 ml/min. - Stage 2: Rate > 90 ml/min.
Heart-events in family	CarInfFam1	No. Yes.	Coded as 'Yes' if any parent or sibling has had a heart attack or chest pain, else 'No'
Family history hypertension	FamHypEv	No. Yes.	Coded as 'Yes' if any parent or sibling has had hypertension, else 'No'.
Physical Activity Indicator (PAI)	PAIlevel	Low. Medium. High	- Low: PAI < 50. - Medium: 49 < PAI < 100. - High: 99 < PAI.
Smoking status	SmoStat	Never. Formerly daily. Daily.	- Never: Never smoked. - Form. daily: Formerly daily smoker. - Daily: Daily smoker.
Municipality geography at invitation	InvMuniciGeo	Fjord. Inland. Coast.	
Marital status	LoveStat	Partner. No partner. Separated. Widow(er).	
Highest level of education	Educ	Secondary school. Upper sec. school. High-school. Higher level, < 4 years. Higher level, > 4 years.	

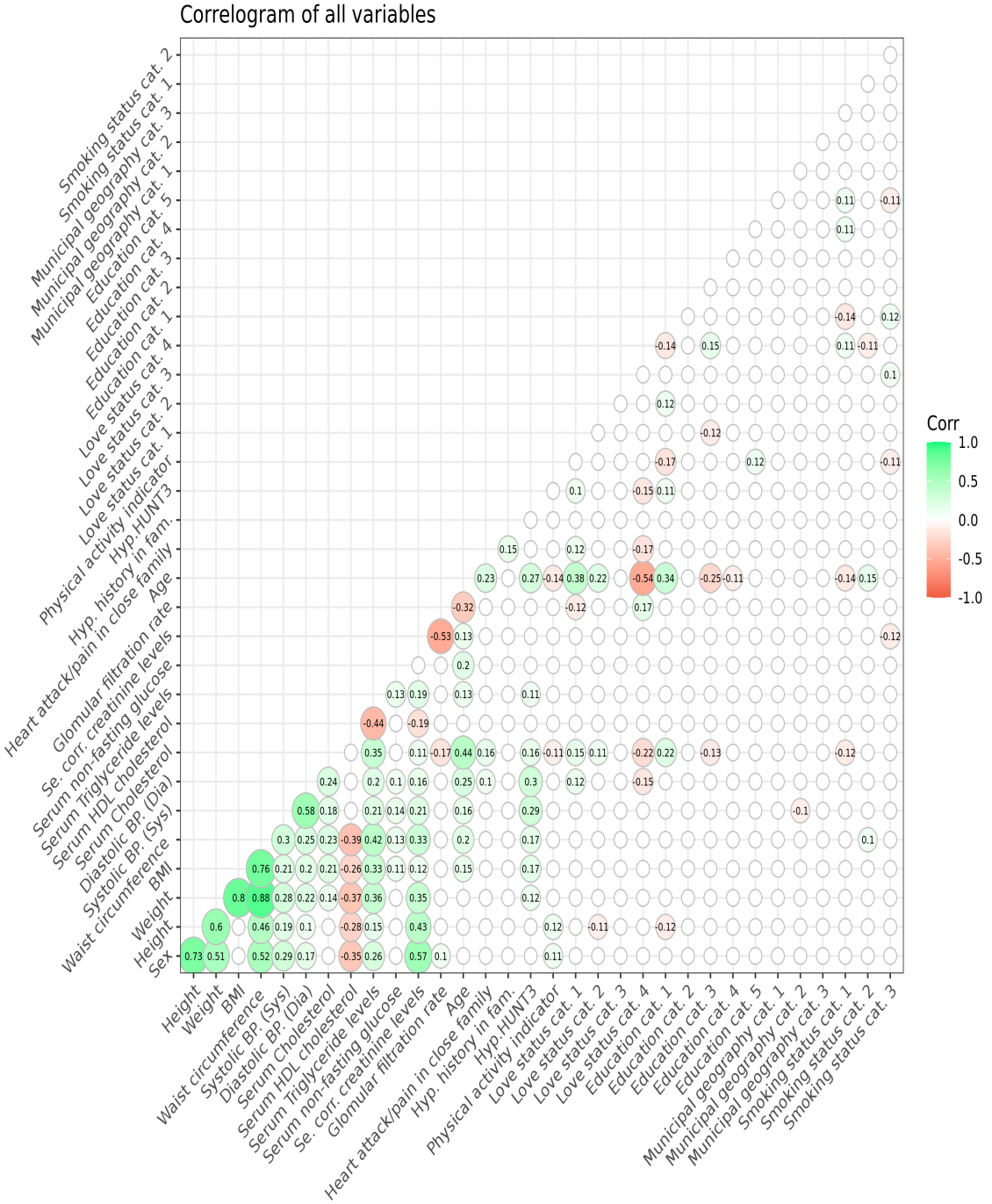


Figure 1: Correlogram for all variables in dataset. All correlations less than 0.1 in absolute value is omitted for easier interpretation. Note some correlations are not meaningful, e.g. correlation between levels in same category.

