DRAFT (AUGUST 13, 2014)

# PREDICTING GLUCOSE CONCENTRATION IN TYPE 1 DIABETES PATIENTS USING ARTIFICIAL NEURAL NETWORKS

NIELS STEENBERGEN (3689700)
SUPERVISORS: VINCENT VAN OOSTROM, GERARD VREESWIJK
UTRECHT UNIVERSITY
BSC THESIS (ECTS: 7,5)

# CONTENTS

## ABSTRACT

Focus: Artificial neural networks may have the potential to predict blood glucose concentration in type 1 diabetes patients who are on an intensive insulin regimen, on the basis of relevant data that a patient can provide themselves.

Methods: A feed-forward back-propagation ANN was constructed and trained with 308 records from a single patient, containing data about time, previous blood glucose values, carbohydrate ingestion, insulin intake, stress and physical activity. It was run multiple times to measure average performance for some feasible parameter choices.

Findings: An ANN containing a single layer of 8 neurons reported a root-mean-square error ($\pm$ standard deviation) of $rmse = 2.156 \pm 0.131$ mmol/l, a mean absolute error percentage ($\pm$ SD) of $mae = 19.4\% \pm 1.3$ pp, and a correlation coefficient ($\pm$ SD) of $r = 0.662 \pm 0.042$. However, the data set does not generalise well.

Conclusion: While limited predictive success has been achieved, much work remains to be done before it can find practical application.

# 1

# INTRODUCTION

Type 1 diabetes mellitus is an auto-immune disease in which the the $\beta$-islet cells in the pancreas cease to produce insulin, a hormone that allows glucose to pass from the bloodstream into the cells. In insulin-deficient or insulin-resistant patients, glucose from carbohydrates lingers in the blood. This is problematic, since it is vital to maintain normal glucose levels to prevent serious complications (see section 2.1.1).

To this end, diabetes patients self-administer insulin. When and how much they do so depends on a great many variables, including the amount and type of carbohydrate ingestion, previous blood glucose readings and the prospect of phsysical effort (see section 2.1.2).

Carbohydrate ingestion and insulin injection must be timed so that glucose and appropriate quantities of insulin are absorbed more or less simultaneously into the bloodstream and so, in a sense, patients are required to emulate the pancreas manually. However, the human body is an intricate ecosystem. It presents physical difficulties through inaccurate blood glucose readings and irregular metabolism; practical difficulties through the sheer volume of data and the disciplin required for continuous control; and theoretical difficulties, since no simple set of rules can hope to capture the full breadth of its complexity. In short, no patient can be expected to carry out this function perfectly. They need to make do with rules of thumb, crude estimations, experience and intuition.

Since patients base decisions on vague but quantifiable information, it seems reasonable to suppose that a model could be constructed which predicts the glycaemic outcome based on the same considerations. While achieving very high accuracy in this way seems infeasible, achieving humanlike accuracy does not.

Many decision support systems are conceivable in which such a model may be useful. On its own, it may be used to warn against impending hypo- and hyperglycaemic episodes, where its merit over the patient's own expectations may be found in convenience, relative precision, consistency, and the potential ability to track and factor in more variables. When coupled with additional analysis, light may be shed on carbohydrate/insulin/blood glucose relationships, on basis of which insulin therapy may be fine-tuned. Finally, it may have educational value to the newly diagnosed, by providing retrospective suggestions to the adjustment of the timing and dosage of meals and insulin injections.

It has been shown that smartphone tracking applications already have the potential for decreasing average blood glucose. [22][13] Since the described system would be low-cost, low-risk, and non-invasive, it could provide permanent guidance by inclusion next to such existing tracking functionality, or perhaps by embedding it in a blood glucose monitoring device.

## 1.1 RESEARCH QUESTION

Specifically, the question is the following: given values for parameters on which is to be decided later in this report, can the course of blood glucose level be projected to some time in the future with acceptable accuracy?

Here, the intention is to stay within the limits of practical applicability in settings outside the hospital — that is, to achieve sufficient accuracy through relatively unobtrusive data recording processes. Efforts shall further be limited to diabetes type 1 patients on an intensive insulin regimen, who self-monitor blood glucose, are familiar with or able to learn about carbohydrate counting, and use a syringe or insulin pen to inject basal and meal insulin seperately.[1]

## 1.2 RELATIONSHIP TO AI

The nature of this problem is dynamic, non-linear and, for all practical purposes, black-boxed. The technique that is used to solve it should furthermore be able to handle multidimensional, noisy input data, and accommodate for the continuous changes to which the metabolism is subject. These characteristics make application of techniques from the field of artificial intelligence particularly appropriate.

The approach that is taken here requires a model to learn, like the human patient, to associate actions with their influence on blood glucose. Since the system is to possess no prior knowledge of the glucose-insulin system, this 'knowledge' (or a subsymbolic representation of something like knowledge) must be derived from the examples that are presented to it. The difficulty here lies not in inferring the rules that lead to optimal predictions, but in developing a system that infers these rules. Therefore, the system will perform a task that would (and does) require intelligence from humans.

1 Therefore, continuous rapid-acting insulin delivery using a pump and catheter will not be taken into consideration. Neither will the additional information that continuous glucose monitors can provide. However, it is conceivable that results are later generalised to apply to other patients and approaches.

## 1.3 OVERVIEW

To get a solid grasp on the problem, it is first explored in its medical context (section 2.1). The approach and performance of other systems (2.2) and the predictive performance of human patients (**??**) are then examined to set the bar on the possibilities that were mentioned in the introduction.

Attention will then shift to the construction of a preliminary model (3). The selected ANN type is a single, simple back-propagation feed-forward neural network, for which parameters are varied (3.1). The problem is simplified, abstracting from the relevant factors mentioned in 2.1, in order to get a workable case (3.2.1). Since time constraints prevent the collection of new data for this case, a pre-existing data set will be used (3.2.2) that is transformed to meet the criteria of this simplified problem (3.2.3).

Experiments are conducted on this data set for various parameter settings (4), and evaluated as discussed in section 3.3.

Finally, these evaluations are interpreted and compared to the performance of existing systems, as mentioned in section 2.2 (5). The issues that have surfaced are addressed, and a conclusion is drawn (6).

### ACKNOWLEDGEMENTS

I am indebted to Vincent van Oostrom and Gerard Vreeswijk for their supervision and valuable advice, and to Tom van der Weide for his help setting up. Gratitude is also expressed to Peter Kok from Delft University of Technology for being so kind as to provide the data set he originally compiled for his research, and to the UCI Machine Learning Repository [3] for hosting a sizable secondary data set.

# 2

# CONTEXT

## 2.1 MEDICAL

This report does neither aim to provide a thorough or rigorous discussion of diabetes, nor to accurately model its biological features. Emphasis is placed instead on the theoretical problem of processing noisy data, and on the practical benefits that such an approach may have for patients.

However, since no medical knowledge is presumed, it is necessary to contextualise the problem somewhat in order to estabilish some feasible input values for a predictive model, to suggest a reasonable method of collecting the corresponding data, and to properly interpret results. This section will therefore discuss the effects of diabetes and the consequent desirability of maintaining euglycaemia. Factors that influence blood glucose level and factors that influence the accuracy of their evaluation will then be considered. Finally, patient ability to assess (future) glycaemia is examined, so system performance can be compared to it.

It is worth noting that diabetes type 1, which is the focus of this survey, is a chronic disease characterised by insulin deficiency, while diabetes type 2 is a non-chronic disease characterised by insulin resistance. For the purposes of the following section, only research articles specifically dealing with type 1 diabetes have been examined; however, it is expected that if a self-learning system can derive the glycaemic profile of patients of one type, it will not be fundamentally unhelpful to the other.

### 2.1.1 *Target glycaemic condition*

It has been shown that reducing high blood glucose levels (hyperglycaemia) in type 1 diabetes significantly reduces the risk and slows the onset of the long-term complications of diabetes, such as eye, foot, kidney and brain damage [43][15][40] and eliminates short-term effects, such as decrements in sensory and cognitive processing. [34]

However, the solution of simply administering enough insulin to decrease glucose levels as much as possible is complicated by the potential adverse effects of low blood glucose level (hypoglycaemia). In addition to acute discomfort through trembling, hunger, sweating, confusion, fatigue, behavioral change or loss of consciousness, hypoglycaemia can be quite dangerous in its severe and prolonged instance, causing permanent nerve damage and possibly death. [34][6]

So, glycaemic levels are to be reduced whilst avoiding hypoglycaemia, approaching the euglycaemic (non-diabetic) range as much as possible, which is approximately 3.9–6.1 mmol/l. [6][15] The acceptable lower bound varies from 3.0 mmol/l in adults to 4.2 mmol/l in children. [34] It is less urgent to define a strict boundary beyond which every value is unacceptably high; however, since there is an exponential relationship between level of blood glucose above the desirable range and risk of complications, [15] straying too far from the euglycaemic upper bound of 6.1–6.7 mmol/l is unsatisfactory. Also, since severe hypoglycaemia has a tendency to occur during sleep, extra care may be taken to avoid nocturnal hypoglycaemia by slightly increasing pre-bedtime glycaemia. [15][34]

### 2.1.2 *Factors affecting glycaemia*

Various metabolic alterations make it difficult to isolate causalities and predict glycemic response, and nutritional advice is often debated. The three most important estabilished factors in decisions about insulin dosage are pre-meal glucose, anticipated carbohydrate intake and physical activity. [39] The following section elaborates on the factors that influence the outcome, without attempting to be exhaustive or universally applicable. These might then either be included in some form as parameters in the model, or be identified as partially contributing to errors in its output.

Additional conditions that complicate maintenance of euglycemia, such as medication, smoking and renal failure, will be disregarded, as will atypical conditions such as pregnancy, illness or surgery. Furthermore, note is taken of intrapersonal fluctuations and age-related changes in hormones and metabolism (puberty, old age) and changes in relevant bodily traits such as weight or muscle mass. Although such attributes are very relevant in determining an initial insulin regimen, it is assumed here that an initial regimen (comprising a rudimentary insulin profile of the basal and meal insulins, number of basal insulin units injected each day, guideline carbohydrate/insulin ratio and guideline insulin/blood glucose change ratio) is available, as suggested by a medical professional. The system suggested here aims merely to fine-tune.

