Predicting Huntington's Disease: Extreme Learning Machine with Missing Values

Emil Eirola[1⋆], Anton Akusok[1], Kaj-Mikael Björk[2], Hans Johnson[3], and Amaury Lendasse[4]

[1] Arcada University of Applied Sciences, Helsinki, Finland
[2] Risklab at Arcada University of Applied Sciences, Helsinki, Finland
[3] Department of Electrical Engineering, The University of Iowa, Iowa City, USA
[4] Department of Mechanical and Industrial Engineering and The Iowa Informatics Initiative, The University of Iowa, Iowa City, USA

Abstract. Problems with incomplete data and missing values are common and important in real-world machine learning scenarios, yet often underrepresented in the research field. Particularly data related to healthcare tends to feature missing values which must be handled properly, and ignoring any incomplete samples is not an acceptable solution. The Extreme Learning Machine has demonstrated excellent performance in a variety of machine learning tasks, including situations with missing values. In this paper, we present an application to predict the onset of Huntington's disease several years in advance based on data from MRI brain scans. Experimental results show that such prediction is indeed realistic with reasonable accuracy, provided the missing values are handled with care. In particular, Multiple Imputation ELM achieves exceptional prediction accuracy.

Keywords: Extreme learning machine, missing values, multiple imputation, Huntington's disease, prediction

1 Introduction

The prevalence of machine learning has been steadily increasing in the current information age. Engineering advances in processor performance and storage capacities have provided an opportunity to make practical use of computational statistics on a large scale. Simultaneously, the research community has contributed by devising new clever algorithms to maximize the amount of relevant information that can be extracted from data. While data sets are large at times, the more common situation is that the number of samples is limited by practical issues, meaning that all of the available data must be used as efficiently as possible in order to achieve the desired results.

One pertinent issue is incomplete data sets, where some samples have missing information [1]. Most methods in machine learning are based on the assumption

⋆ Corresponding author

that data is available as a fixed set of measurements for each sample. However, this is not always true in practice, as several samples may have incomplete records for any of a number of reasons. These could include measurement error, device malfunction, operator failure, non-response in a survey, etc. Simply discarding the samples or variables which have missing components often means throwing out a large part of data that could be useful for the model. It is relevant to look for better ways of dealing with missing values in such cases.

If the fraction of missing data is sufficiently small, a common pre-processing step is to perform imputation to fill in the missing values and proceed with conventional methods for further processing. Any errors introduced by inaccurate imputation may be considered insignificant in terms of the entire processing chain. With a larger proportion of measurements being missing, errors caused by the imputation are increasingly relevant as errors propagate in non-obvious ways, and the missing value imputation cannot be considered a separate step. Instead, the task should be seen from a holistic perspective, and the statistical properties of the missing data should be considered more carefully [2].

A particularly important area where incomplete data is commonplace is in healthcare, where varying procedures and equipment affect which data is available. Studies generally include a limited number of subjects, and often requires expensive equipment and highly trained professionals, meaning that discarding data samples with a few unknown values would not be cost-effective, and all the data must be used to its maximal potential.

Recently, significant results in machine learning have been achieved with methods based on the Extreme Learning Machine (ELM) [3]. Several modified approaches have been published with the goal of using datasets with missing values [4–8]. In this paper, we describe an application of ELM with multiple imputation [5] to predict the onset of Huntington's disease from early brain scans.

The structure of this paper is as follows: Section 2 describes the application scenario and data used. The modelling procedure is detailed in Section 3, and results are presented in Section 4.

2 Application: Predicting Onset of Huntington's Disease

Huntington's disease (HD) is an inherited condition caused by a genetic disorder. It affects muscle coordination and leads to mental decline and behavioral symptoms, and ultimately death. All patients with a sufficiently severe form of the disorder will eventually get the disease. Physical symptoms can begin at any age, but usually begin between 35 and 44 years of age. No cure is known, but therapy can considerably mitigate symptoms, especially if started at an early stage

While identifying the disease can be achieved early by testing for the genetic disorder, it is more difficult to predict how quickly symptoms will manifest, as the progression of the disease is not fully understood in detail [9–12]. It has been observed that subtle changes in brain structure can be identified several years

before diagnosis [9].The main research question here is to study how well MRI data allows to predict when symptoms will appear, up to 10 years in advance.