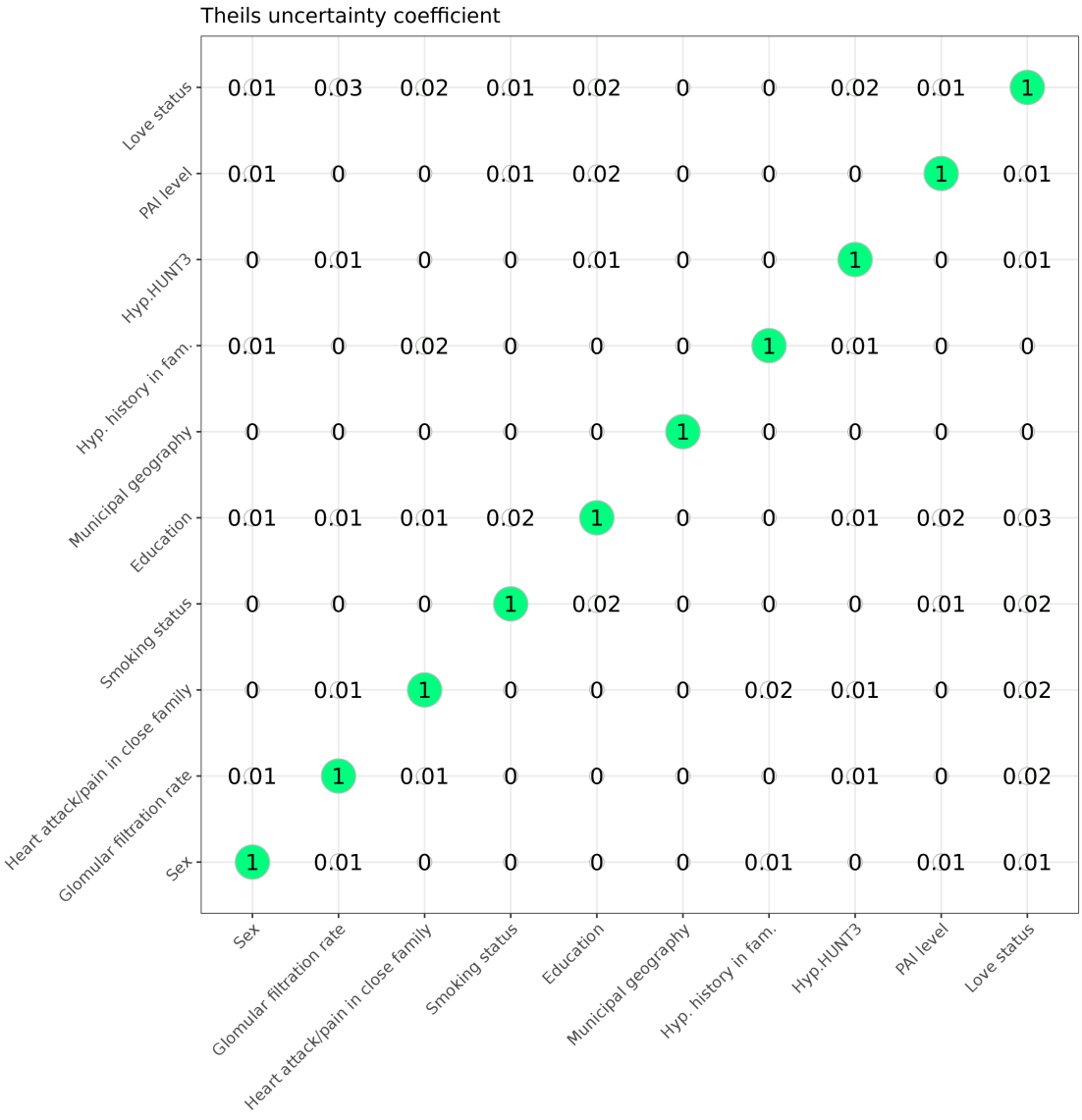


Figure 2: Theil's uncertainty coefficients between discrete variables. Score is ratio of bits for variable on X-axis explained by corresponding variable on Y-axis, in the range of $[0, 1]$.

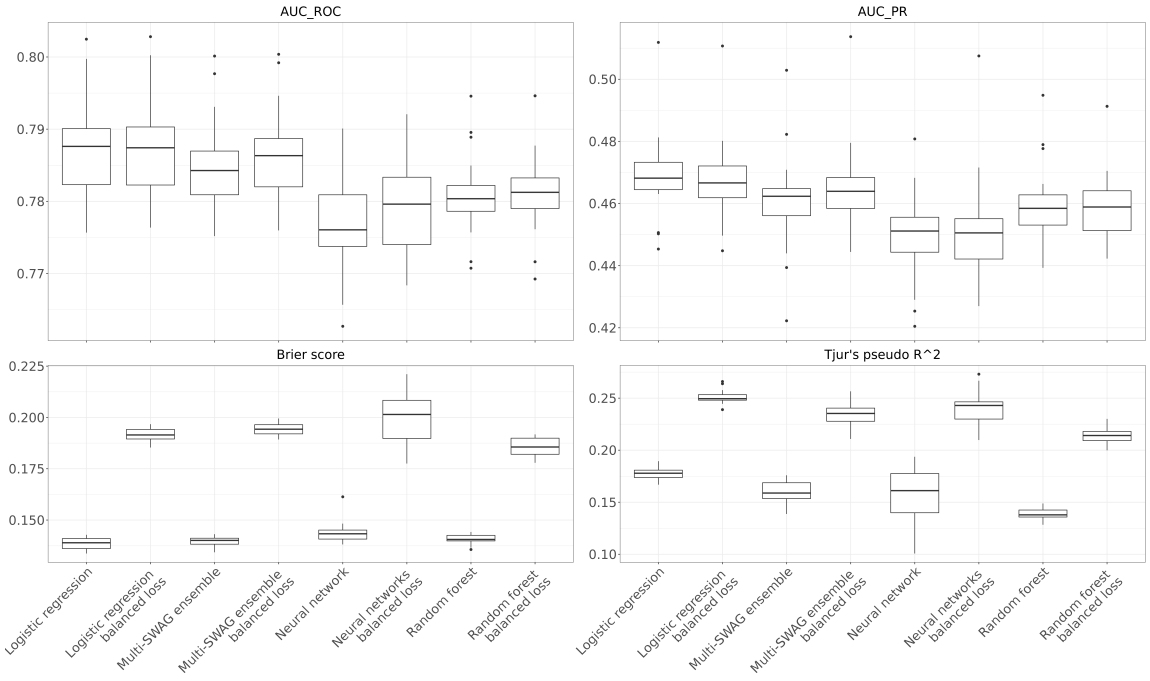


Figure 3: Results for models of the different model families, divided by usage of balanced loss in training, using the full feature set.

Table 4: Summary statistics for features coefficient sizes in the best performing logistic regression models. Calculated on the coefficients of the 20 logistic regression models fitted **without** balanced loss.

Feature	Full	Reduced	Mini
Systolic BP. (Sys)	0.5914 ± 0.0171	0.5413 ± 0.0175	0.5528 ± 0.0303
Diastolic BP. (Dia)	0.4276 ± 0.0145	0.4446 ± 0.0135	0.4705 ± 0.0214
Age	0.5042 ± 0.0145	0.5602 ± 0.011	0.5558 ± 0.0317
Hyp. history in fam.	0.3409 ± 0.0222	0.3593 ± 0.0209	
Waist circumference	0.1986 ± 0.0117	0.203 ± 0.0072	
Serum Cholesterol	0.0167 ± 0.0096		
Serum HDL cholesterol	-0.0547 ± 0.0143		
Serum Triglyceride levels	-0.0136 ± 0.0118		
Serum non-fasting glucose	0.0203 ± 0.008		
Se. corr. creatinine levels	-0.0233 ± 0.0161		
Glomular filtration rate	-0.0228 ± 0.0294		
Sex	-0.2721 ± 0.0263		
Heart attack/pain in close family	0.001 ± 0.0101		
Physical activity indicator	-0.037 ± 0.0154		
Love status cat. 2	0.0282 ± 0.0462		
Love status cat. 3	$5e\text{-}04 \pm 0.0087$		
Love status cat. 4	-0.2265 ± 0.024		
Education cat. 2	0.0484 ± 0.0223		
Education cat. 3	-0.003 ± 0.0118		
Education cat. 4	-0.039 ± 0.0315		
Education cat. 5	-0.0633 ± 0.0481		
Municipal geography cat. 2	0.4382 ± 0.0502		
Municipal geography cat. 3	0.2025 ± 0.0689		
Smoking status cat. 2	0.0104 ± 0.0134		
Smoking status cat. 3	0.0018 ± 0.0066		

Table 5: Summary statistics for features coefficient sizes in the best performing logistic regression models. Calculated on the coefficients of the 20 logistic regression models fitted **with** balanced loss.

Feature	Full	Reduced	Mini
Systolic BP. (Sys)	0.5964 ± 0.0159	0.5443 ± 0.0174	0.5515 ± 0.0228
Diastolic BP. (Dia)	0.4322 ± 0.0137	0.4511 ± 0.0126	0.474 ± 0.0178
Age	0.5455 ± 0.0162	0.6156 ± 0.0103	0.6058 ± 0.0269
Hyp. history in fam.	0.3683 ± 0.0229	0.3753 ± 0.0297	
Waist circumference	0.2097 ± 0.012	0.2139 ± 0.0079	
Serum Cholesterol	0.0288 ± 0.0088		
Serum HDL cholesterol	-0.0693 ± 0.0155		
Serum Triglyceride levels	-0.0163 ± 0.0109		
Serum non-fasting glucose	0.0258 ± 0.0077		
Se. corr. creatinine levels	-0.0252 ± 0.0178		
Glomular filtration rate	-0.0368 ± 0.0306		
Sex	-0.2675 ± 0.0311		
Heart attack/pain in close family	0.0028 ± 0.0093		
Physical activity indicator	-0.0451 ± 0.0157		
Love status cat. 2	0.0481 ± 0.0628		
Love status cat. 3	0.0063 ± 0.0179		
Love status cat. 4	-0.2356 ± 0.023		
Education cat. 2	0.035 ± 0.0206		
Education cat. 3	-0.0095 ± 0.0257		
Education cat. 4	-0.087 ± 0.0295		
Education cat. 5	-0.121 ± 0.0434		
Municipal geography cat. 2	0.4547 ± 0.0344		
Municipal geography cat. 3	0.2345 ± 0.0463		
Smoking status cat. 2	0.0053 ± 0.0136		
Smoking status cat. 3	0.006 ± 0.0141		

Table 6: Tuning parameters selected for logistic regression models. Splitted by usage of balanced loss and feature set used in modelfitting. Listed as: 'Regularisation type: (λ, γ)' per the notation in section 2.1.3.1.