#### *Food consumption*

The most apparent action that increases blood glucose is ingestion of carbohydrates; the quantity of carbohydrates is therefore seen as the primary predictor of post-meal glycemia. [40]

However, qualitative information about food may be valuable as well, since it influences the extent to which and rate at which glucose will appear in the bloodstream. For this, the source of the carbohydrate (e.g. starch versus sucrose) is relevant, as are the nutrients co-

ingested with the carbohydrate. [40] For instance, fat slows glucose absorption, delaying the glycaemic peak. [42] Proteins may augment insulin release which increases glucose clearance, [40] and they may delay the glycaemic peak. [42] The physical form of the food is of influence; for example, glucose response to juice peaks quicker than it does to its whole counterpart, and fruit contains more fructose as it ripens. [40]

Of special interest is alcohol ingestion, following which insulin sensitivity is reduced with effects potentially lasting until the next day. [21][36]

*Insulin injection*

Variations in rate of insulin uptake exist between types of insulin or insulin analogues, and are classified generally into rapid-acting insulin (meal insulin, intended to counteract carbohydrate intake), long-acting insulin (basal insulin, usually injected at bedtime and intended to counteract fasting glucose increase) and intermediate-acting. [39] Between these classifications, there also is variation in the absorption rate and general effect on blood glucose of different insulin solutions (see table 1). [14] It may be problematic to time insulin absorption to occur simultaneously with glucose entrance in the bloodstream, especially given the variability of glucose absorption and the fact that there tends to be a single type of meal insulin at one's disposal at a single time.

| insulin | onset | peak | duration |
|---|---|---|---|
| regular | 30 min | 120–300 min | 360–480 min |
| lispro | 15 min | 15 min | 120–240 min |
| aspart | 15 min | 45 min | 120–300 min |
| NPH | 120–180 min | 300–420 min | 780–960 min |
| lente | 120–180 min | 420–720 min | up to 1080 min |

Table 1.: Variation between some insulins and insulin analogues. Table adapted from Shah & Zinman (2003). [39]

The site of injection has an effect on insulin uptake: insulin uptake peaks higher and quicker when injected in the deltoid and abdominal sites than it would when injected in the anterior thigh or buttocks. [14] The same goes for the depth and angle of the injection, where intermuscular injections peak quicker than subcutaneous injections. [30] Another aspect of injection technique comes into play when injections are repeatedly performed on the same site or with the same needle; tissue damage may develop, which modifies the rate and level of insulin absorption on the affected site. [45]

### *Exercise*

During exercise, muscle glucose uptake is stimulated and insulin sensitivity is increased. [34] The acute effects on blood glucose during and immediately after exercise are large and long-lasting; as well as decreasing blood glucose immediately, there's a late decrease of glycaemia 6-15 hours after exercise, with potential measurable effects up to 24 hours later. [46] Therefore, late-onset hypoglycaemia can occur even if insulin administration is appropriately decreased to avoid early hypoglycaemia.

The effect on glucose levels in response to exercise is relatively reliable within a single subject, [29] but varies with the intensity of the exercise, the time of insulin injection and the nutritional and glycaemic status pre-exercise. [46] For example, high-intensity exercise is associated with a smaller drop in blood glucose, due to an increase in glucose production in the liver. [10][46] This reaction to high-intensity exercise may also prompt post-exercise hyperglycaemia in patients with poor glycemic control. [46]

### *Circadian rhythm*

Insulin sensitivity varies across the day. Most importantly, it tends to be lower at breakfast than it is for a later meal, and lower in the second part of the night than it is diurnally. [49]

The abnormal metabolism in diabetes patients cues an increase of blood glucose even in the fasting state. [40] The fasting glycaemic increase necessitates the administration of a long-acting basal insulin, usually once every 24 hours at bedtime, although for some patients the imperfect 24-hour distribution of long-acting insulin requires extra long-acting insulin before breakfast. [39]

### *Other factors affecting blood glucose*

Psychological conditions such as stress and depression also contribute to glycaemia to an extent that cannot be wholly explained by behavioural differences in subjects that suffer from such conditions. [27][32]

Hyperglycaemia per se may negatively influence the continued uptake of glucose [52] and prompts urinary glucose excretion from the kidney. [25] Insulin has an additional effect on the liver, inhibiting its glucose production. [10] Low glycemic values prompt the body itself to produce counterregulatory hormones like glucagon, which is secreted through the $\alpha$-cells in the pancreas when blood glucose levels fall below approximately 3.6–3.9 mmol/l. [6] However, these responses are gradually lost in almost all patients with long-standing diabetes type 1 and are thus relevant mostly in new patients, or if glucagon is exogenously provided. [41][46] In newly diagnosed type

1 diabetes, some residual $\beta$-cell function may also influence blood glucose. [39]

### 2.1.3 *Noise in glycaemia evaluation*

Isolation of only those factors that influence actual blood glucose level does not suffice, if what is desired is an overview of possible variables that affect the accuracy of a model's outcome. This is due to noise in the input introduced by inaccurate readings of the blood glucose meter.

With the advent of continuous glucose measuring (CGM) devices, a steady and consistent collection of data from which to draw blood glucose values can be readily available. However, such devices as of yet often require sensor implantation, introduce more delays and inaccuracies, and are available to few patients for general use, especially amongst those who use multiple daily injections. [17] Since this is not consistent with the aim of this report, focus will be on widely used hand-held glucose meters using finger pricks. This means that the quantity of glycaemic data will be very limited.

According to the ISO 15197 quality standard, 95% of reported blood glucose values in such devices must be within $\pm$ 0.83 mmol/L of a reference method for concentrations less than 4.2 mmol/L and within 20% otherwise. [5] Although this is acceptable for home testing, significant variability of reported against actual blood glucose level is to be expected.

The accuracy of the devices may be further affected by elevation, temperature and humidity. [20] It may be amplified by user-level error such as contamination of the sampling site, applying pressure to the finger from which blood drops are squeezed, use of an expired or damaged test strip, or use of a poorly calibrated device. [5][19]

## 2.2 RELATED WORK

Models for glucose prediction can be roughly divided into qualitative physiological models wherein input is run on an explicit mathematical description of the body's metabolism to produce some output, and quantitative data-driven models wherein implicit relationships are derived from given input and output. The latter category has been researched because the complexity of the glucose-insulin system imposes limits on the former category. [16]

This report will take a data-driven approach, but it is still valuable to look at physiological models and artificial pancreata for comparison, such as the AIDA system, which attempts to simulate glycaemic curves through a physiological model and thus achieves a root-mean-square deviation ranging from 0.8 to 4.6 mmol/l, with a mean error $\pm$

standard deviation of $1.93 \pm 0.86$ mmol/l. [1]. Data for was extracted from 24 patients over 4-5 days. Predictive window is unknown.

As of now, the methods that are examined here seem to be aimed at possible inclusion in a closed automatic system in a more clinical setting, as opposed to being accessible in real-time and understandable for smartphone users. Approaching from the consumer side, the available smartphone applications tend to be focused on data tracking, training and nutritional reference, rather than on adaptive analysis. [47][7][12] A number of apps do provide insulin dosage suggestions through non-machine learning or unknown (proprietary) approaches. [12]

*Sandham et al (1998)*

A recurrent artificial neural network was used by Sandham et al [38] through back-propagation in a 95-neuron Elman network, using insulin, carbohydrate content, exercise, blood glucose concentration and a binary value for 'other complications' as input. The data set contained only 122 samples, and results were 'mostly' within a 1.5 mmol/l error.

*Kok (2004)*

Kok created a back propagation feed-forward neural network split into seperate networks for morning, afternoon, evening and night. [23] Input consisted of a varying set of variables (blood glucose, carbohydrate intake, rapid-acting insulin, stress, exercise) at the start and during an interval; the output was the predicted blood glucose level at the end of that interval. The architectures of the networks were static and manually varied to find that, for the given parameters, performance differed only slightly as the number of units increased. Single- and double layer networks performed comparably. Using a single hidden layer network with 11 hidden nodes, on-line training with learning rate $0.1 \leq \eta \leq 0.2$ and momentum $0.5 \leq \alpha \leq 0.7$, on data gathered from 1 patient over 77 consecutive days, a reasonable average root-mean-squared deviation of about 2.3, 2.1, 2.4 and 2.3 mmol/l resulted for respectively the morning, afternoon, evening and night intervals.

*Pappada et al (2008)*

Pappada et al created a back-propagation feed-forward neural network with a memory component. [33] Genetic algorithms were used to optimise the number of neurons, the learning rate and the momentum. They recorded detailed data on 18 patients over 3 to 9 days using CGM technology and a real-time electronic diary. This resulted

in a mean absolute error percentage of 18.7 to 25.8% on a predictive window of 100 minutes, but hypoglycaemic values were strongly overestimated, with an error percentage of 44 to 61.6%.

*Zainuddin et al (2009)*

Zainuddin et al improved on Kok's approach by using a wavelet neural network with various wavelet families. [53] The dimensionality of the input was decreased through principal component analysis. On the same data set, they achieved a scaled root-mean-square deviation of 0.017 to 0.045.

# 3 METHODOLOGY

The feed-forward neural network will be customly written in the C# programming language. This choice of language was made since it is intuitive to express a neural network in an object-oriented way, while leaving room for multiparadigmatic approaches. A secondary consideration is binary compatibility across all major platforms.

For educational purposes, no pre-existing external tools for neural networks were used. To facilitate experiments with principal component analysis, the Accord.NET framework was deployed. [44]

While an explanation on the general operation of neural networks is omitted here, commented source code to the companion software is available.

For a user-level manual to the application, call `$ bin/ann.exe -h` on Windows using the .NET framework or `$ mono bin/ann.exe -h` on any platform using the Mono framework.

The network will be trained on a Linux machine through the `Mono JIT compiler version 3.4.0`. The software package `gnuplot` will be used for illustrating performance.

## 3.1 PARAMETERS

### 3.1.1 *Architecture*

A single back-propagation feed-forward neural[1] network was selected as the machine learning method. This is appropriate because a complicated function is to be derived from examples that are representable as real-valued vectors.

This type of network can theoretically be used to approximate any function to arbitrary accuracy, using a finite number of neurons and layers, [31, 105] but this fact does not address the question of how many neurons and layers such a network will require. [35] To reduce complexity, no algorithm was employed to find the optimal answer. Trial and error will yield some acceptable architecture.

More complex types of neural networks have been researched for this purpose by others in the past, [38][53] but it bears merit to reassess the performance and re-examine possibilities (such as principal component analysis and $k$-fold cross-validation) within a simpler

1 The nodes or units of the network will be referred to as *neurons*, because it is an evocative term with immediate clarity in this context. It is acknowledged that the resemblance to their biological namesake requires leniency.

framework. Additions to this basal network were considered an option if time were to allow for it — it is now left to others.

Others have also suggested making four seperate networks: one for each part of the day. [23][53] This was justified by the notion that there would not be a great correlation between the glycaemic course of different intervals. [53] However, the solutions here were condensed into a single network so as to allow for experiments with increased volume of training data and decreased dependence on day-to-day structure of the patient.

### 3.1.2 *Training*

On-line training was used to train the network. Batch training was unexamined, since neither theoretical considerations [51] nor empirical results [23] appeared to show much promise for this method.

### 3.1.3 *Partitioning, error and stop criterion*

For each run, the data set was randomly partitioned into one set for use during training (the *training set* covering $n \times 100\%$), and one set used for evaluating performance after training was done (the *comparison set* covering $m \times 100\% = 1 - n \times 100\%$). The sizes of these sets can be varied, but experiments indicated that the comparison sets should not be much smaller than 0.3. This is due to the small size of the data set itself. The training set should be as large as possible, but the reported performance needs to still be meaningful. It was found that $n = 0.65, m = 0.35$ yielded acceptable results. Therefore, unless otherwise noted, these numbers are used during partitioning.

Since the size of the data set is so limited, $k$-fold cross-validation was used to determine an acceptable stopping point. This eliminates the need for an additional set to measure unbiased performance after back-propagating, which means that all training data may be used for training — there is no seperate validation set.