The data consists of a number of measurements related to the patients. Each sample corresponds to one session with a patient. For many patients, measurements (sessions) are available *before and after* they have been diagnosed with the disease, and this is crucial for studying the progression in detail. As most patients attended several sessions, it is important to consider that several samples are associated to the same patient. The sessions were planned to be conducted approximately once every two years, but in reality the data is available at very irregular intervals, differing for each patient.

There are a total of 3729 sessions and 1370 patients. There is a control group of 288 patients, which do not have the genetic disorder, and as such do not contract the disease.

The measurements (variables) consist of key metrics derived from an MRI scan, e.g., volume or length of specific structures. There are 561 variables in total. In addition, the data contains a target variable representing whether the physician diagnosed the patient with Huntington's disease or not.

The data has been collected at several different locations, by different people, on a variety of equipment. The varying procedures and equipment mean that many values are missing for a large number of patients. For each session, the number of available measurements varies from 95 to 561, and only 10% of sessions have no missing values. No measurement available for all sessions. Overall, 45% of values are missing in the data.

3 Model

The goal of the model is to predict the progression of the disease several years in advance. However, the data directly includes only the diagnosis at the time of the measurement session. Fortunately, the majority of the patients return for follow-up sessions, meaning that some information about the progression can be inferred.

Ideally, the output variable Y should contain the state of the patient up to 10 years in the future, but this information is not fully available. For example, consider the sample related to a visit in 2001, at which the patient does not show symptoms. Say the other available sessions for this patient are as follows:

- 2005 (not diagnosed)
- 2008 (diagnosed)
- 2009 (diagnosed)

By assuming the progression is monotonic (i.e., once diagnosed, a patient will never return to a non-diagnosed state), we can conclude that for the years 2001–2005 (0–4 years in the future) the patient should be considered as not diagnosed, and 2008→ (7+ years in the future) should be considered diagnosed. Information for 5–6 years is however still not available. We construct a *prediction trajectory*,

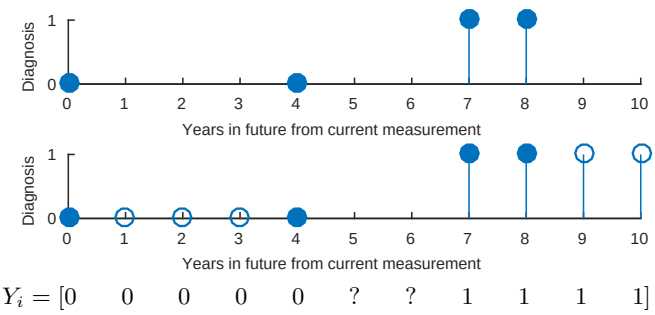

Fig. 1. Constructing the output as a prediction trajectory for a sample patient with data from three future sessions available. Filled points represent available data, and non-filled points inferred information. The state at 5 and 6 years remains unknown.

such that the output is 0 for years with no diagnosis, 1 for years with diagnosis, and missing values when the state is unknown (see Figure 1).

As such, the output is an 11-dimensional vector, and there are missing values in both input and output variables of the data. Note that several other types of particular situations occur in the data:

Diagnosed at current visit	$Y_i = [\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,]$
Infrequent visits	$Y_i = [\,0\,?\,?\,?\,?\,?\,?\,1\,1\,1\,1\,]$
Single visit, not diagnosed	$Y_i = [\,0\,?\,?\,?\,?\,?\,?\,?\,?\,?\,?\,]$
Not diagnosed after several visits	$Y_i = [\,0\,0\,0\,0\,0\,?\,?\,?\,?\,?\,?\,]$
Control group subject	$Y_i = [\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,]$