	Full			Reduced			Mini		
Standard loss	Reg. type	α	γ	Reg. type	α	γ	Reg. type	α	γ
1	Elastic	6.3096	0.6	Elastic	25.1189	0	Elastic	63.0957	0
2	L1	6.3096	-	None	-	-	None	-	-
3	L1	15.8489	-	None	-	-	None	-	-
4	L1	15.8489	-	Elastic	6.3096	0.4	None	-	-
5	Elastic	15.8489	0	L2	39.8107	-	Elastic	158.4893	0
6	L1	6.3096	-	None	-	-	None	-	-
7	L1	6.3096	-	L1	1.5849	-	None	-	-
8	Elastic	6.3096	0.9	L2	25.1189	-	None	-	-
9	Elastic	10	0.7	Elastic	63.0957	0	None	-	-
10	L1	15.8489	-	None	-	-	L2	398.1072	-
11	L1	10	-	L2	25.1189	-	None	-	-
12	Elastic	6.3096	0.4	None	-	-	None	-	-
13	Elastic	6.3096	0.7	None	-	-	None	-	-
14	Elastic	10	0.8	L1	10	-	L1	6.3096	-
15	Elastic	10	0.8	None	-	-	None	-	-
16	Elastic	10	0.9	Elastic	25.1189	0	Elastic	39.8107	0
17	Elastic	10	0.7	None	-	-	Elastic	158.4893	0
18	Elastic	10	0.8	Elastic	6.3096	0.4	None	-	-
19	L1	10	-	None	-	-	Elastic	251.1886	0
20	L1	10	-	L2	25.1189	-	None	-	-
Balanced loss	Reg. type	α	γ	Reg. type	α	γ	Reg. type	α	γ
1	Elastic	6.3096	0.9	L2	25.1189	-	Elastic	63.0957	0
2	Elastic	10	0.7	Elastic	100	0	None	-	-
3	Elastic	15.8489	0.7	Elastic	25.1189	0.1	Elastic	100	0.1
4	L1	10	-	L1	15.8489	-	None	-	-
5	Elastic	15.8489	0	None	-	-	L2	100	-
6	Elastic	15.8489	0.5	None	-	-	Elastic	39.8107	0.2
7	Elastic	10	0.3	Elastic	39.8107	0	L2	398.1072	-
8	Elastic	10	0.8	Elastic	6.3096	0.4	Elastic	100	0.5
9	Elastic	10	0.9	Elastic	100	0	Elastic	158.4893	0.1
10	L2	15.8489	-	L2	39.8107	-	L2	251.1886	-
11	L1	15.8489	-	None	-	-	L1	63.0957	-
12	Elastic	10	0.7	None	-	-	None	-	-
13	Elastic	10	0.8	Elastic	10	0.7	Elastic	158.4893	0.3
14	Elastic	15.8489	0.9	Elastic	6.3096	0.8	None	-	-
15	Elastic	10	0.9	None	-	-	None	-	-
16	L1	10	-	L2	15.8489	-	None	-	-
17	Elastic	15.8489	0.8	None	-	-	None	-	-
18	L1	6.3096	-	Elastic	15.8489	0.6	None	-	-
19	Elastic	10	0.9	None	-	-	Elastic	100	0.1
20	Elastic	15.8489	0.8	L2	63.0957	-	Elastic	158.4893	0.1

Table 7: Pruning intensity α and the resulting average of terminal nodes per decision tree, for each random forest model. Splitted by usage of balanced loss and feature set used in modelfitting. Note that there was 100 trees per random forest model. 'Endnodes per tree' reported as mean $\pm$ std.