With this method, the network is trained $k$ times, from scratch, but with equal initial conditions where possible. Each time, a random selection of $\frac{n}{k} \times 100\%$ of the original data set is used as validation and the other $\frac{n(k-1)}{k} \times 100\%$ is used for training, such that each sample acts as validation sample exactly once. During each of these training sessions, the best iteration $i_j$ is recorded, where the *best* iteration is defined simply as the one at which the lowest *rmse* on the validation set occurred. The network is then trained one final time, using *all* training examples, for $\bar{i} = \frac{\Sigma_j i_j}{k}$ iterations; it is assumed that this iteration will be an acceptable stopping point. [31, 112]

The $k$ training sessions stop simply when they reach a maximum amount of iterations $M$. Since it is quite expensive to re-train a net-

work, $k$ is kept low at the cost of some accuracy for $\bar{i}$. Unless otherwise noted, $k = 2$.

Unless otherwise noted, the maximum is set to $M = 5000$; this limit was chosen since the best performances in preliminary experimentation, for reasonable parameter values, tended to occur in the general area between 100 to 3000 epochs. Finding the best performance after this point was thus deemed unlikely.

### 3.1.4 *Committee*

Output values may be averaged over a committee of $N$ networks that were trained using the same data and set to the same parameter values, but that were initialised with different weights. Dependence on chance is in this way reduced (output is less likely to stray far from expected values) and it has been shown that combining the 'ambiguity' between predictors that have arrived in different local minima (where one is not necessarily better than the other) leads to better generalisation. [24]

For all results where $N > 1$, the committee size will be reported. This is the case at least for all results for which parameters are *chosen* rather than *considered*.

### 3.1.5 *Activation function*

The back-propagation algorithm requires a function $\sigma$ as a final transformation on the output of a neuron $j$, such that its output $y_j = \sigma(\Sigma_i x_{ij} w_{ij})$. Here, $x_{ij} w_{ij}$ is the product of the value of input neuron $i$ of $j$ and the strengths of the connections between the corresponding neurons (weights).

The function $\sigma$ should be easily differentiable, such that it is straightforward to calculate the gradient on the error surface with respect to these weights. This gradient is necessary to adjust these weights in such a way that at each sample presentation, the network output tends a little more toward the desired output for that sample (see section 3.1.8).

Two such functions were examined: the hyperbolic tangent $\tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$ (derivative $\tanh'(x) = 1 - \tanh^2(x)$) and the logistic sigmoid $S(x) = \frac{1}{1+e^{-x}}$ (derivative $S'(x) = S(x)(1 - S(x))$). Since the logistic sigmoid was found to produce better results during superficial testing, it was used without further commentary.

### 3.1.6 *Momentum*

A momentum term $\alpha \in [0, 1]$ was used as a means overstep local minima on the error surface (see section 3.1.8). It works by retaining

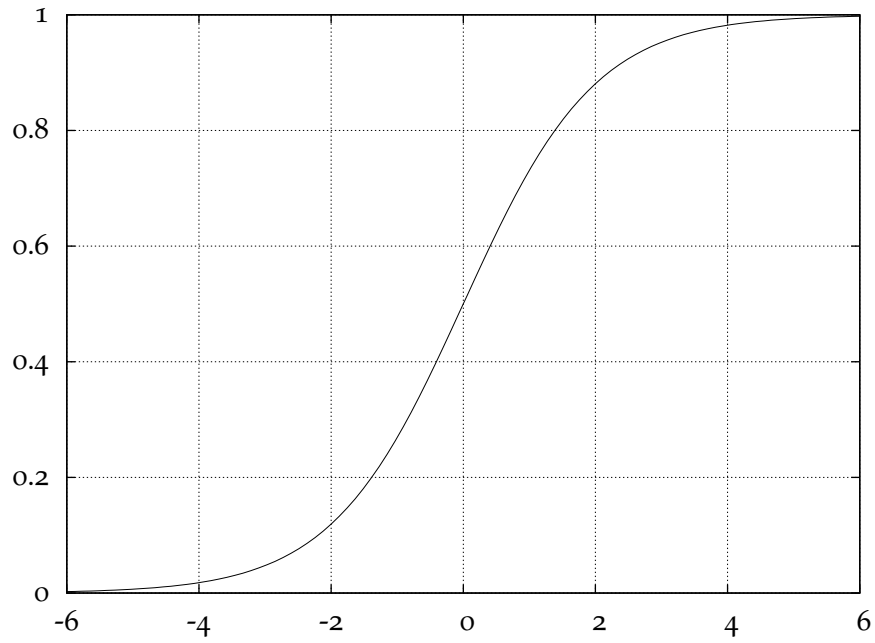


Figure 1.: The logistic sigmoid $S(x) = \frac{1}{1+e^{-x}}$.

a fraction $\alpha$ of the movement along the error surface in the past time step: like a ball, the solution will then roll over small pits instead of getting stuck in them. This term will be varied in the experiments.

### 3.1.7 *Learning rate*

The learning rate $\eta$ determines the rate by which the weights adapt to each sample (see section 3.1.8). In the system, it is given by a linear function $\eta'_d$ on the layer depth $i$ of the neuron that is being adjusted; the reason for this is to counteract the nonlinear error flow inside a network of $\hat{i} > 1$ layers and to obtain more or less equal learning rates for all the network parameters. [35] The rule of thumb suggested in literature is to increase the learning rate from the output layer to the input layer by a factor $d$ of 2 to 5. [35] The weight update rule then uses $\eta = \eta'_d(i) = \hat{\eta}(1 + \frac{i}{\hat{i}-1}(d-1))$, where $\hat{\eta}$ is the initial learning rate on the output layer, $i$ is the layer depth starting at 0 at the output layer and $\hat{i}$ is the number of layers in the network. Both $\hat{\eta}$ and $d$ will be varied in the experiments.

### 3.1.8 *Weight update rule*

The following section explains the back-propagation rule exactly as it is implemented in the companion neural network application. The concept was adapted from chapter 4 in Mitchell. [31]

*Weight change*

The weights $w_{ij}$ connecting neurons $i$ to neurons $j$ of the network, are changed at each epoch or time step $t$ by $\Delta_t w_{ij}$, where $\Delta_t w_{ij}$ is based on the back propagation weight update rule with momentum: $\Delta_t w_{ij} = \eta'_d(i)\delta_j x_{ij}(1-\alpha) + \alpha\Delta_{t-1}w_{ij}$. At each sample presentation, the output will tend a little more toward the desired output for that sample.

Here, $\eta'_d(i)$ is the learning rate on layer $i$, $\alpha\Delta_{t-1}w_{ij}$ is the momentum generated from change during the previous epoch $t-1$, $x_{ij}$ is the $i$'th input value to neuron $j$, and $\delta_j$ is its error term.

*Error term*

The error term of $j$ is coupled with each of its inputs $i$ to express the gradient $\frac{\partial E}{\partial w_j}$ along the error surface $E$ with respect to the corresponding weights $w_{ij}$, and so it is defined $\delta_j x_{ij} = \frac{\partial E}{\partial w_{ij}}$. Using this gradient, the tools are present to move the solution along the error surface in the hypothesis space (that is, the space of all possible weight vectors).

The error surface $E$ that was used, is $E = \frac{1}{2}(y^* - y)^2$, for easy differientability and to assign a relatively steeper slope for larger errors. Here, $y_j^*$ and $y_j$ are respectively the desired and the actual outputs for $j$. It is also useful to note the output of $j$ before transformation by $\sigma$, namely $u_j = \sigma^{-1}(y_j) = \Sigma_i x_{ij} w_{ij}$, and the derivative of the activation function, $\sigma'(z) = \frac{\partial\sigma(z)}{\partial z}$.

When the weight has a connection to the output layer, $\frac{\partial E}{\partial w_{ij}}$ is simply given by differentiation using the error at the corresponding output neuron. Therefore, $\delta_j x_{ij} = \frac{\partial E}{\partial w_{ij}} = (y_j^* - y_j)\sigma'(u_j)x_{ij}$.

For all other weights, the term is calculated by back-propagating: summing the error terms on neurons $k$ of the next layer, $\delta_k$, weighted by the extent to which neuron $j$ contributes to the errors on that layer, and differentiating there: $\delta_j x_{ij} = \frac{\partial E}{\partial w_{ij}} = \Sigma_k \delta_k w_{jk} \sigma'(u_j) x_{ij}$.

When $\sigma'(u_j)$ can be expressed in terms of $\sigma(u_j)$, it can clearly be expressed in terms of $y_j$, since $y_j = \sigma(u_j)$. For example, for the logistic sigmoid, $\sigma'(u_j) = y_j(1-y_j)$. This is convenient because each $y_j$ has to be calculated already when the network error vector $y^* - y$ is determined.

## 3.2 DATA

### 3.2.1 *Simplifications and assumptions*

The selection of components should make medical sense, and since the goal is applicability outside a clinical setting, additional constraints are put on the available input. These are not as limiting and oversimplifying as they may seem, since humans are subject to them as well. Therefore, the volume and quality of the data need not be less than what the patient or physician normally bases decisions on.

It is expected that many of the factors explicated in section 2.1.2 can be disregarded in a practical model, because some can be considered immutable in a single instance. For example, weight, muscle mass, age, residual pancreas function and insulin type are unlikely to change often or very quickly. Hence, the model may treat them as constant and adapt to them as it is expected to do to the general metabolism of the subject. Furthermore, internal processes, when they are neither easily quantifiable nor directly influenced by outside action, remain hidden and cannot be taken into consideration even if they are relevant. Therefore, liver glucose absorption and production, renal excretion, and other metabolic fluctuations will be disregarded.

Mistakes and errors, though inevitable, cannot be parametrised. Therefore, it is assumed that all injections are subcutaneous, that injection sites are often circulated, that fingers are washed before blood sampling, that the blood glucose meter is calibrated and used in a normal environment and that all reported blood glucose values are within 20% of their true value.

Since carbohydrate ingestion is such a significant factor, input is to be based on meals (or their absence). There will be three slots a day for the main meals, plus a nightly slot. This provides a way to deal with arhythmicity in meal timing, and implicitly deals with circadian rhythm. Patients are expected to report blood glucose before bed, immediately before each meal, and optionally some time afterward. Each slot will cover the blood glucose measurements at the start of the meal until the start of the next meal.

Estimating the full effect of a meal is thought to be too complex, so a meal will be fully characterised by its carbohydrate content. Although accurate carbohydrate counting requires some education on the part of the patient, [37] nutritional reference works [2] and smartphone applications [28][26] exist to facilitate the process. Alcohol will be left out of consideration.

It will be assumed that only two (invariable) insulin types are used: one basal insulin and one meal insulin. The relatively modest effect of insulin injection site (buttocks and thighs versus abdomen) is omitted.

Psychological condition will be considered to be too complex and rare to take into account at the current time. An exception is stress level, which may be taken into account if information is present.