3.1 Extreme Learning Machine

The prediction model is realised using the Extreme Learning Machine (ELM) [3], which is a single hidden-layer feed-forward neural network where *only* the output weights β_k are optimised, and all the weights w_{kj} between the input and hidden layer are assigned randomly. With input vectors $\boldsymbol{x}_i$ and the targets collected as a vector $\boldsymbol{y}$, it can be written as

$$\mathbf{H}\boldsymbol{\beta} = \boldsymbol{y} \quad \text{where} \quad H_{ik} = h\big(\boldsymbol{w}_k^T \boldsymbol{x}_i\big)\,. \tag{1}$$

Here $h(\cdot)$ is a non-linear activation function applied elementwise. Training this model is simple, as the optimal output weights β_k can be calculated by ordinary least squares. The method relies on the idea of random projection: mapping the data randomly into a sufficiently high-dimensional space means that a linear model is likely to be relatively accurate. As such, the number of hidden-layer neurons needed for achieving equivalent accuracy is often much higher than in a multilayer perceptron trained by back-propagation, but the computational burden for training the model is still considerably lower.

The optimal weights $\boldsymbol{\beta}$ can be calculated as the least squares solution to Eq. (1), or formulated by using the Moore–Penrose pseudoinvorse as follows:

$$\boldsymbol{\beta} = \mathbf{H}^{+}\boldsymbol{y} \tag{2}$$

A high number of neurons in the hidden layer introduces concerns of overfitting, and regularised versions of the ELM have been developed to remedy this issue. These include the *optimally pruned ELM* (OP-ELM) [13], and its Tikhonov-regularised variant TROP-ELM [14]. In the current case, Tikhonov regularisation is applied when solving the least square problem in Eq. (1). The value of the regularisation parameter is selected by minimising the leave-one-out error (efficiently calculated via the PRESS statistic [14]) by a MATLAB minimisation procedure[5].

3.2 Multiple Imputation ELM for incomplete data

To handle the missing value problem, Multiple Imputation ELM (MI ELM) [5] is used. The method is based on the established procedure of multiple imputation [15], which is a principled approach to modelling incomplete data sets while avoiding any additional bias.

For ELM, the multiple imputation procedure is as follows. First generate a set of M imputations of the data $\boldsymbol{X}$, denote these as $\boldsymbol{X}_k$, for $1 \leq k \leq M$. The imputations should be randomly drawn from a distribution representing the data. In this case, we fit a Gaussian distribution to the data set by using the EM algorithm [16,17]. Having the distribution allows us to generate imputed versions of the data by drawing random samples from the conditional distribution of each missing value.

For each imputed version of the data, calculate the corresponding hidden layer representation $\mathbf{H}_k = h\big(\mathbf{W}^T\boldsymbol{X}_k\big)$, using the same set of hidden layer weights $\mathbf{W}$. Then solve for the output weights $\boldsymbol{\beta}_k = \mathbf{H}_k^{+}\boldsymbol{y}$.

When applying the model to a new set of (testing) data $\boldsymbol{X}_t$, in principle each trained ELM is used to generate a separate prediction $\hat{\boldsymbol{y}}_k = \boldsymbol{\beta}_k\mathbf{H}_t$, where $\mathbf{H}_t = h\big(\mathbf{W}^T\boldsymbol{X}_t\big)$, and these are then averaged to produce the final result:

$$\hat{\boldsymbol{y}} = \frac{1}{M}\sum_k \hat{\boldsymbol{y}}_k \tag{3}$$

However, the equivalent result can be obtained more efficiently by first averaging the weights as

$$\boldsymbol{\beta} = \frac{1}{M}\sum_k \boldsymbol{\beta}_k \tag{4}$$

and then applying the model

$$\hat{\boldsymbol{y}} = \boldsymbol{\beta}\mathbf{H}_t \tag{5}$$

[5] `fminsearch`: `https://www.mathworks.com/help/matlab/ref/fminsearch.html`

In particular, the average weight in Eq. (4) can be calculated during the training phase, without having access to the testing data. The trained model consists only of the random weight matrix $\mathbf{W}$ and the (average) output layer weight vector $\boldsymbol{\beta}$, just as in the conventional ELM. The multiple imputation approach is only used in the training phase to more accurately find $\boldsymbol{\beta}$ in the presence of missing values. The number of multiple imputations can be dynamically chosen in accordance with available resources, the guiding principle being that a larger number of imputations leads to a more accurate model.