	Full		Reduced		Mini	
Standard loss	α	Endnodes per tree	α	Leaves per tree	α	Leaves per tree
1	$4{\cdot}10^{-4}$	51.4 ± 9.95	$5{\cdot}10^{-4}$	22.34 ± 4.65	$4{\cdot}10^{-4}$	25.51 ± 5.99
2	$4{\cdot}10^{-4}$	51.22 ± 12.08	$4{\cdot}10^{-4}$	35.19 ± 6.79	$5{\cdot}10^{-4}$	18.82 ± 4.12
3	$5{\cdot}10^{-4}$	29.58 ± 7.91	$4{\cdot}10^{-4}$	33.59 ± 7.68	$6{\cdot}10^{-4}$	15.96 ± 3.49
4	$4{\cdot}10^{-4}$	51.44 ± 10.84	$5{\cdot}10^{-4}$	23.95 ± 4.84	$5{\cdot}10^{-4}$	19.71 ± 4.65
5	$4{\cdot}10^{-4}$	50.01 ± 11.77	$4{\cdot}10^{-4}$	36.08 ± 7.21	$4{\cdot}10^{-4}$	25.43 ± 6.19
6	$3{\cdot}10^{-4}$	127.55 ± 16.36	$4{\cdot}10^{-4}$	35.95 ± 7.71	$6{\cdot}10^{-4}$	15.77 ± 3.37
7	$4{\cdot}10^{-4}$	51.52 ± 12.52	$4{\cdot}10^{-4}$	37.26 ± 7.17	$6{\cdot}10^{-4}$	16.89 ± 3.95
8	$6{\cdot}10^{-4}$	21.4 ± 5.79	$5{\cdot}10^{-4}$	22.75 ± 4.41	$4{\cdot}10^{-4}$	23.53 ± 5.43
9	$5{\cdot}10^{-4}$	31.13 ± 7.96	$4{\cdot}10^{-4}$	35.94 ± 7.46	$5{\cdot}10^{-4}$	19.24 ± 3.87
10	$5{\cdot}10^{-4}$	29.85 ± 7.97	$4{\cdot}10^{-4}$	33.49 ± 6.65	$4{\cdot}10^{-4}$	24.29 ± 5.49
11	$5{\cdot}10^{-4}$	28.6 ± 6.76	$5{\cdot}10^{-4}$	22.85 ± 4.79	$5{\cdot}10^{-4}$	18.33 ± 5.14
12	$4{\cdot}10^{-4}$	50.12 ± 10.99	$4{\cdot}10^{-4}$	35.2 ± 6.67	$4{\cdot}10^{-4}$	24.35 ± 5.96
13	$4{\cdot}10^{-4}$	50.83 ± 11.38	$4{\cdot}10^{-4}$	35.42 ± 6.49	$4{\cdot}10^{-4}$	25.15 ± 5.87
14	$4{\cdot}10^{-4}$	50.21 ± 11.08	$4{\cdot}10^{-4}$	34.49 ± 7.64	$4{\cdot}10^{-4}$	25.34 ± 5.74
15	$5{\cdot}10^{-4}$	29.86 ± 8.84	$4{\cdot}10^{-4}$	33.7 ± 5.69	$5{\cdot}10^{-4}$	18.71 ± 4.06
16	$4{\cdot}10^{-4}$	48.92 ± 11.1	$4{\cdot}10^{-4}$	33.46 ± 7.25	$4{\cdot}10^{-4}$	24.94 ± 5.56
17	$3{\cdot}10^{-4}$	123.63 ± 17.74	$5{\cdot}10^{-4}$	22.98 ± 5.11	$4{\cdot}10^{-4}$	24.35 ± 5.7
18	$5{\cdot}10^{-4}$	31.59 ± 8.25	$4{\cdot}10^{-4}$	36.88 ± 6.7	$5{\cdot}10^{-4}$	18.91 ± 4.37
19	$4{\cdot}10^{-4}$	50.02 ± 12.15	$5{\cdot}10^{-4}$	24.23 ± 4.8	$4{\cdot}10^{-4}$	25.74 ± 6.21
20	$5{\cdot}10^{-4}$	28.74 ± 6.33	$4{\cdot}10^{-4}$	35.74 ± 7.88	$6{\cdot}10^{-4}$	14.61 ± 3.27
Balanced loss	α	Endnodes per tree	α	Leaves per tree	α	Leaves per tree
1	$5{\cdot}10^{-4}$	94.45 ± 14.52	$7{\cdot}10^{-4}$	33.23 ± 6.11	$7{\cdot}10^{-4}$	26.35 ± 6.19
2	$4{\cdot}10^{-4}$	168.14 ± 17.18	$6{\cdot}10^{-4}$	41.96 ± 7.83	$7{\cdot}10^{-4}$	26.57 ± 6.25
3	$5{\cdot}10^{-4}$	92.46 ± 15.8	$8{\cdot}10^{-4}$	26.84 ± 5.65	$6{\cdot}10^{-4}$	32.37 ± 7.1
4	$6{\cdot}10^{-4}$	59.5 ± 11.58	$7{\cdot}10^{-4}$	33.08 ± 6.99	$8{\cdot}10^{-4}$	22.22 ± 4.51
5	$7{\cdot}10^{-4}$	44.57 ± 10.6	$7{\cdot}10^{-4}$	31.97 ± 5.75	$9{\cdot}10^{-4}$	19.76 ± 5.03
6	$6{\cdot}10^{-4}$	59.69 ± 11.34	$7{\cdot}10^{-4}$	33.64 ± 6.15	$8{\cdot}10^{-4}$	22.49 ± 4.42
7	$5{\cdot}10^{-4}$	92.5 ± 15.34	$6{\cdot}10^{-4}$	42.93 ± 8.79	$6{\cdot}10^{-4}$	30.98 ± 6.86
8	$6{\cdot}10^{-4}$	59.58 ± 11.66	$7{\cdot}10^{-4}$	33.52 ± 6.05	$7{\cdot}10^{-4}$	27.1 ± 5.69
9	$6{\cdot}10^{-4}$	61.08 ± 12.94	$6{\cdot}10^{-4}$	41.75 ± 6.79	$6{\cdot}10^{-4}$	31.09 ± 6.26
10	$5{\cdot}10^{-4}$	93.9 ± 15.35	$7{\cdot}10^{-4}$	31.44 ± 6.04	$7{\cdot}10^{-4}$	25.96 ± 5.5
11	$7{\cdot}10^{-4}$	41.79 ± 8.99	$8{\cdot}10^{-4}$	26.97 ± 5.61	$7{\cdot}10^{-4}$	25.25 ± 4.96
12	$5{\cdot}10^{-4}$	91.07 ± 16.8	$6{\cdot}10^{-4}$	41.92 ± 6.94	$11{\cdot}10^{-4}$	17.1 ± 3.31
13	$7{\cdot}10^{-4}$	43.52 ± 10.56	$7{\cdot}10^{-4}$	33.48 ± 6.24	$7{\cdot}10^{-4}$	26.11 ± 5.52
14	$4{\cdot}10^{-4}$	169.84 ± 19.39	$7{\cdot}10^{-4}$	32.1 ± 6.4	$9{\cdot}10^{-4}$	20.38 ± 4.35
15	$5{\cdot}10^{-4}$	92.64 ± 15.53	$6{\cdot}10^{-4}$	40.97 ± 6.93	$7{\cdot}10^{-4}$	25.06 ± 5.43
16	$6{\cdot}10^{-4}$	58.23 ± 11.96	$8{\cdot}10^{-4}$	26.82 ± 5.89	$9{\cdot}10^{-4}$	19.25 ± 4.44
17	$5{\cdot}10^{-4}$	94.05 ± 15.74	$10{\cdot}10^{-4}$	21.03 ± 4.23	$9{\cdot}10^{-4}$	20.12 ± 4.72
18	$5{\cdot}10^{-4}$	93.77 ± 14.47	$6{\cdot}10^{-4}$	42.45 ± 8.03	$8{\cdot}10^{-4}$	22.6 ± 4.44
19	$6{\cdot}10^{-4}$	61.75 ± 12.12	$7{\cdot}10^{-4}$	33.46 ± 4.98	$7{\cdot}10^{-4}$	27.09 ± 6.6
20	$8{\cdot}10^{-4}$	33.99 ± 8.18	$7{\cdot}10^{-4}$	32.89 ± 6.55	$7{\cdot}10^{-4}$	26.27 ± 6.08

Table 8: Neural network tuning parameters found via the Bayesian tuning parameter search. LR = Learning rate.

Standard loss	Width	Depth	Activation function	π_{Drop}	LR	LR decay per epoch
Iter. 1	8	5	'GELU'	0	0.001	None
Iter. 2	128	2	'Sigmoid'	0.2	0.02354	0.5 %
Iter. 3	8	2	'GELU'	0	0.00747	2.5 %
Iter. 4	128	5	'ReLU'	0	0.001	5 %
Iter. 5	128	5	'GELU'	0.4	0.001	None
Iter. 6	128	5	'Tanh'	0.5	0.001	5 %
Iter. 7	64	4	'GELU'	0	0.01306	0.5 %
Iter. 8	16	2	'Sigmoid'	0.4	0.01292	2.5 %
Iter. 9	128	5	'Tanh'	0	0.001	5 %
Iter. 10	32	1	'Sigmoid'	0.4	0.01519	5 %
Iter. 11	128	4	'Sigmoid'	0.4	0.00296	0.5 %
Iter. 12	128	5	'ReLU'	0	0.001	None
Iter. 13	32	3	'GELU'	0	0.01267	2.5 %
Iter. 14	128	1	'Sigmoid'	0	0.05277	5 %
Iter. 15	128	1	'Sigmoid'	0	0.02596	5 %
Iter. 16	128	5	'GELU'	0.2	0.001	5 %
Iter. 17	32	1	'Sigmoid'	0	0.03057	5 %
Iter. 18	128	1	'Sigmoid'	0	0.1	5 %
Iter. 19	16	1	'Sigmoid'	0.3	0.03256	None
Iter. 20	16	2	'GELU'	0	0.0079	0.5 %
Balanced loss	Width	Depth	Activation function	Dropout_{π}	LR	LR decay per epoch
Iter. 1	128	1	'Tanh'	0	0.001	5 %
Iter. 2	128	5	'ReLU'	0	0.001	None
Iter. 3	8	1	'Sigmoid'	0	0.04843	0.5 %
Iter. 4	128	3	'GELU'	0.9	0.00274	0.5 %
Iter. 5	128	5	'Tanh'	0.3	0.001	None
Iter. 6	16	3	'GELU'	0.3	0.01152	0.5 %
Iter. 7	32	5	'GELU'	0	0.00824	5 %
Iter. 8	32	4	'GELU'	0	0.00121	None
Iter. 9	64	2	'ReLU'	0.5	0.00403	0.5 %
Iter. 10	128	1	'ReLU'	0	0.001	5 %
Iter. 11	128	2	'Sigmoid'	0.5	0.02595	0.5 %
Iter. 12	128	3	'ReLU'	0	0.001	5 %
Iter. 13	128	3	'GELU'	0	0.001	5 %
Iter. 14	128	1	'GELU'	0	0.001	None
Iter. 15	16	4	'ReLU'	0	0.00852	2.5 %
Iter. 16	128	1	'GELU'	0	0.001	5 %
Iter. 17	128	3	'GELU'	0	0.001	5 %
Iter. 18	128	1	'ReLU'	0.5	0.001	None
Iter. 19	128	5	'Tanh'	0	0.001	5 %
Iter. 20	128	5	'GELU'	0	0.001	None