Physical effort is the last of the significant factors. In keeping with the rest of the input, intensity level is associated with some interval around a meal. The way in which the level of physical activity is codified to a numerical value is irrelevant as long as it is done consistently. To account for both immediate- and late-onset hypoglycaemia, a measure of physical effort over the current interval is given seperately from the one over 6-15 hours before.

Reviewing section 2.1, no factor (that is also taken into account in the parameters here) seems to have a effect duration that is significant for more than a few hours, save for physical effort and basal insulin. It therefore seems sensible to take into account actions and measurements of no more than one meal slot back, plus some measure of the basal insulin over the past 24 hours and past physical effort.

Feedback effects such as the self-limitation of hypo- and hyperglycaemia are dealt with internally, since previous glucose values in time act as input.

### 3.2.2 *Data collection and quality assessment*

Unfortunately, it is beyond the scope of this report to collect data. A pre-existing data set will therefore be used, from which input will be compiled into the desired form.

The primary data set was provided by Peter Kok and is the same as the one used in his research assignment. [23] The data was gathered from 1 patient, covering at most 8 daily measure points (middle of the night, before breakfast, after breakfast, before lunch, after lunch, before dinner, after dinner and before bed time) over 77 consecutive days (308 records). Additional data was discarded. Each record may or may not have a value for date, time, blood glucose in mmol/l, short acting insulin, long acting insulin, carbohydrate intake, exercise and stress. Exercise and stress are expressed on a 1-5 scale. They were only recorded on the 'before' records and cover the entire interval until the next 'before' record. The patient was not under a special food or exercise regimen and used NovoRapid as short acting insulin and Lantus as long-acting insulin. This set will be used to train towards parametrisation of the individual blood glucose profile. It is expected that the size of the set is large enough to be able to train an acceptable neural network, while not covering so much time that the glucose metabolism is likely to have significantly changed during the interval.

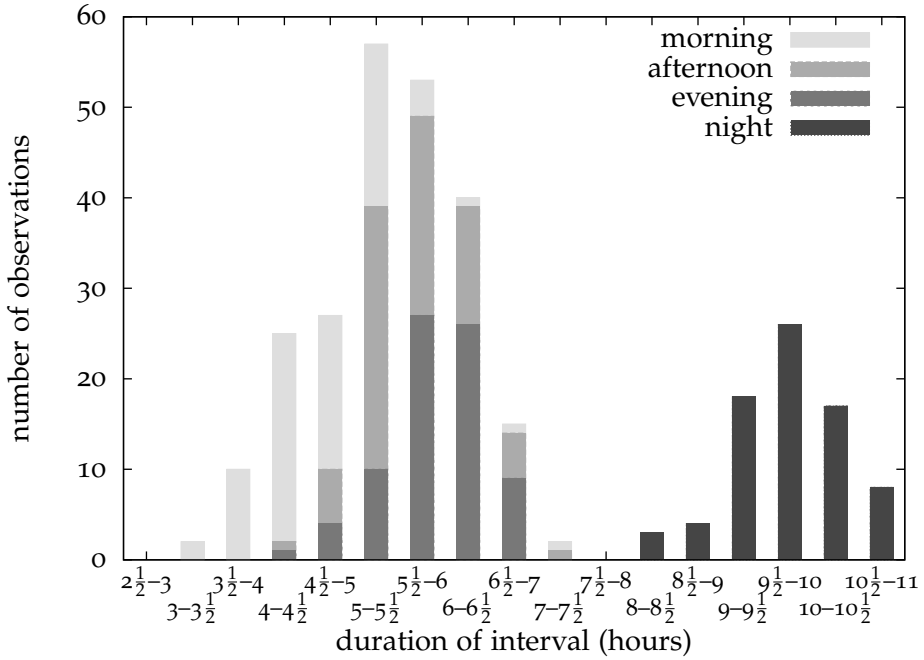


Figure 2.: Duration of the intervals (predictive windows) in Kok's data set, split out for interval type.

### 3.2.3 *Conversion and preprocessing*

*Output selection*

The output of the model is demarcated by the research question — it must be the blood glucose level at some point in the future. This point lies at the end of an interval spanning at most until the next meal. This predictive window seems fitting, because anything beyond it is surely indeterminable considering the additional glucose intake; and although significant changes may surely occur during it, such changes are deemed beyond the scope of prediction with the tools currently at hand.

*Input selection*

The attributes of the input vectors, however, may be varied. They are drawn from the simplified information that is necessary and available to us.

The dimensionality of these input vectors should be high enough to exhibit a correlation, but not so high as to harm the generalising power of the network. This is an especially critical concern when dealing with a relatively small data set, as ours will be.

With the previous assumptions, the input for a training sample is reduced to:

- moment of the day (morning, noon, evening, night);
- interval duration;
- pre-meal *and* post-meal blood glucose levels;
- carbohydrate ingestion during *and* after main meal;
- rapid insulin intake at *and* after main meal;
- sum of basal insulin dosage over the past 24 hours;
- physical activity level during the interval;
- sum of physical activity level over the previous 2 intervals;
- glucose level at the start of the previous interval;
- total carbohydrate ingestion over the previous interval.

*Transforming data into input/output samples*

Kok's data is given as 8 data points a day. Most of this structure is retained in transforming that data into an input vector for the neural network: three consecutive points (a point from before a meal, after that meal and before the beginning of the next meal) are simply taken together to extract input data (first two points) and associated output data (last point). This compound information will be referred to as an *observation*.

| | nt | bb | ab | bl | al | bd | ad | bn |
|---|---|---|---|---|---|---|---|---|
| time | | 7:14 | | 12:04 | 15:38 | 17:22 | | 23:09 |
| blood glucose | | 9.3 | | 3.7 | 4.2 | 7.4 | | 8.5 |
| rapid insulin | | 8 | | 5 | | 6 | | 0 |
| slow insulin | | 22 | | 0 | | 0 | | 3 |
| carbohydrate | | 86 | 6 | 117 | 46 | 88 | 30 | 14 |
| exercise | | 3 | | 5 | | 2 | | 2 |
| stress | | 2 | | 3 | | 2 | | 1 |

Table 2.: Raw input data from Kok over a single day (April 21st, 2004). [23] The highlighted area indicates an example observation.

Training samples draw their attributes from multiple subsequent observations.

*Normalisation*

Proper functioning of the neural network requires that the inputs be on an appropriate interval for $\sigma$.

There are various methods for achieving this: a non-linear scaling method may be appropriate if the data is exponentially distributed;

a linear transformation to $z$-scores may be used so that each component has a mean of 0 and a unit standard deviation; and a simple linear mapping on each input component $i$ from $[\min_i, \max_i]$ to the interval $[l, u]$ is possible. The latter is used here since it is the most straightforward. Since the activation function is the logistic sigmoid $\sigma = S$, which has asymptotes at $y = 0$ and $y = 1$, $[l, u]$ is set to $[\frac{1}{10}, \frac{9}{10}]$.

## 3.3 PERFORMANCE

There exists no single, standard way to express all aspects of the performance of a glycaemic model (see section 5.2). The networks in this report are evaluated through Pearson's correlation coefficient, the mean absolute error percentage and the root-mean-squared deviation, as well as performance and convergence plots.

To obtain an unbiased indication of the performance, a dedicated portion of the data, seperate from all data that has been used to train, will be used to evaluate the system. This data is referred to as comparison or test data.

Note that comparisons of networks are meaningful only if the network is instantiated repeatedly, since learning is a stochastic process that depends on the initial conditions. [35] Therefore, all reported performance measures, unless they are only sensible in the context of a single partitioning, or when they are specifically used to illustrate the variance caused by the initial conditions as in section 4.1, are averages of $n$ results of the algorithm run with identical parameters, but different initialisation values. This means that for instances where outputs to specific samples are irrelevant, as they are in convergence plots, performance is measured by the course of the performance of multiple instances of those networks on varying but equally-sized partitions of the data. Since performance plots require testing against one specific comparison set, they are an example of an instance where performance of network(s) is measured on one fixed partitioning. The networks may still be averaged over a committee of multiple networks.

In the following equations, $\dot{x}$ are the target output values (references), $\dot{y}$ are the actual output values (predictions), $\overline{x}$ and $\overline{y}$ are their respective means, $n$ is the size of the data set, and $s_{\dot{z}} = \sqrt{\frac{1}{n-1}\Sigma_i(\dot{z}_i - \overline{z})^2}$ is the sample standard deviation of $\dot{z}$.

### *Root-mean-squared deviation*

The root-mean-squared deviation is a generalised average of the network's error, given by $rmse = \sqrt{\frac{\Sigma_i(\dot{y}_i - \dot{x}_i)^2}{n}}$. This has the desirable property of assigning more weight to one large error than to a multitude of small errors. Taking the square root ensures that the result is

expressed in the same unit. When the network exhibits a high *rmse*, performance is bad, but the inverse is not necessarily true: a small *rmse* can also result from improper distribution of the comparison data, when all data points lie close together. [23]

Note also the following:

$$\begin{aligned} rmse_{\overline{x}} &= \sqrt{\frac{\Sigma_i(\dot{y}_i - \dot{x}_i)^2}{n}} = \sqrt{\frac{1}{n}}\sqrt{\Sigma_i(\dot{x}_i - \overline{x})^2} \\ &= \sqrt{\frac{n-1}{n}}\sqrt{\frac{1}{n-1}}\sqrt{\Sigma_i(\dot{x}_i - \overline{x})^2} = \sqrt{\frac{n-1}{n}} \cdot s_{\dot{x}} \\ \text{where } \lim_{n\to\infty} &\sqrt{\frac{n-1}{n}} = 1 \end{aligned}$$

From this observation follows that, for any large data set, any system that in all cases simply returns the mean blood glucose value ($\dot{y} = \overline{x}$), already achieves an *rmse* that is close to the standard deviation $s_{\dot{x}}$. Since the standard deviation of the target output values of the observations in Kok's data set is $s_{\dot{x}} = 2.84$ mmol/l, with mean $\overline{x} = 9.24$ mmol/l, a system that returns 9.24 mmol/l for all inputs will have $rmse \approx 2.84$ mmol/l. This performance may be called the token performance, since it is the upper bound on the performance of this trivial model. It is illustrative of the expressive power of the root-mean-squared error.

*Correlation coefficient*

The correlation coefficient is a measure of the linear correlation between two variables, in this case the predicted output and the reference output. It is given by $r = \frac{1}{n-1}\Sigma_i(\frac{\dot{x}_i - \overline{x}}{s_{\dot{x}}})(\frac{\dot{y}_i - \overline{y}}{s_{\dot{y}}})$.

If the network performs well, it will exhibit an $r$-value close to 1. Note that a high correlation does not necessarily entail that it is the correct one, since any linear function has $r = 1$. However, this is of no great concern here since any slanted correlation can be corrected by a linear transformation on the predicted values.

*Mean absolute error percentage*

The mean absolute difference is reported, given by $mae = \Sigma_i|1 - \frac{\dot{x}_i}{\dot{y}_i}|$. This value is reported only to maximise comparability between research.