If the data in the test set also have missing values, as in the current case, these can also be handled by multiple imputation. I.e., generate several imputed copies, calculate predictions for each copy, and average the results. Note that the multiple imputation procedure for training and testing can be conducted entirely separately from each other, and the number of imputations need not be the same.

3.3 Variable Selection

As the data is high dimensional, and contains redundant information, a variable selection procedure is applied to condense the problem. First, variables which are highly correlated (correlation coefficient with other variables above 0.99) are discarded. However, this only reduces the dimensionality from 561 to 483, and further reductions are needed.

While many methods for variable selection have been developed, only a few of them can be applied when the data contains missing values. One which is applicable is the Delta test [18, 19], which only requires identifying the nearest neighbor of each sample. This can be accomplished by first estimating distances with another method, filling in missing values and accounting for the uncertainty [2]. By again applying the previously calculated Gaussian distribution of the data, the conditional mean and variance can be calculated for each missing value. Replacing each missing value by its conditional mean produces an imputed version of the data, denoted by $\tilde{\boldsymbol{X}}$. Let the conditional variance for each missing sample be notated as $\sigma_{i,d}^2 = \mathrm{Var}(x_{i,d})$, with $\sigma_{i,d}^2 = 0$ if $x_{i,d}$ is known. Then the expected (squared) distance between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is

$$\mathrm{E}\left[\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right] = \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j\|^2 + \sum_d \sigma_{i,d}^2 + \sum_d \sigma_{j,d}^2 \tag{6}$$

These distances can then be used to identify nearest neighbours and calculate the Delta test. As the candidate space is too large for an exhaustive search, optimising the Delta test is done by applying genetic algorithms [20]. The end result is a set of 29 selected variables, and these are used for the remainder of the experiments.

3.4 Entire procedure

The complete training procedure can be summarised as follows:

1. Pre-processing
   (a) Standardise input variables to zero mean and unit variance
   (b) Discard too highly correlated variables
   (c) Construct outputs $\boldsymbol{y}_i$ for each sample (prediction trajectory)
2. Fit Gaussian distribution to the incomplete data set using the EM algorithm
3. Variable selection:
   (a) Generate imputed data with uncertainties, and calculate distances
   (b) Use Genetic Algorithm to select variables which minimise the Delta test
4. Multiple imputation ELM
   (a) Generate weights using a fixed value of 1000 neurons
   (b) Generate multiple imputed copies by drawing from the conditional Gaussian distribution
   (c) Select regularisation parameter by minimising the leave-one-out error
   (d) For each copy, train ELM using selected variables
   (e) Average the weights to get one model

4 Experiments

Five methods are compared in the accuracy of predicting the diagnosis 0–10 years ahead:

- ELM with multiple imputation
- ELM with missing values imputed with the conditional mean (using the Gaussian distribution)
- ELM using only samples for which all variables are known
- Support Vectors Machines (SVM) [21] using only samples for which all variables are known
- Nearest neighbor classifier (1-NN) with imputation

Each method evaluated on a test set of 30% of the patients, while the remaining data is used for training the models. The experiment is repeated 250 times to obtain more reliable measures of expected performance. Since the output also contains missing values, the training and testing for each prediction horizon is conducted only on those samples where the output is known.