Table 9: Main properties of reviewed studies

		Data			Setup		
Authors, year	Country	Baseline age mean ± sd. (range)	Exclusion criteria	Participants / outcomes (datapoints)	HT definition	Study type. Follow up	Methods used
Chae et al. 2001 [65]	Korea	51.86	nr	18277 / 9103	ESC	Cross-sectional	LR, DTMs
Ture et al. 2005 [66]	Turkey	HT: 48.2 ± 8.6, NT: 46.5 ± 8.2	nr	694 / 452	ESC	Cross-sectional	LR, NNs, DTMs
Akdag et al. 2006 [67]	Turkey	HT: 54.47 ± 9.4, NT: 49.72 ± 7.7	nr	1761 / 678	ESC*	Cross-sectional	DT
Parikh et al. 2008 [8]	USA	42 ± 9.6 (20 - 69)	CVD, high serum creatinine values, diabetes	1717 / 796 (5814)	ESC	Prospective median 3.8 years	WR
Kivimäki et al. 2009 [68]	UK	44.6 ± 6.4 (35-68)	Diabetes, CVD	6704 / 2043 (13679)	ESC*	Prospective median 5.6 years	WR
Paynter et al. 2009 [69]	USA	median: 51 (45-inf)	Only women. Pre-HT.	14822 / 3003	ESC*	Prospective minimum 8 years	LR

Abbreviations:
HT: *Hypertension/hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **nr**: *Not reported*, **CVD**: *Cardiovascular-disease*, **ESC**: *In accordance with ESC/ESH guidelines [33]*, **ESC***: *ESC and/or HT medication usage*, **ESC****: *ESC* and/or any physician diagnosis ever*, **LR**: *Logistic regression*, **NN**: *Neural networks*, **DTM**: *Decision-tree methods*, **WR**: *Weibull or Cox regression*

Continued on next page

Table 9 : Main properties of reviewed studies – *Continued from previous page*

Authors, year	Country	Baseline age mean ± sd. (range)	Exclusion criteria	Participants / outcomes (datapoints)	HT definition	Study type. Follow up	Methods used
Muntner et al. 2010 [39]	USA, multiethnic	58.5 ± 9.7 (45 - 84)	Diabetes, CVD history, pregnant and more.	3013 / 849 (7619)	ESC*	Prospective 1) median 1.6 yrs 2) median 4.8 yrs	Not clear.
Chien et al. 2010 [48]	Taiwan	HT: 54.0 ± 11.7 NT: 51.5 ± 12.1, (35 - inf)	nr	2506 / 1029	ESC**	Prospective median 6.15 years	WR
Kshirsagar et al. 2010 [64]	USA	56 ± 9 (45-inf)	nr	11407 / 3795	ESC*	Prospective 3 years 6 years 9 years	LR
Huang et al. 2010 [60]	China	51.4 ± 12 (35-inf)	CVD, diabetes, chronic renal disease, history of long term illness.	3054 / 823	ESC**	Cross-sectional	LR, NN
Samant et al. 2013 [70]	India	nr (19-73)	nr	981 / nr	ESC	Cross-sectional	NN

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **nr**: *Not reported*, **CVD**: *Cardiovascular-disease*, **ESC**: *In accordance with ESC / ESH guidelines [33]*, **ESC***: *ESC and / or HT medication usage*, **ESC****: *ESC* and / or any physician diagnosis ever*, **LR**: *Logistic regression*, **NN**: *Neural networks*, **DTM**: *Decision-tree methods*, **WR**: *Weibull or Cox regression*

Continued on next page

Table 9 : Main properties of reviewed studies – *Continued from previous page*

Authors, year	Country	Baseline age mean ± sd. (range)	Exclusion criteria	Participants / outcomes (datapoints)	HT definition	Study type. Follow up	Methods used
Lim et al. 2013 [2]	Korea	50.48 ± 8.44 (40-69)	CVD, high serum creatinine levels	4747 / 819	ESC	Prospective 4 years	WR
Fava et al. 2013 [44]	Sweden	45.2 ± 7.4	nr	17688 / nr	ESC*	Prospective 23 years	LR
Carson et al. 2013 [50]	USA, multiethnic	24.9 ± 3.6 (18-30)	Diabetes	4388 / 1179 (15166)	ESC*	Prospective median 5 years	WR
Zheng et al. 2013 [42]	China	47.9 ± 10.2 (35-inf)	Diabetes, CVD history	24434 / 8675 (48868)	ESC*	Prospective 2 years, 4 years	WR
Golino et al. 2014 [71]	Brazil	23.14 ± 6.03 (16-63)	nr	400 / 142	Women: sys. BP. > 120. Men: sys. BP. > 140	Cross-sectional	LR, CART
Wang et al. 2015 [43]	USA	HT: 59.8 ± 13.4 NT: 49.3 ± 14.8 (18-inf)	nr	308711 / 108260	Self-reported	Cross-sectional	LR, NN

Abbreviations:
HT: *Hypertension/hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **nr**: *Not reported*, **CVD**: *Cardiovascular-disease*, **ESC**: *In accordance with ESC/ESH guidelines [33]*, **ESC***: *ESC and/or HT medication usage*, **ESC****: *ESC* and/or any physician diagnosis ever*, **LR**: *Logistic regression*, **NN**: *Neural networks*, **DTM**: *Decision-tree methods*, **WR**: *Weibull or Cox regression*

Continued on next page

Table 9 : Main properties of reviewed studies – *Continued from previous page*

Authors, year	Country	Baseline age mean ± sd. (range)	Exclusion criteria	Participants / outcomes (datapoints)	HT definition	Study type. Follow up	Methods used
Sathish et al. 2015 [51]	India	36.1 ± 13.7 (15-64)	Pregnancy	297 / 70	ESC*	Prospective 7.1 ± 0.2 years	LR
Niiranen et al. 2015 [46]	Finland	52.7 ± 14.8 (30-inf)	nr	CS: 5402/nr. Pros: 2045/nr	ESC*	Cross-sectional and Prospective 11 years	LR
Lu et al. 2015 [45]	China	48.56 ± 9.7 (35-74)	CVD	7724 / 2559	ESC*	Prospective 7.9 years	LR
LaFreniere et al. 2016 [72]	Canada	56.69 ± 14.65 (18-75)	nr	379027 / 185371	ESC	Cross-sectional	NN
Ramezankhani et al. 2016 [41]	Iran	38.6 (20-inf)	CVD, pregnancy	2763 men, 3442 women / 731 men, 736 women	ESC*	Prospective median 8.7	DTM
Ye et al. 2018 [73]	USA	nr (18-inf)	nr	1504437 / 152577	ESC*	Prospective 1 year	DTM

Abbreviations:
HT: *Hypertension/hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **nr**: *Not reported*, **CVD**: *Cardiovascular-disease*, **ESC**: *In accordance with ESC/ESH guidelines [33]*, **ESC***: *ESC and/or HT medication usage*, **ESC****: *ESC* and/or any physician diagnosis ever*, **LR**: *Logistic regression*, **NN**: *Neural networks*, **DTM**: *Decision-tree methods*, **WR**: *Weibull or Cox regression*