*Visualisations*

Performance plots of reference versus predicted values will be provided. In the optimal case, the points should reveal a $1:1$ relationship, e.g. lie on the line given by $y = x$. A simple linear regression

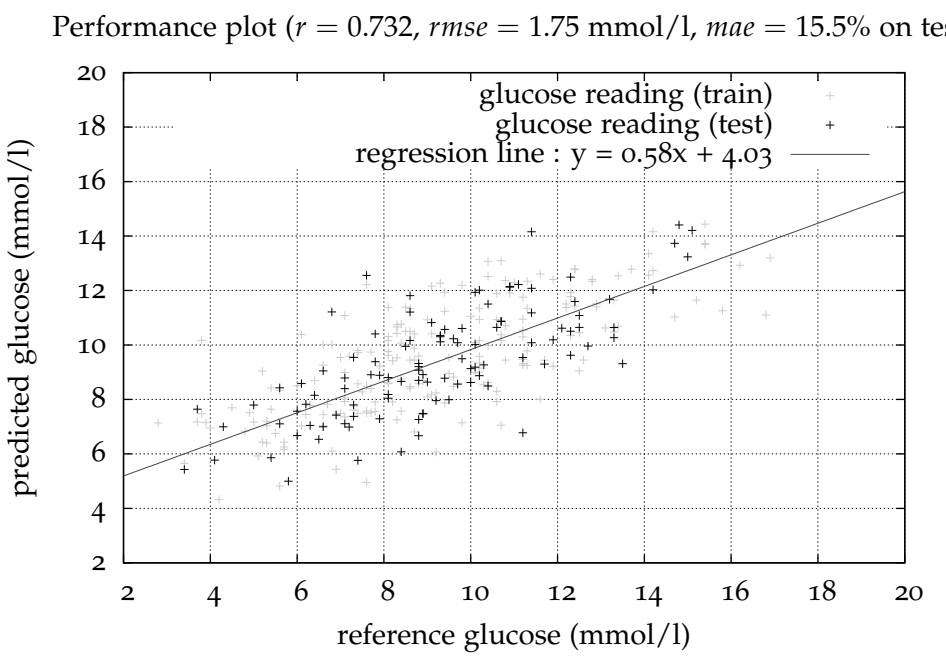


Figure 3.: Example performance plot. One hidden layer of 15 neurons, $\alpha = 0.5$ , $\eta = 0.5$ , $d = 1$.

line using the method of least squares (the line $l$ minimising the mean square error between it and $y$) is drawn to illustrate the correlation. This line is given by $l : y = \frac{s_y}{s_x} \cdot r(x - \overline{x}) + \overline{y}$. [18, 126]

Finally, for each data partition, plots of $r$ and $rmse$ against the training epoch will illustrate the rate of convergence and learning of the network over time. To get a feel for how much the convergence plots may vary over multiple runs, an area of 2 standard deviations around the averaged convergence plot for the comparison data has been highlighted.

Note that in later epochs of the performance plot, the averaged performance measures will stabilise as more networks become stationary at their stop condition.

# 4

# RESULTS

It is infeasible to test all possible parameters in all possible configurations, due to the combinatorial complexity of doing so. Therefore, results of experiments are given for some reasonable configurations.

The best network found had $rmse = 1.749$ mmol/l, and was plotted in figure 3. However, as can be deduced from section 4.1, single network performance is not immediately meaningful. As it will turn out, the following performance plot represents a fairly typical instance of the system.

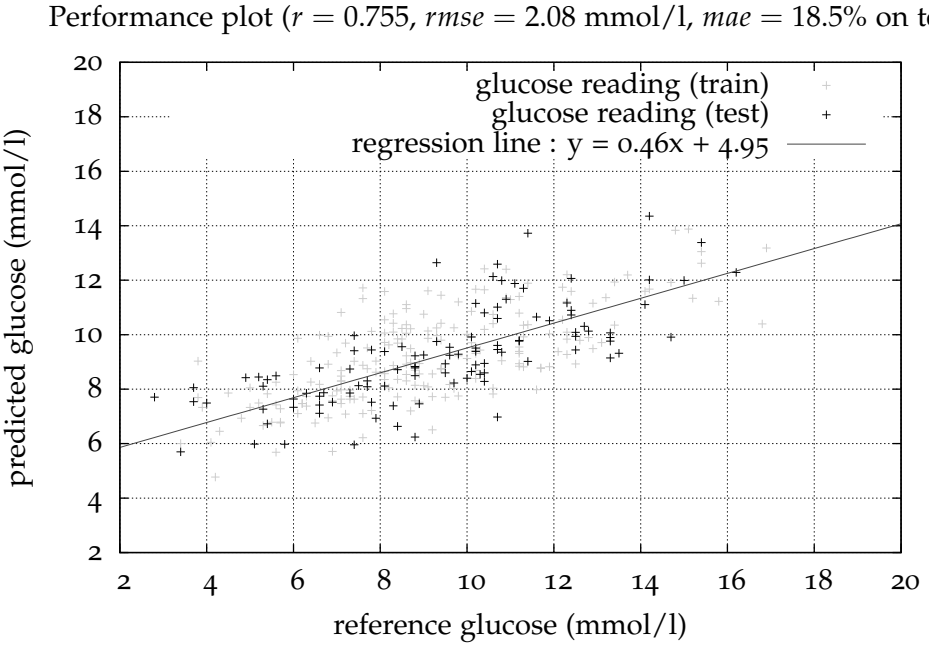


Figure 4.: A performance plot obtained with the recommended variables.

In the committee of 20 from which this plot was obtained, each network consisted of a single 8-neuron hidden layer and had its parameters set to $\alpha = 0.5$, $\eta = 0.5$ and $d = 1$. The typical results[1] for such a committee are as follows: $rmse = 2.156 \pm 0.131$ mmol/l, $mae = 19.4\% \pm 1.3$ pp, and $r = 0.662 \pm 0.042$. These results were obtained from 20 runs of these 20 networks.

1 Note that the notation $\overline{x} \pm s_{\dot{x}}$ here reports the mean and the sample standard deviation. The value after $\pm$ is not the margin of error. No assumptions are made about the distribution of sample committees.

The figures were broken down for each interval type. Results can be inspected in figure 3.

Since the number of samples from a specific interval type in the comparison set was smaller and could vary per run, the performances that are reported here display more variation.

| | | performance | | |
|---|---|---|---|---|
| | | $\bar{r} \pm s_r$ | $\overline{rmse} \pm s_{rmse}$ | $\overline{mae} \pm s_{mae}$ |
| interval | *combined* | $0.655 \pm 0.043$ | $2.169 \pm 0.107$ | $19.6\% \pm 1.4$pp |
| | morning | $0.738 \pm 0.078$ | $2.147 \pm 0.223$ | $22.9\% \pm 2.9$pp |
| | afternoon | $0.661 \pm 0.076$ | $1.853 \pm 0.182$ | $17.1\% \pm 2.9$pp |
| | evening | $0.348 \pm 0.115$ | $2.416 \pm 0.247$ | $18.3\% \pm 1.9$pp |
| | night | $0.580 \pm 0.067$ | $2.160 \pm 0.147$ | $20.1\% \pm 1.6$pp |

Table 3.: Breakdown of performance over interval types. Better values are darkened.

## 4.1 STOCHASTIC ELEMENTS

There are 3 random factors that cause variability in the output.

- The initial weights on the connections. They are initialised to small random values on the $[-\frac{1}{2}, \frac{1}{2}]$ interval.
- The selection of samples that make up the training and comparison partitions.
- The order in which these samples are presented to the network. Of course, if the partitions differ, the presentation order must also be different.

Experimentation revealed that varying network sizes causes neglible change when compared to the change caused by varying the distribution of the data over the partitions. To illustrate the stochastic volatility of the results, the performance was measured of one 8-neuron, 1-hidden layer network with $\alpha = 0.5, \eta = 0.5, d = 1$ and averaged[2] over 40 runs in each one of the following situations:

- when keeping only the network weights constant across the rest of the runs;
- when keeping only the partition distribution constant across the rest of the runs;
- when keeping everything but the network weights constant across the rest of the runs;
- when keeping everything but the presentation order constant across the rest of the runs;
- when repeatedly initialising all variables randomly, as is customary.

Evidently, the same result will occur endlessly if *all* random variables are kept constant.

2 Note that the plots themselves are only examples, since the values that are kept constant are still very much randomly initialised. By the nature of these experiments, proper averaging would have to occur on a (very computationally expensive) higher level.

*Reference performance*

Figure 5 displays the convergence plot as it might appear when random variables are all allowed to change across runs. The *rmse* here is at a typical $2.239 \pm 0.134$ mmol/l.

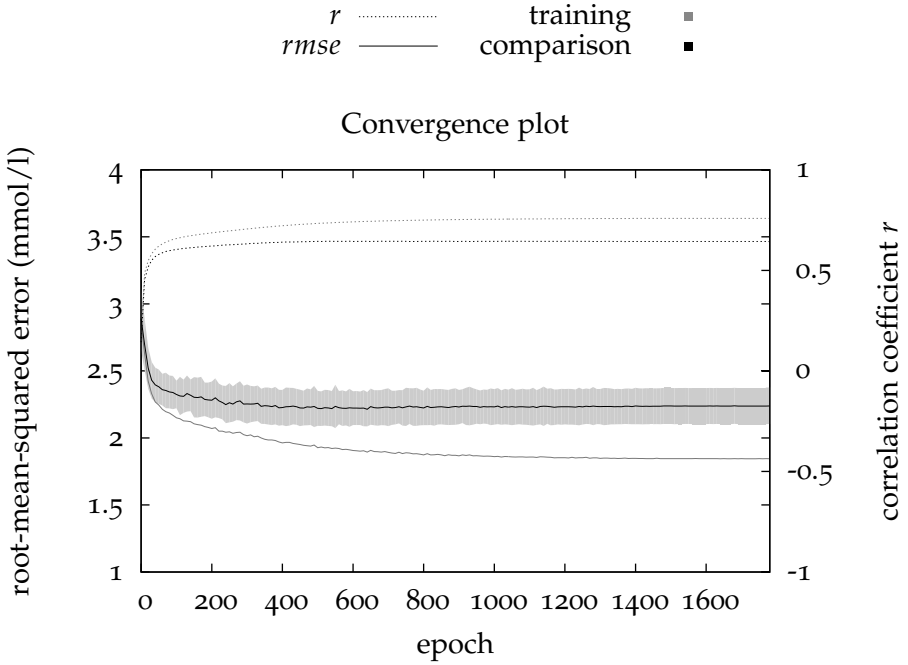


Figure 5.: Allowing random variables to stay random across runs.

*Influence of sample distribution and presentation*

In figure 6, the algorithm sticked to the weights as they were chosen at the beginning. At an *rmse* of $2.221 \pm 0.140$ mmol/l, the course does not look significantly different from the one that preceded it. The standard deviation is more or less the same. Apparently, the factor that is left — weight initialisation — is not such a significant contributor to variability — not, at least, in the face of variable sample distribution and shuffled training samples.