4.1 Results

Overall classification accuracy for 0–10 years from the date of the session is presented in Figure 2. The imputation-based ELM variants both consistently achieve accuracies above 90%, whereas the other models have poorer performance. However, with unbalanced classes and different costs for false positives and false negatives, it is crucial to study precision and recall separately, and these are shown in Figure 3. Alternatively, models can be compared by their F-score, which gives a more balanced assessment of the performance than the

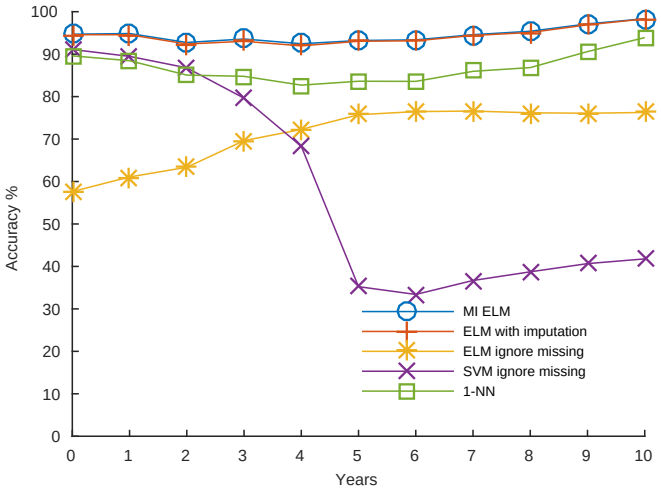

Fig. 2. Results in terms of average classification accuracy for each method for 0–10 years ahead.

overall classification accuracy [22]. The F-score, or F_1 measure, can be defined through the precision and recall as

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

The results as measured by the F-score are shown in Figure 4. The same values are presented in Table 1, along with standard deviations. A statistical significance analysis is also done to determine which differences in accuracy can be considered significant.

It can be seen that the Multiple Imputation ELM procedure gives the best results for 1–9 years ahead, and notably is significantly better than ELM with (single) imputation. For 0 and 10 years ahead, the accuracies between the two methods are not statistically distinguishable. In all cases, the two methods perform clearly better than the other three methods (1-NN, SVM, and ELM when ignoring samples with missing values).

5 Conclusions

In this paper we study how well variants of the Extreme Learning Machine can be used to predict the diagnosis of Huntington's disease from early MRI scans. The results clearly show that informative predictions are possible with satisfactory accuracy, and predicting onset of symptoms 10 years in advance is realistic.

The Extreme Learning Machine is able to model the scenario accurately. The best results are achieved by applying the principled multiple imputation

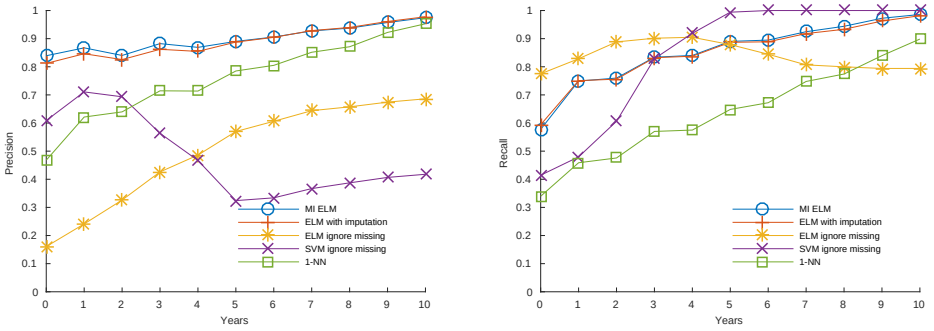

Fig. 3. Results in terms of average precision and recall for each method for 0–10 years ahead.

procedure. Indeed, properly accounting for the missing values is crucial for the machine learning task to perform reliably.

Further investigation is still required to more precisely analyse which variables (or combinations of variables) are the most informative in enabling the early prediction, and whether further refinements to the modelling procedure could lead to even more accurate predictions.