Continued on next page

Table 9 : Main properties of reviewed studies – *Continued from previous page*

Authors, year	Country	Baseline age mean ± sd. (range)	Exclusion criteria	Participants / outcomes (datapoints)	HT definition	Study type. Follow up	Methods used
Sakr et al. 2018 [61]	USA, multiethnic	49 ± 12 (17-95)	nr	23095 / 8090	nr	Prospective 10 years	NN, DTM
Du et al. 2018 [74]	China	HT: 59.65 ± 11.65 NT: 44.25 ± 15.76 (15-inf)	nr	13554 / 2571	Med. usage or past diagnosis	Cross-sectional	LR
Patnaik et al. 2018 [75]	Korea	nr	Only men.	7988 / 3994	Sys. BP > 130 or dia. BP > 90	Prospective 1 year	LR, DTM, NN, others
Kadomatsu et al. 2019 [76]	Japan	HT: 56.1 ± 8.01 NT: 50.9 ± 9.17 (15-inf)	CVD, cerebrovascular disease	3936 / 324	ESC*	Prospective median 5 years	LR
Alzubi et al. 2019 [47]	UK	nr	nr	3501 / 2001	nr	Cross-sectional	NN, others

Abbreviations:
HT: *Hypertension/hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **nr**: *Not reported*, **CVD**: *Cardiovascular-disease*, **ESC**: *In accordance with ESC/ESH guidelines [33]*, **ESC***: *ESC and/or HT medication usage*, **ESC****: *ESC* and/or any physician diagnosis ever*, **LR**: *Logistic regression*, **NN**: *Neural networks*, **DTM**: *Decision-tree methods*, **WR**: *Weibull or Cox regression*

Continued on next page

Table 9 : Main properties of reviewed studies – *Continued from previous page*

Authors, year	Country	Baseline age mean ± sd. (range)	Exclusion criteria	Participants / outcomes (datapoints)	HT definition	Study type. Follow up	Methods used
Fitriyani et al. 2019 [52]	Brazil	23.14 ± 6.03 (16-63)	nr	325 / 104	Women: sys. BP. > 120. Men: sys. BP. > 140	Cross-sectional	NN, LR, DTM, others
Kanegae et al. 2019 [62]	Japan	46.4 ± 12.1	nr	18258 / nr	ESC*	Prospective 1 year	LR, DTM
Xu et al. 2019 [63]	China	nr (35-74)	Disability, severe infectious disease, cancer, CVD, history of chronic kidney disease	8319 / nr	ESC**	Prospective 6 years	WR, NN, DTM, others

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **nr**: *Not reported*, **CVD**: *Cardiovascular-disease*, **ESC**: *In accordance with ESC / ESH guidelines [33]*, **ESC***: *ESC and / or HT medication usage*, **ESC****: *ESC* and / or any physician diagnosis ever*, **LR**: *Logistic regression*, **NN**: *Neural networks*, **DTM**: *Decision-tree methods*, **WR**: *Weibull or Cox regression*

Table 10: Performance of reported hypertension risk models

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Chae et al. 2001 [65]	LR CHAID C5.0	Internal split, 3:1 train:test	19	nr nr nr	nr nr nr	0.633 0.523 0.591	0.644 0.763 0.593	nr nr nr	nr nr nr	0.638 0.6406 0.5922
Ture et al. 2005 [66]	LR MLP RBF CHAID CART QUEST	Internal split, 3:1 train:test	9	nr nr nr nr nr nr	nr nr nr nr nr nr	0.641 0.718 0.667 0.7 0.7 0.6	0.905 0.905 0.952 0.882 0.824 0.902	nr nr nr nr nr nr	nr nr nr nr nr nr	0.778 0.815 0.815 0.831 0.789 0.817
Akdag et al. 2006 [67]	DT	nr	7	nr	nr	0.971	0.816	0.947	0.894	0.911
Parikh et al. 2008 [8]	WR, - Framingham model (FR)	Bootstrap simulations	8	0.788	HL: 4.35	nr	nr	nr	nr	nr

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

Authors, year	Method, modelname	Validation	Number of features	Measures						
				AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Kivïmaki et al. 2009 [68]	WR	Internal split, 3:2 train:test	8	0.804	HL: 14.3	nr	nr	nr	nr	nr
	FR validation	-	-	0.803	HL: 11.5	nr	nr	nr	nr	nr
Paynter et al. 2009 [69]	LR,	Internal split, 2:1 train:test								
	- model 1		2	0.676	HL: 8.8	nr	nr	nr	nr	nr
	- model 2		5	0.703	HL: 12.3	nr	nr	nr	nr	nr
	- model 3		6	0.705	HL: 20.7	nr	nr	nr	nr	nr
	- model 4		10	0.705	HL: 24.6	nr	nr	nr	nr	nr
Muntner et al. 2010 [39]	Unclear,	nr								
	- model 1		8	0.788	nr	nr	nr	nr	nr	nr
	- model 2		1	0.768	nr	nr	nr	nr	nr	nr
	- model 3		1	0.699	nr	nr	nr	nr	nr	nr

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

Authors, year	Method, modelname	Validation	Number of features	Measures						
				AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Chien et al. 2010 [48]	WR, - model 1	5 fold cross-validation	5	0.74	nr	nr	nr	nr	nr	nr
	- model 2	nr	8	0.735	nr	nr	nr	nr	nr	nr
	FR validation	-	-	0.709	nr	nr	nr	nr	nr	nr
Kshirsagar et al. 2010 [64]	LR, - 3 year model	Internal split, 2:1 train:test	10	0.742	nr	nr	nr	nr	nr	nr
	- 6 year model			0.75	nr	nr	nr	nr	nr	nr
	- 9 year model			0.791	nr	nr	nr	nr	nr	nr
	- Ever model			0.775	nr	nr	nr	nr	nr	nr
Huang et al. 2010 [60]	LR	Internal split, 4:1 train:test	10	0.732	nr	0.838	0.817	0.506	0.958	0.834
	NN			0.9	nr	0.979	0.701	0.952	0.848	0.877
Samant et al. 2013 [70]	NN	Multiple splits	13	nr	nr	nr	nr	nr	nr	0.9285

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Lim et al. 2013 [49]	WR, - model 1 - model 2 FR validation (recalibrated)	Internal split, 3:2 train:test -	8 1 -	0.791 0.707 0.79	HL: 4.17 nr HL 29.73	nr nr nr	nr nr nr	nr nr nr	nr nr nr	nr nr nr
Fava et al. 2013 [44]	LR, - model 1 - model 2	nr nr	15 34	0.662 0.664	nr nr	nr nr	nr nr	nr nr	nr nr	nr nr
Carson et al. 2013 [50]	WR, - model 1 - model 2 - FR mimick	nr nr nr	1 1 8	0.71 0.81 0.84	nr HL: 12.9 HL: 14.6	nr nr nr	nr nr nr	nr nr nr	nr nr nr	nr nr nr

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

Authors, year	Method, modelname	Validation	Number of features	Measures						
				AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Zheng et al. 2013 [42]	FR validation, - 2 year model - 4 year model	- -	- -	0.537 0.61	HL: 2287.7 HL: 8227.1	nr nr	nr nr	nr nr	nr nr	nr nr
Golino et al. 2014 [71]	CART, - model women - model men LR, - model women - model men	2 fold cross-validation nr nr	4 4 3 3	nr nr 0.566 0.68	nr nr nr nr	nr nr nr nr	nr nr nr nr	nr nr nr nr	nr nr nr nr	0.81 0.84 nr nr
Wang et al. 2015 [43]	LR NN	nr Internal split, 7:3 train:test	11 nr	0.74 0.77	nr nr	0.864 0.846	0.447 0.489	nr nr	nr nr	0.72 0.721