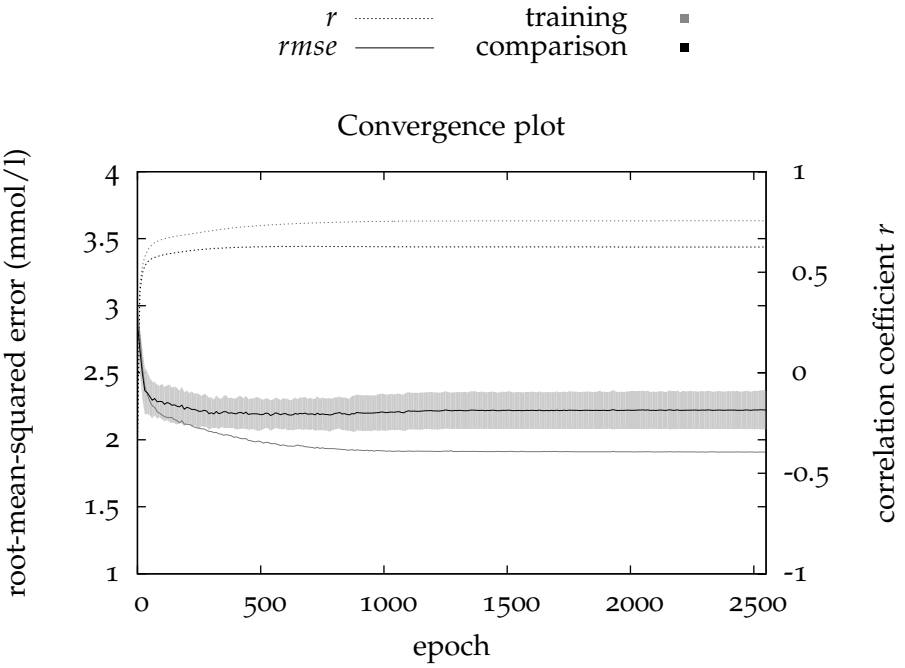


Figure 6.: Keeping constant the initial weights across 40 runs.

*Influence of sample presentation order*

When the the weights and the partition distribution are kept constant, as they were in figure 7, the only factor that causes variability is the sample presentation order. The standard deviation drops; the *rmse* is at $2.276 \pm 0.056$ mmol/l. Therefore, sample presentation order can be said to have a comparatively small effect on variability.

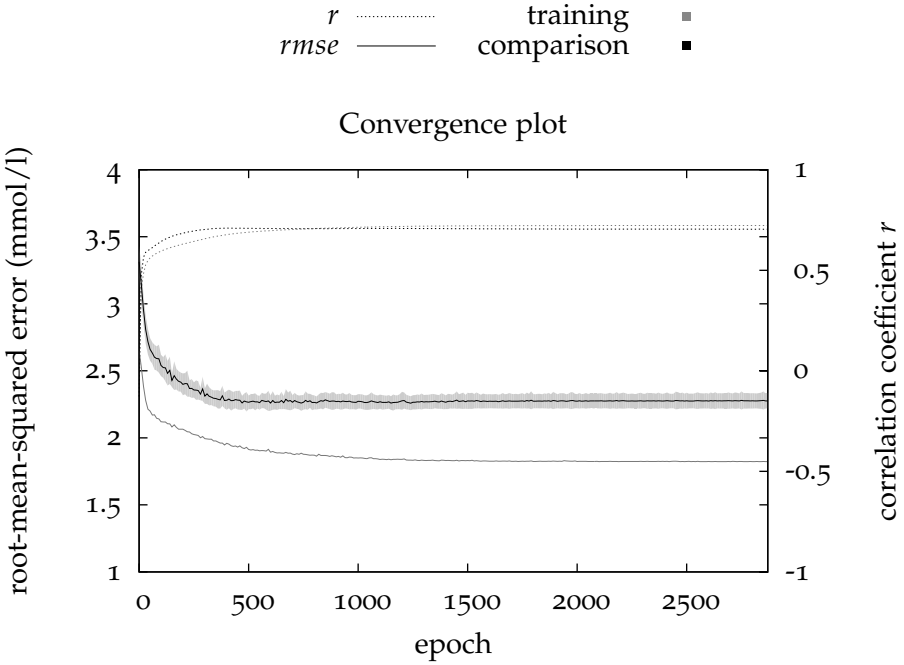


Figure 7.: Keeping constant everything but the sample presentation order across 40 runs.

*Influence of weight initialisation*

Figure 8 demonstrates what happens if the partition distribution and presentation order are kept constant.

This figure is the most interesting, because it illustrates that the make-up of the training data is almost the sole culprit of end variability. Weight initialisation plays secondary role.

The curve lacks much of the smoothness of previous plots; despite the randomness that might be expected from random weight initialisation, all of the 40 networks are guided toward the same general course. After some number of epochs, the standard deviation does begin to increase, but it does so painstakingly; it ends at $2.342 \pm 0.047$ mmol/l.

Given that the *rmse* is at the high end of the spectrum, it is assumed that the initial partition distribution was not a particularly good one in this case. It can be concluded that distributing the training data wisely is by far the most fruitful way of obtaining a good performance — but doing so puts a bias on the reported performance and consequently defeats the purpose of this survey.

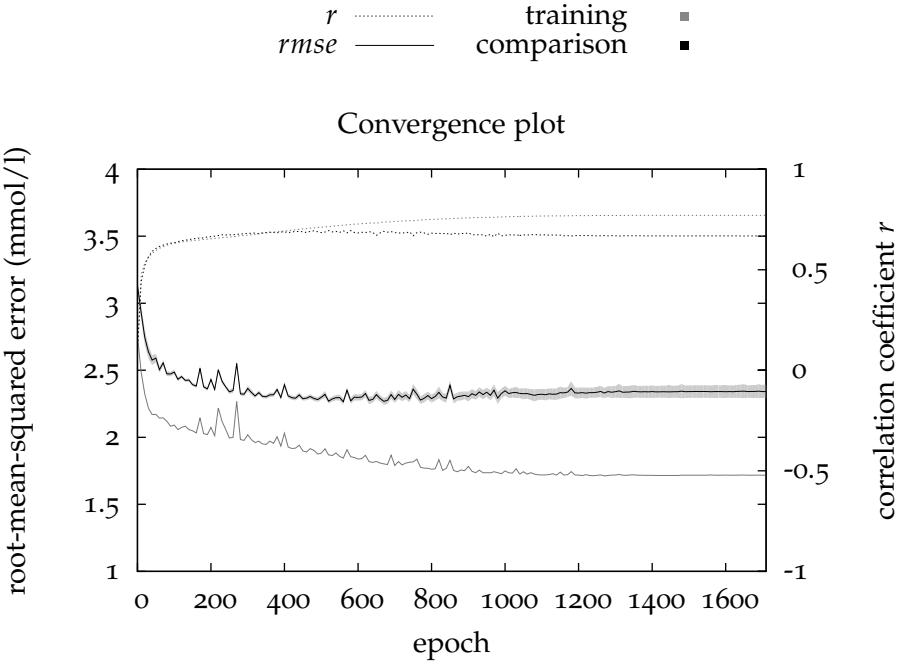


Figure 8.: Keeping everything but weight initialisation constant across 40 runs.

*Combined influence of weight initialisation and sample presentation order*

Finally, the result of keeping only the partitions constant will look like figure 9. As might be expected, the variance is more or less a combination of the previous two figures, with the *rmse* at $2.251 \pm 0.094$ mmol/l.

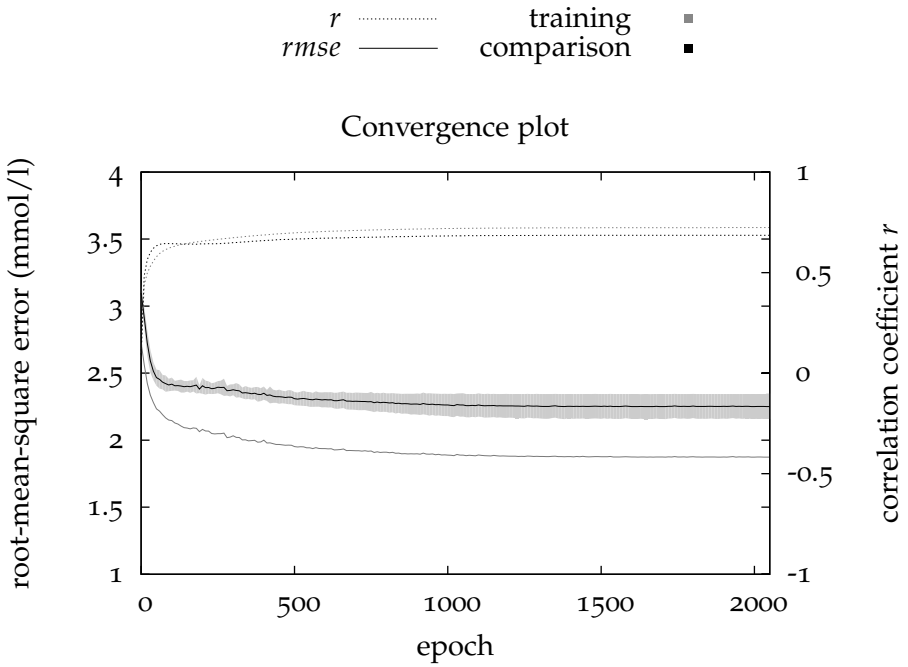


Figure 9.: Keeping the partition distribution constant across 40 runs.

## 4.2 PARAMETER VARIATIONS

It was found that network configuration has a limited effect on performance. The recommended configuration is therefore a conservative combination of 1 8-neuron hidden layer using $\alpha = 0.5$, $\eta = 0.5$, $d = 1$.

*Learning terms*

On a modest network of one hidden layer of 15 neurons, the learning rate $\eta$, momentum term $\alpha$ and layer increase factor $d$ were varied such that all solutions satisfying $(\alpha = 0.15 \veebar \alpha = 0.5 \veebar \alpha = 0.7) \wedge (\eta = 0.15 \veebar \eta = 0.5 \veebar \eta = 0.7) \wedge (d = 1 \veebar d = 3 \veebar d = 5)$ were evaluated. Results were averaged over 20 runs. Table 4 summarises the results in terms of the root-mean-squared error $\pm$ its standard deviation.

| | | $\alpha = 0.15$ | $\alpha = 0.5$ | $\alpha = 0.7$ |
|---|---|---|---|---|
| $d = 1$ | $\eta = 0.15$ | $2.236 \pm 0.171$ | $2.252 \pm 0.147$ | $2.207 \pm 0.113$ |
| | $\eta = 0.5$ | $2.203 \pm 0.133$ | $2.183 \pm 0.130$ | $2.264 \pm 0.160$ |
| | $\eta = 0.7$ | $2.198 \pm 0.103$ | $2.165 \pm 0.103$ | $2.204 \pm 0.133$ |
| $d = 3$ | $\eta = 0.15$ | $2.186 \pm 0.134$ | $2.235 \pm 0.144$ | $2.246 \pm 0.087$ |
| | $\eta = 0.5$ | $2.208 \pm 0.160$ | $2.277 \pm 0.113$ | $2.217 \pm 0.097$ |
| | $\eta = 0.7$ | $2.224 \pm 0.119$ | $2.191 \pm 0.089$ | $2.179 \pm 0.118$ |
| $d = 5$ | $\eta = 0.15$ | $2.273 \pm 0.104$ | $2.284 \pm 0.114$ | $2.243 \pm 0.140$ |
| | $\eta = 0.5$ | $2.245 \pm 0.116$ | $2.230 \pm 0.130$ | $2.273 \pm 0.172$ |
| | $\eta = 0.7$ | $2.240 \pm 0.120$ | $2.221 \pm 0.120$ | $2.204 \pm 0.079$ |

Table 4.: $\overline{rmse} \pm s_{rmse}$ in mmol/l as resulting from concurrent variation of learning variables $\eta$, $\alpha$ and $d$. Lower (better) values are darkened.