References

1. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Second edn. Wiley-Interscience (2002) doi: 10.1002/9781119013563.
2. Eirola, E., Doquire, G., Verleysen, M., Lendasse, A.: Distance estimation in numerical data sets with missing values. Information Sciences **240** (2013) 115–128 doi: 10.1016/j.ins.2013.03.043.
3. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: Theory and applications. Neurocomputing **70**(1–3) (2006) 489–501 doi: 10.1016/j.neucom.2005.12.126.
4. Yu, Q., Miche, Y., Eirola, E., van Heeswijk, M., Séverin, E., Lendasse, A.: Regularized extreme learning machine for regression with missing data. Neurocomputing **102** (2013) 45–51 doi: 10.1016/j.neucom.2012.02.040.
5. Sovilj, D., Eirola, E., Miche, Y., Björk, K., Nian, R., Akusok, A., Lendasse, A.: Extreme learning machine for missing data using multiple imputations. Neurocomputing **174, Part A** (2016) 220–231 doi: 10.1016/j.neucom.2015.03.108.
6. Gao, H., Liu, X.W., Peng, Y.X., Jian, S.L.: Sample-based extreme learning machine with missing data. Mathematical Problems in Engineering **2015** (2015) doi: 10.1155/2015/145156.
7. Xie, P., Liu, X., Yin, J., Wang, Y.: Absent extreme learning machine algorithm with application to packed executable identification. Neural Computing and Applications **27**(1) (2016) 93–100 doi: 10.1007/s00521-014-1558-4.
8. Yan, Y.T., Zhang, Y.P., Chen, J., Zhang, Y.W.: Incomplete data classification with voting based extreme learning machine. Neurocomputing **193** (2016) 167–175 doi: 10.1016/j.neucom.2016.01.068.

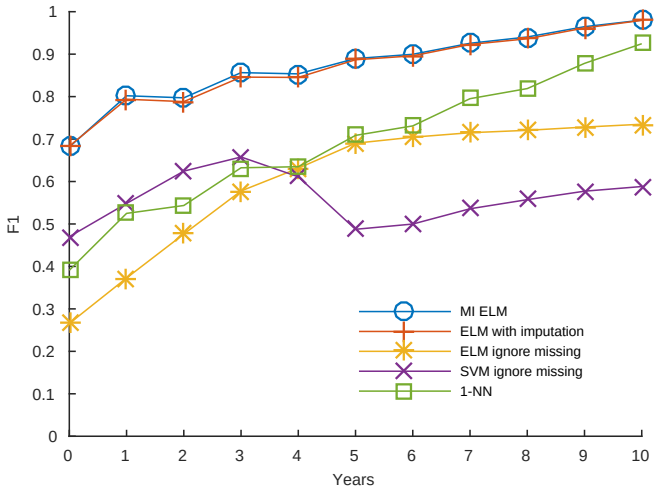

Fig. 4. Results in terms of average F-score for each method for 0–10 years ahead.

9. Paulsen, J.S., Langbehn, D.R., Stout, J.C., Aylward, E., Ross, C.A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L.J., Duff, K., Kayson, E., Biglan, K., Shoulson, I., Oakes, D., Hayden, M.: Detection of Huntington's disease decades before diagnosis: the Predict-HD study. Journal of Neurology, Neurosurgery & Psychiatry **79**(8) (2008) 874–880 doi: 10.1136/jnnp.2007.128728.
10. Paulsen, J.S., Long, J.D., Ross, C.A., Harrington, D.L., Erwin, C.J., Williams, J.K., Westervelt, H.J., Johnson, H.J., Aylward, E.H., Zhang, Y., et al.: Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. The Lancet Neurology **13**(12) (2014) 1193–1201 doi: 10.1016/S1474-4422(14)70238-8.
11. Matsui, J.T., Vaidya, J.G., Wassermann, D., Kim, R.E., Magnotta, V.A., Johnson, H.J., Paulsen, J.S.: Prefrontal cortex white matter tracts in prodromal Huntington disease. Human brain mapping **36**(10) (2015) 3717–3732 doi: 10.1002/hbm.22835.
12. Sturrock, A., Laule, C., Wyper, K., Milner, R.A., Decolongon, J., Santos, R.D., Coleman, A.J., Carter, K., Creighton, S., Bechtel, N., et al.: A longitudinal study of magnetic resonance spectroscopy Huntington's disease biomarkers. Movement Disorders **30**(3) (2015) 393–401 doi: 10.1002/mds.26118.
13. Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., Lendasse, A.: OP-ELM: Optimally-pruned extreme learning machine. IEEE Transactions on Neural Networks **21**(1) (2010) 158–162 doi: 10.1109/TNN.2009.2036259.
14. Miche, Y., van Heeswijk, M., Bas, P., Simula, O., Lendasse, A.: TROP-ELM: a double-regularized ELM using LARS and Tikhonov regularization. Neurocomputing **74**(16) (2011) 2413–2421 doi: 10.1016/j.neucom.2010.12.042.
15. Rubin, D.B.: Multiple imputation for nonresponse in surveys. John Wiley & Sons (1987)
16. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39**(1)