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Sathish et al. 2015 [51]	LR	nr	4	0.802	nr	0.652	0.786	0.411	0.908	nr
Niiranen et al. 2015 [46]	Cross-sectional									
	- model 1	nr	9	0.8	nr	nr	nr	nr	nr	nr
	- model 2	nr	10	0.803	nr	nr	nr	nr	nr	nr
	Prospective									
	- model 3	nr	10	0.731	nr	nr	nr	nr	nr	nr
	- model 4	nr	11	0.733	nr	nr	nr	nr	nr	nr
Lu et al. 2015 [45]	LR,									
	- model 1	nr	3	0.65	nr	nr	nr	nr	nr	nr
	- model 2	nr	4	0.655	nr	nr	nr	nr	nr	nr
	- model 3	nr	7	0.683	nr	nr	nr	nr	nr	nr
	- model 4	nr	8	0.687	nr	nr	nr	nr	nr	nr
	- model 5	nr	9	0.774	nr	nr	nr	nr	nr	nr
	- model 6	nr	10	0.777	nr	nr	nr	nr	nr	nr

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
LaFreniere et al. 2016 [72]	NN	Internal split, 70:15:15 train:val:test	11	nr	nr	0.829	0.816	0.82	0.825	0.823
Ramezankhani et al. 2016 [41]	CART,	Internal split, 7:3 train:test								
	- model all		20	0.78	BS: 0.14	0.73	0.72	0.45	0.89	nr
	- model men		20	0.73	BS: 0.17	0.64	0.7	0.44	0.84	nr
	- model women		13	0.81	BS: 0.12	0.81	0.68	0.48	0.91	nr
	QUEST,									
	- model all		20	0.73	BS: 0.15	0.73	0.69	045	0.88	nr
	- model men		20	0.7	BS: 0.18	0.64	0.65	0.44	0.85	nr
	- model women		13	0.79	BS: 0.12	0.79	0.71	0.45	0.91	nr
	C5.0,									
	- model all		20	0.77	BS: 0.14	0.69	0.69	0.43	0.88	nr
	- model men		20	0.72	BS: 0.17	0.67	0.65	0.44	0.83	nr
	- model women		13	0.81	BS: 0.12	0.78	0.69	0.44	0.91	nr
Ye et al. 2016 [73]	XGBoost	External data, 3:2 train:test	798	0.87	nr	nr	nr	nr	nr	nr

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Sakr et al. 2018 [61]	RF	10 fold cross-validation.	13	0.93	nr	0.917	0.7	0.817	nr	nr
		Internal split, 70:30 train:test.		0.88	nr	0.856	0.743	0.735	nr	nr
		Internal split, 80:20 train:test.		0.89	nr	0.862	0.75	0.73	nr	nr
	NN	10 fold cross-validation.		0.67	nr	0.88	0.301	0.574	nr	nr
		Internal split, 70:30 train:test.		0.72	nr	0.865	0.395	0.612	nr	nr
		Internal split, 80:20 train:test.		0.74	nr	0.884	0.4	0.652	nr	nr
Du et al. 2018 [74]	LR	nr	7	0.81	nr	0.642	0.832	0.352	nr	nr
Patnaik et al. 2018 [75]	LR	Unclear	Unclear	nr	nr	0.789	0.802	0.794	nr	0.798
	RF			nr	nr	0.837	0.755	0.8243	nr	0.788
	NN			nr	nr	0.7647	0.712	0.754	nr	0.733

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Kadomatsu et al. 2019 [76]	LR, - model 1 - model 2	Internal split, 3:2 train:test, resampled 100 times	10 11	0.826 0.83	HL: 12.2 HL: 7.85	nr nr	nr nr	nr nr	nr nr	nr nr
Alzubi et al. 2019 [47]	SVM NN KNN LDA NB Ensemble: SVM+KNN+NB	Stratified 5 fold cross-validation	417523	nr nr nr nr nr nr	nr nr nr nr nr nr	0.969 0.933 0.942 0.972 0.731 0.973	0.833 0.862 0.69 0.798 0.902 0.878	0.953 0.916 0.780 0.955 0.715 0.960	nr nr nr nr nr nr	0.911 0.903 0.834 0.895 0.804 0.932

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Fitriyani et al. 2019 [52]	Stacked classifier: SVM + NN + DT -> LR - model men - model women	10 fold cross-validation	6	0.87 0.76	nr nr	nr nr	0.849 0.818	0.936 0.756	nr nr	0.857 0.759
Kanegae et al. 2019 [62]	XGBoost LR Ensemble: LR + XGBoost	Internal split, 3:1 train:test	17	0.877 0.859 0.881	nr nr nr	nr nr nr	0.317 0.29 0.253	0.601 0.638 0.635	nr nr nr	nr nr nr

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Continued on next page

Table 10 : Reported model performance – *Continued from previous page*

				Measures						
Authors, year	Method, modelname	Validation	Number of features	AUC_{ROC}	Calibration	True negative rate	True positive rate	Positive predictive value	Negative predictive value	Accuracy
Xu et al. 2019 [63]	WR,	Internal split, 1.36:1 train:test								
	- model men		7	0.771	HL: 6.31	nr	nr	nr	nr	nr
	- model women		7	0.765	HL: 6.78	nr	nr	nr	nr	nr
	CART,									
	- model men		1	0.722	HL: 5.25	nr	nr	nr	nr	nr
	- model women		1	0.698	HL: 19.73	nr	nr	nr	nr	nr
	NN,									
	- model men		5	0.773	HL: 29.3	nr	nr	nr	nr	nr
	- model women		5	0.756	HL: 4.7	nr	nr	nr	nr	nr
	NB,									
	- model men		5	0.76	HL: 82.3	nr	nr	nr	nr	nr
	- model women		5	0.761	HL: 189.8	nr	nr	nr	nr	nr

Abbreviations:
FR: *Framingham model*, **nr**: *Not reported*,
LR: *Logistic regression*, **NN**: *Neural networks* **MLP, RBF**: *Types of NNs*
RF: *Random forest*, **WR**: *Weibull or Cox regression*, **DTM**: *Decision-tree method*
CHAID, QUEST, C5, CART, XGBoost: *DTMs*, **SVM, KNN**: *Other ML methods*

Table 11: Predictors of reported hypertension risk models

Authors	Modelname	Predictor selection method	Predictors used
Chae et al.	All models	nr	Age, Ever: {*Stroke, diabetes, HT.*}, Family ever: { *CVD, diabetes* }, BMI, urin content, blood sugar, total chol., exercise, tobacco, alcohol, diet info
Ture et al.	All models	nr	Age, sex, family HT history, smoking, lipoprotein, triglyceride, uric acid, total chol., BMI
Akdag et al	DT	Variable importance of DT	Sex, current HT in family, BMI, WHR, sex, triglyceride, total chol., dietary habits
Parikh	Framingham model (FR)	Stepwise by WR significance values.	Age, sex, sys. BP, dia. BP, smoking, current HT in family, BMI, age$\times$dia. BP
Kivimäki et al.	WR	FR mimick	Same as FR
Paynter et al.	Model 1 Model 2 Model 3 Model 4	Simplest Simpler Simple Bayes Info Crit.	sys. BP, dia. BP + age, BMI, race + cholesterol different protein levels, diet habits