In this case, the variations did not lead to spectacular differences and its implications are not clear-cut. Most results are within a close margin of eachother. In general, performance seems to increase slightly when $\eta$ is on the higher end. The layer increase term $d$ should be used sparingly. From this data, no claims can be made about the use of momentum term $\alpha$, other than that it does not seem to affect the output much.

*Architecture*

An appropriate number of layers and neurons has to be found. To measure the influence that the amount of neurons has, the plot in figure 10 visualises the mean performance $\pm$ standard deviation of 40 runs of a neural network with one hidden layer, containing 1–100 neurons. The learning terms were set to $\eta = 0.6$, $\alpha = 0.3$, $d = 1.5$.

From this plot, it can be seen that performance does not change dramatically with network size. It is only with very small or relatively large networks that performance starts to worsen.

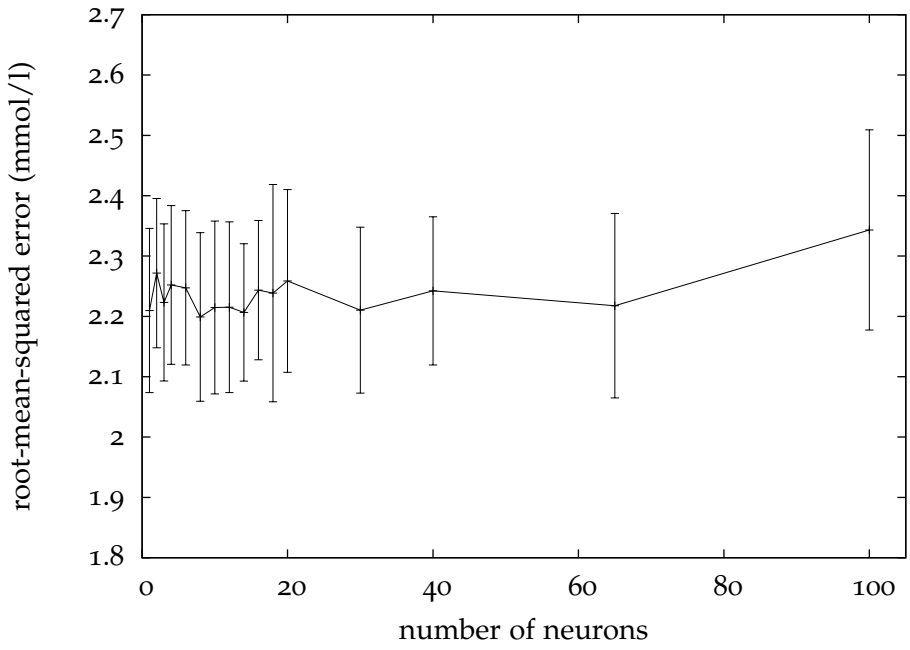


Figure 10.: Performance consequences of varying the amount of neurons in a single-hidden-layer ANN.

A cursory exploration of multi-hidden-layer architectures suggests that the performance displayed by networks of greater depth is similarly flat. This can be seen in table 5, which was derived from 40 runs of a network with $\alpha = 0.5$, $\eta = 0.5$, $d = 2$.

| | | hidden layer 1 | | |
|---|---|---|---|---|
| | | 8 | 15 | 30 |
| hidden layer 2 | 0 | $2.244 \pm 0.134$ | $2.193 \pm 0.134$ | $2.266 \pm 0.163$ |
| | 8 | $2.281 \pm 0.134$ | $2.262 \pm 0.117$ | $2.271 \pm 0.136$ |
| | 15 | $2.295 \pm 0.140$ | $2.255 \pm 0.121$ | $2.268 \pm 0.150$ |
| | 30 | $2.271 \pm 0.128$ | $2.258 \pm 0.114$ | $2.263 \pm 0.127$ |

Table 5.: $\overline{rmse} \pm s_{rmse}$ in mmol/l as resulting from concurrent variation of a multi-hidden-layer architecture. Lower (better) values are darkened.

The possibility cannot be ruled out that somewhere on these surfaces lies a 'perfect point' for which performance is wholly different from the points that were measured for figure 10 and table 5, but it was not deemed sufficiently promising to warrant further inspection.

For the same reason, no larger, more complicated (computationally expensive) network has been examined.

When multiple solutions are possible, the simplest should be chosen. Because observed performance was more or less the same for any network with less than tens of neurons, it is suggested that 1 hidden layer of only 8 neurons should be sufficient.

# 5 DISCUSSION

## 5.1 STRENGTHS

Some limited predictive power can be ascribed to the network, with a correlation coefficient that is in almost all cases $r > 0.6$ and sometimes $r > 0.7$.

The mean absolute error percentage is very similar to the one reported in Pappada et al, [33] with 19.4% in this report versus 22.7% in theirs. However, they base predictions on CGM data, which means that their predictive windows can be influenced and kept constant at will. With a predictive window of 50 minutes, their *mae* could get as low as 6.7%. It must also be noted that their data set contains many more hypoglycaemic episodes to drag down their performance, while Kok's data set contains very few. It is therefore a dubious practice to assign too much weight to comparisons between these measures.

While the mean *rmse* reported here is higher than the one reported for the AIDA system ($1.93 \pm 0.86$ mmol/l versus $2.16 \pm 0.13$ mmol/l), [1] our variability is much lower. The companion system rarely ever reports an *rmse* above 3 mmol/l, while AIDA's maximum *rmse* was 4.6 mmol/l. Again, the predictive window with which this AIDA's research dealt is not known and the data set was wholly different, so much care should be taken when interpreting these observations.

Comparisons between this and research of Sandham et al [38] and Zainuddin et al [53] are even harder to make due to lack of quantification.

The main improvement over Kok's research [23] was the use of $k$-fold cross-validation. This is believed to be the main contributor to the slight increase of performance: mean performance on the morning, afternoon, evening and night intervals of this report are at 2.15, 1.85, 2.42 and 2.16 mmol/l respectively, versus 2.3, 2.1, 2.4 and 2.3 mmol/l in the earlier report. Although these superior results occur fairly consistently, a side note has to be made that the improvements are of such magnitude that statistical flukes are not entirely unlikely.

Research on the same data set that was used in this report, tends to train different networks for each time slot. [53][23] This was justified by the hypothesis that there would not be a great correlation between the intervals. However, here we condense everything into one network so as to increase the available training data and to decrease dependence on the daily habits of the patient. Neither significant improvement nor significant deterioration was found to result from this decision.

Repeating the experiments over many instances and many data combinations unveiled the great influence of stochastical components. While this conclusion may not generalise to other data sets, it does emphasize the importance and difficulty of obtaining a sufficient data set. It also puts under greater scrutiny values that were reported earlier.

## 5.2 LIMITATIONS

*Medical*

Despite acceptable performance when compared to previous research, the error still seems too high to be of practical use in the near future. The improvement in *rmse* relative to the token error (obtained by always predicting the mean blood glucose) is only about 25.4%, and a worst-case *rmse* exists on the afternoon interval of over 2.4 mmol/l. Even then, these figures are only relevant to the particular data set that was used — in practice, it may be worse.

With these figures, it is not at all unlikely that a healthy, high-end predicted blood glucose value will correspond to a dangerously hypoglycaemic actual blood glucose level. Even in the best-performing ANN that was visualised in figure 3, a sample in the training data can be easily spotted that has a predicted blood glucose value of $> 10$ mmol/l and a reference blood glucose value of $< 4$ mmol/l. In a medical system, this is unacceptable.

The discontinuous nature of measurement points and predictive output means that sudden decrease or increase in the blood glucose level cannot be modelled, detected or prevented by the system. For example, a hypoglycaemic event that stabilises before the end of an interval will go unnoticed. This is a problem because it is known that in practice, significant changes do occur in time frames of less than 1 hour. [9]

Also consider the bias that may be introduced when patients self-monitor without steady rhythmicity: it seems probable that subjects will test more often when they are suffering from symptoms, e.g. mostly when blood glucose values are atypical. In addition, nocturnal monitoring of glucose is not possible without interrupting sleep or using a continuous glucose measuring device. This objection is not specific to the current examination. Continuous glucose monitoring may eliminate the bias that is introduced in this way.

More generally, as for any sub-symbolic system, it is difficult for humans to interpret the inner workings. This is problematic in a medical setting, because bad advice can lead to severe complications. It is hard to guard against them and it is unclear how to assign responsibilities.

Assumptions, simplifications and generalisations have had to be made, out of practical considerations such as deployability in a real life setting, and because of time constraints. Combined, this places another limit on the medical significance of the system.

Finally, it deserves mentioning that the process of recording data may be a very tedious process. If the system does not present a great improvement, it may not warrant the meticulous, labour-intensive data collection requirements that it imposes on the patient. In this

case, attention may be better focused on more automated approaches, such as those involving — again — continuous glucose monitoring devices.

### *Technical*

#### *Data*

The size of the data set was small. Data was extracted from a single person over a short period of time. This cannot necessarily be generalised to a larger population or a greater length of time. This is illustrated by the high dependence of the performance on the data subsets that were used for training and performance evaluation, as can be seen in section 4.1. This suggests that the sample size is too small to be treated as an accurate reflection of the general case.

This is supported by the fact that the number of hypoglycaemic episodes in the data set is staggeringly low, at 9 observations (when hypoglycaemia is defined as an end-of-interval blood glucose level at or below 4.0 mmol/l). This ties in to another problem: the goal of the network is to reduce hyper- and hypoglycaemic events to boundary cases. Since neural networks are to abstract general rules rather than exceptions, this presents a paradoxal difficulty: as dysglycaemia occurs less often in the data set, the network will become worse at its detection.

A related observation can be made on the difference in performance between intervals (see table 3). This difference may not be solely attributable to circadian rhythm, in which case it is possible that the daily rhythm of the particular subject facilitates or hampers prediction, or — again — that the sample size is too small to filter out the noise. Another contributing factor is the fact that duration varies between interval type (see chart 2).

It is noted that lack of sufficiently large data sets is one of the prime limits on data-driven models of diabetes. [4]

#### *Function complexity*

The fact that a very small network was selected to give the best results seems to imply that the function is too complex to be approximated from the limited amount of training data. A rudimentary network then gives the best results, in the same way that returning the mean error gives the best prediction when choice is limited to methods that return a constant value.

This idea gains credibility from the fact that the *rmse* of the training data set struggles to get below 1.5 mmol/l. This means that the training set can not be properly fitted into the ANN to produce an error score below some limit. Once again, this suggests that the underly-

ing function is too complex to be satisfyingly abstracted in neural networks of the sizes that were examined.