Table 1. The mean and standard deviation of the F-score for each method for 0–10 years ahead. The best result for each prediction horizon is in bold font, as well as any values not statistically significant (in a paired t-test at significance level 0.05).

Years	MI ELM	ELM with imputation	ELM ignore missing	SVM ignore missing	1-NN with imputation
0	**0.6814** ±0.0455	**0.6834** ±0.0462	0.2656 ±0.0297	0.4701 ±0.0815	0.3909 ±0.0474
1	**0.8020** ±0.0352	0.7938 ±0.0354	0.3728 ±0.0365	0.5483 ±0.1012	0.5249 ±0.0478
2	**0.7971** ±0.0275	0.7882 ±0.0262	0.4762 ±0.0363	0.6245 ±0.0770	0.5435 ±0.0420
3	**0.8565** ±0.0226	0.8459 ±0.0241	0.5778 ±0.0387	0.6577 ±0.0618	0.6323 ±0.0398
4	**0.8536** ±0.0243	0.8452 ±0.0237	0.6312 ±0.0408	0.6112 ±0.0727	0.6349 ±0.0405
5	**0.8896** ±0.0231	0.8868 ±0.0225	0.6902 ±0.0416	0.4875 ±0.0438	0.7087 ±0.0354
6	**0.8997** ±0.0209	0.8960 ±0.0217	0.7044 ±0.0390	0.4998 ±0.0372	0.7310 ±0.0340
7	**0.9256** ±0.0179	0.9228 ±0.0182	0.7149 ±0.0479	0.5358 ±0.0385	0.7951 ±0.0289
8	**0.9404** ±0.0160	0.9362 ±0.0160	0.7206 ±0.0489	0.5573 ±0.0396	0.8189 ±0.0272
9	**0.9647** ±0.0110	0.9616 ±0.0122	0.7282 ±0.0503	0.5772 ±0.0363	0.8783 ±0.0233
10	**0.9801** ±0.0075	**0.9796** ±0.0085	0.7349 ±0.0493	0.5882 ±0.0405	0.9246 ±0.0163

(1977) 1–38
17. Eirola, E., Lendasse, A., Vandewalle, V., Biernacki, C.: Mixture of gaussians for distance estimation with missing data. Neurocomputing **131** (2014) 32–42 doi: 10.1016/j.neucom.2013.07.050.
18. Eirola, E., Liitiäinen, E., Lendasse, A., Corona, F., Verleysen, M.: Using the Delta test for variable selection. In: Proceedings of ESANN 2008, European Symposium on Artificial Neural Networks, Bruges (Belgium). (2008) 25–30
19. Eirola, E., Lendasse, A., Corona, F., Verleysen, M.: The Delta test: The 1-NN estimator as a feature selection criterion. In: 2014 International Joint Conference on Neural Networks (IJCNN), IEEE (2014) 4214–4222 doi: 10.1109/IJCNN.2014.6889560.
20. Sovilj, D.: Multistart strategy using delta test for variable selection. In: International Conference on Artificial Neural Networks, Springer Berlin Heidelberg (2011) 413–420 doi: 10.1007/978-3-642-21738-8_53.
21. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20** (1995) 273–297 doi: 10.1023/A:1022627411411.
22. Rijsbergen, C.J.V.: Information Retrieval. 2nd edn. Butterworth-Heinemann (1979)