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*,
Sys.: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*,
CVD: *Cardiovascular disease*, **HDL**: *High-density lipid*, **BMI**: *Body mass index*,
FR: *Framingham model*, **WHR**: *Waist-heigh ratio*

Continued on next page

Table 11 : Model predictors – *Continued from previous page*

Authors	Modelname	Predictor selection method	Predictors used
Muntner et al.	Model 1 Model 2 Model 3	FR mimick Simpler, medical custom Simpler, medical custom	Same as FR Sys. BP categorized Dia. BP categorized
Chien et al.	Model 1 Model 2	Stepwise signif. levels Expanded with biochemical features	Age, sex, BMI, sys. BP, dia. BP + white blood cell count, fasting glucose, uric acid
Kshirsagar et al.	All models	Stepwise by LR significance levels	Age, sex, sys. BP, dia. BP, smoke, family ever HT, BMI, diabetes, age×dia. BP, exercise
Huang et al.	All models	LR significance levels	Sedentary work, family ever HT, education, alcohol, diet habits, exercise, overweight, dysareriotony
Samant et al.	NN	nr	Age, pulse, sys. BP, dia. BP, blood protein levels, albumin, cholesterol, triglycerides, other blood measures
Lim et al.	Model 1 Model 2	FR mimick Simplified	Same as FR Prehypertension status

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*,
Sys.: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*,
CVD: *Cardiovascular disease*, **HDL**: *High-density lipid*, **BMI**: *Body mass index*,
FR: *Framingham model*, **WHR**: *Waist-heigh ratio*

Continued on next page

Table 11 : Model predictors – *Continued from previous page*

Authors	Modelname	Predictor selection method	Predictors used
Fava et al.	Model 1	LR significance levels	Sex, age, age^2, sex$\times$age, heart rate, obesity, diabetes, triglycerid, prehypertension, family ever HT, sedentary job, alcohol problem, relationship status, smoking, genetic risk score
	Model 2	Extended	+ individual SNPs
Carson et al.	Model 1	Simple	Prehypertension status
	Model 2	Simple	Age$\times$dia. BP
	FR mimick	FR mimick	Same as FR
Zheng et al.	All models	FR model validation	Same as FR
Golino et al.	CART models	Exhaustive search	BMI, Waist circumference, hip circumference, waist-hip ratio
	LR models	nr	BMI, waist circumference, hip circumference
Wang et al.	All models	LR significance values	Age, sex, exercise, diabetes, hyperlipemia, age, marriage, education, income, weights, height, drink
Sathish et al.	LR	Forward likelihood ratio method	Age > 35, smoking, prehypertension, obesity

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **CVD**: *Cardiovascular disease*, **HDL**: *High-density lipid*, **BMI**: *Body mass index*, **FR**: *Framingham model*, **WHR**: *Waist-heigh ratio*

Continued on next page

Table 11 : Model predictors – *Continued from previous page*

Authors	Modelname	Predictor selection method	Predictors used
Niiranen et al.	CS: Model 1 CS: Model 2 Pro: Model 3 Pro: Model 4	LR significance levels	Age, sex, smoking, diabetes, education, hypercholesterolemia, exercise, BMI + 38 SNP genetic risk score Model 1 + mean arterial pressure + 38 SNP genetic risk score
Lu et al.	Model 1 Model 2 Model 3 Model 4 Model 5 Model 6	Simplest Simplest extended Simple Simple, extended LR significance levels LR significance levels ext.	Age, sex, BMI + 22 SNP genetic risk score Model 1 + smoking, drinking, pulse, education + 22 SNP genetic risk score Model 3 + sys. BP, dia. BP + 22 SNP genetic risk score
LaFreniere	NN	nr	Age, gender, BMI, sys. BP, dia. BP, lipoproteins, triglycerides, cholesterol, microalbumin, urine albumin/creatinine ratio

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*,
Sys.: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*,
CVD: *Cardiovascular disease*, **HDL**: *High-density lipid*, **BMI**: *Body mass index*,
FR: *Framingham model*, **WHR**: *Waist-heigh ratio*

Continued on next page

Table 11 : Model predictors – *Continued from previous page*

Authors	Modelname	Predictor selection method	Predictors used
Ramezankhani et al.	Model all	Correlation-based, Consistency based	Age, duration of urban living, BMI, circumference of waist and wrist, glucose levels, triglyceride, total and HDL cholesterol, eGlomerular filtration rate, sys. BP, dia. BP, sex, education, CVD history in female family member, diabetes history in close family, former smoker, usage of blood aspirine,
	Model men		Age, duration of urban living, BMI, circumference of waist and hip, glucose levels, triglyceride, total cholesterol, eGlomuerular filtration rate, sys. BP, dia. BP, education, marital status, CVD history in female family member, diabetes history in close family, physical activity levels, smoking, participation in life-style intervention group
	Model women		Age, circumference of waist and wrist, glucose levels, triglyceride, HDL cholesterol, sys. BP, dia. BP, education, usage of blood glucose drugs, usage of pregnancy prevention, menstruation status

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **CVD**: *Cardiovascular disease*, **HDL**: *High-density lipid*, **BMI**: *Body mass index*, **FR**: *Framingham model*, **WHR**: *Waist-heigh ratio*

Continued on next page

Table 11 : Model predictors – *Continued from previous page*

Authors	Modelname	Predictor selection method	Predictors used
Ye et al.	XGBoost	Univariate correlation filtering vs target, Cochran.Armitage trend test and univariate LR p-values for cont values	Too many to list (798)
Sakr et al.	All models	Information gain ranking	Age, METS, Resting Systolic Blood Pressure, Peak Diastolic Blood Pressure, Resting Diastolic Blood Pressure, HX Coronary Artery Disease, Reason for test, History of Diabetes, Percentage HR achieved, Race, History of Hyperlipidemia, Aspirin Use, Hypertension response
Du et al.	LR	Univariable LR, backward selection using $p < 0.1$	Age, minority population, marital status, alcohol consumotion, sex, various comorbidities, high BMI
Patnaik et al.	All models	nr	Unclear. Age, sys. BP, dia. BP, BMI, family history of hypertension, income and geographic informatino are mentioned.
Kadomatsu	Model 1 Model 2	LR significance levels Extended	Age, sex, smoking, drinking, diabetes, BMI, sys. BP, dia. BP, parhyp, medicine use + biochemical marker

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*, **Sys.**: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*, **CVD**: *Cardiovascular disease*, **HDL**: *High-density lipid*, **BMI**: *Body mass index*, **FR**: *Framingham model*, **WHR**: *Waist-heigh ratio*

Continued on next page

Table 11 : Model predictors – *Continued from previous page*

Authors	Modelname	Predictor selection method	Predictors used
Alzubi et al.	All models	Conditional mutual info. minimization	417523 SNPS
Fitriyani et al.	All models	nr	Age, obese, bmi, wc, hc, whr
Kanegae et al.	All models	nr	Age, sex, BMI, sys. BP, dia. BP, CAVI sys. BP, CAVI dia. BP, high/low lipoprotein, uric acid, fasting glucose, diabetes, chronic kidney disease, smoking, drinking
Xu et al et al.	All models	Univariate cox models by significance	Age, sys. BP, dia. BP, waist-circumference (WC), history of hypertension in family, age×WC, age×dia. BP

Abbreviations:
HT: *Hypertension / hypertensives*, **NT**: *Normotensives*,
Sys.: *Systolic*, **Dia.**: *Diastolic*, **BP**: *Blood pressure*,
CVD: *Cardiovascular disease*, **HDL**: *High-density lipid*, **BMI**: *Body mass index*,
FR: *Framingham model*, **WHR**: *Waist-heigh ratio*