*Sub-optimality of input vector*

Although some experimentation on the composition of the input vector has been performed, the available information per data point is limited and it was not varied in any rigorous way to find a good combination. The reported results are therefore entirely based on 'best effort'. It is possible and likely that inclusion, exclusion, recombination or preprocession of some input component will yield superior results. It has been suggested that principal component analysis is of use in preprocessing the input so as to partially alleviate this problem, by automatically reducing the number of input attributes to only the most relevant among them. [53][11]

However, some input components were simply not present in the data, even though they are known to contribute to the glycaemic condition. For example, even though glycaemic response to a composite meal is not a simple sum of the types of carbohydrates that the meal contains, it is known that crude estimates based on such values as the glycaemic index or the glycaemic load can increase reliability of predictions. [8] The data required to evaluate this possibility was not available.

*Sub-optimality of parameters*

Similarly, the multitude of parameters (learning rate, momentum, etcetera) cannot feasibly be varied and tested in conjunction due to the combinatorial complexity of doing so. Therefore, only a small subset of configurations was tested. It is possible that a configuration with unusually high performance was overlooked in this way.

*Committee size*

In this report, whenever interest focuses on the relative difference between performances, and the output to specific samples is unimportant, the committee size $N$ is generally set to 1. This is because the reported performance measure is then already an average of multiple runs in which the weights were varied, alongside the partition distribution. Considering the strong influence on variability that data distribution has (see also section 4.1), it was deemed that averaging over many possible partitions was to take precedence over the gains that may be had from combining network output.

However, it is not guaranteed that the decrease in generalisation error that can be achieved by assembling a committee is the same across all configurations. This means that the relative difference may be off, and there may therefore be a chance — however slight — that acceptable parameter choices for $N = 1$ are not acceptable parameter

choices for $N = 40$. This is a concession that was made to save computation time.

*Back-propagation*

A more general issue of sub-optimality occurs in the backpropagation algorithm. It is only guaranteed to converge to some local minimum on the error surface, that is, the surface that is defined by the error of the network in terms of its weights. This problem is partially overcome through the addition of a momentum term, through the use of varying random weights at network initialisation, and through the use of stochastic rather than true gradient descent; however, no conclusive proof of optimality is provided, because not all possible weight vectors can be reached. [31, 104]

*Performance*

Errors are measured over one interval. This means, for instance, that a predictive window of 3 hours will be judged by the same standards as a one of 10 hours. Clearly, this introduces a problem with the concept of performance measurement, because there exists a large degree of chaos in the problem and predictability is expected to rapidly decrease as the time period gets larger. A fixed predictive window would give more meaning to the performance measure, but the necessary data for this was unavailable.

Since most of the variability in duration is is dependent on interval type (see chart 2), an acceptable solution is to break down the score for each interval as was done in table 3. However, since a high degree of variability remains, this eliminates the problem only partially. Furthermore, due to the smaller size of the sample against which performance is measured and due to the expected particularities of the patient, further noise will be introduced in the measure and it will be hard to assign much meaning to the resulting values.

*Performance comparison*

It is challenging to find a good performance measure for a system. There exists no standard against which to compare systems, and a good rating in one method does not necessarily mean that the system is superior to one with a lesser rating.

Additionally, measures tend to express general predictive power, while what is of most interest to diabetics is the accuracy of predictions *outside the euglycaemic range*, since these are the situations which are sought to avoid. For instance, a patient who expects a blood glucose level of 8 mmol/l when it is actually 4, is in less danger and has better sense of the body than a patient who predicts 4.5 mmol/l when it is 2.5 — even though the error in the latter case is smaller both in relative and in absolute terms. Analogously, a model that predicts

relatively poorly in a normal situation might still be preferable if it is relatively good at detecting boundary cases.

One approach to alleviating this problem may be to transform the network from a predictor into a classifier of normal, hypo- and hyperglycaemic situations. The performance may then be measured by some function of the number of correctly predicted dysglycaemic situations and their severity. A related approach, taken by Pappada et al, [33] is to evaluate performance in dysglycaemic situations seperately.

However, these approaches are very dependent on the defined boundaries for euglycaemia and on the availability of dysglycaemic data. The latter is particularly troubling, since the data set that was used here contained very limited instances of hypoglycaemia. As such, performance remains hard to compare.

The most important performance measure, however, is the comparison against human performance; after all, it is this performance that is to be improved. Automated predictions have their advantages, but patients have access to unquantifiable knowledge, like the "gut feeling" from subjective symptoms and a more complete integration of context information. It is known that patient blood glucose *estimation* is very inaccurate, [50] but it is not known how patients compare at the *prediction* task that the model aims to perform. Without this information, there exists no clear bar against which to measure the usefulness of the ANN.

*Learning rate variation*

The learning rate $\eta$ already varies with the layer depth. However, it has been suggested that gradual reduction of the value of the learning rate will reduce the risk of overstepping the minimum in the error surface. [31, 92] No such gradual decrease function is part of $\eta'_d$ because the amount of variables, and hence the complexity of the experiments, necessitated keeping constant the less dominating amongst them.

*Dealing with deficient data*

Noise and missing features are not only to be expected in the output of the network, but also in the input and target outputs. Neural networks are themselves robust and will work in spite of this. However, when the prevalence of unrecorded information is high across different components of the samples, or the level of uncertainty per component varies greatly, the effects on output error may be significant. [48]

The algorithm extrapolates from the data with which it is fed and it does not understand the meaning of the word 'unknown'. Since uncertainty is inevitable and patients cannot be expected to record their lives to the meticulous extent that would be required if deficiency was not to occur, it may be valuable to find the best way of codifying defi-

cient data so as to minimise its influence on output error. [48] Due to time constraints, this has not been pursued in the present research.

# 6 CONCLUSION

In this report, it was shown that some limited predictive capabilities for blood glucose level in type 1 diabetic patients ($r > 0.65$) can be gained from a very rudimentary feed-forward artificial neural network of a single hidden layer containing 8 neurons, when it is fed with data about time, previous blood glucose, carbohydrate ingestion, insulin intake, stress and physical activity.

With parameters set at a learning rate of $\eta = 0.5$ and a momentum of $\alpha = 0.5$, it exhibited a performance comparable to the performance for similar models that were mentioned in literature. Performance reports indicate a root-mean-squared error of $rmse = 2.156 \pm 0.131$ mmol/l, a mean absolute error percentage of $mae = 19.4\% \pm 1.3$ pp and a correlation coefficient of $r = 0.662 \pm 0.042$.

However, the meaning of these findings is debatable, since many different variables were to be integrated into the model from a relatively small data set: one set was available, containing only 308 records from a single patient, with predictive windows varying from 3 to 11 hours. Performance was measured on a subset 35% the size of the original data set. This may not generalise well. This observation was illustrated by the fact that the most potent contributor to variability existed within the data set itself; the particular neural network was of secondary importance. Therefore, care has to be taken when drawing conclusions.

Errors are as of yet likely far too large to have any application beyond educational purposes. While this is no reason to write off the use of artificial neural networks in glycaemic prediction, it is likely that more sophisticated training methods must be applied and — more urgently — that detailed, voluminous data sets are essential.

## 6.1 IMPLICATIONS FOR AI

On the other hand, the combination of a large number of variables and a small amount of data is also present in knowledge-based approaches to figuring out the individual glucose profile.

The fact that a very rudimentary model such as the one put forward in this report could closely approach the results obtained from a complicated physiological model like AIDA, emphasizes the power that artificial intelligence techniques have in problems such as these. It is believed that the complexity of the function makes abstraction from limited data very hard, but this is a blessing as well as a curse

because it enables the ANN to put to use a type of guesswork at which it excels.

## 6.2 RECOMMENDATIONS

It is unclear whether any practical size and level of detail for a data set will be large enough for neural network methods to give results that have clinical significance. For this reason, efforts should be focused either on methods that can abstract generalities from a smaller amount of data, or on methods that can generate more data to abstract from.

Other suggestions include: finding a better way to deal with deficiencies in the data set in the way suggested by Tresp et al [48], and finding a better selection of inputs. To reduce the number of inputs, an automatic process such as principal component analysis may be used as suggested by others. [53][11] Increase or preprocessing must be done through a process of speculation and trial-and-error — after all, before knowing the optimal selection of input and how to present it, the problem would need to be known to such an extent that it would not be a black box anymore, in which case a neural network would not be likely to be helpful. One such addition might be to include a more exact specification of meals, by way of glycaemic load or glycaemic index.

Finding better parameter values for the specific method that was used here should be a lesser concern; even though the values chosen here are sub-optimal, it is expected that the gains from perfecting the parameters to this inherently imperfect model will be marginal.

In any case, more data from multiple subjects are required to definitively judge performance of the current model and, by extension, the other models currently in existence.

# A

## BIBLIOGRAPHY


[1] AIDA. Aida v4 freeware diabetes software blood glucose simulator program – technical guide. `http://www.2aida.org/aida/technical.htm#i`, May 2014.

[2] Fiona S. Atkinson, Kaye Foster-Powell, and Jennie C. Brand-Miller. International tables of glycemic index and glycemic load values: 2008. *Diabetes Care*, 31(12):2281–2283, 2008.

[3] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[4] Troy Bremer and David A. Gough. Is blood glucose predictable from previous values? a solicitation for data. *Diabetes*, 48(3):445–451, 1999.

[5] S. F. Clarke and J. R. Foster. A history of blood glucose meters and their role in self-monitoring of diabetes mellitus. *British Journal of Biomedical Science*, 69(2):83–93, 2012.

[6] Philip E. Cryer. Prevention and correction of hypoglycemia. In Mark A. Sperling, editor, *Type 1 Diabetes: Etiology and Treatment*, volume 32, pages 145–161. Humana Press, Totowa, New Jersey, USA, 2003.

[7] Andrew P. Demidowich, Kevin Lu, Ronald Tamler, and Zachary Bloomgarden. An evaluation of diabetes self-management applications for android smartphones. *Journal of telemedicine and telecare*, 18(4):235–238, 2012.

[8] Hayley Dodd, Sheila Williams, Rachel Brown, and Bernard Venn. Calculating meal glycemic index by using measured and published food values compared with directly measured meal glycemic index. *The American Journal of Clinical Nutrition*, 94(4):992–996, 2011.

[9] Timothy C. Dunn, Richard C. Eastman, and Janet A. Tamada. Rates of glucose change measured by blood glucose meter and the glucowatch biographer during day, night, and around mealtimes. *Diabetes Care*, 27(9):2161–2165, 2004.

[10] D. S. Edgerton, M. Lautz, M. Scott, C. A. Everett, K. M. Stettler, D. W. Neal, C. A. Chu, and A. D. Cherrington. Insulin's direct effects on the liver dominate the control of hepatic glucose production. *J. Clin. Invest.*, 116(2):521–527, Feb 2006.

[11] A Karim El-Jabali. Neural network modeling and control of type 1 diabetes mellitus. *Bioprocess and biosystems engineering*, 27(2):75–79, 2005.

[12] Donna S. Eng and Joyce M. Lee. The promise and peril of mobile health applications for diabetes and endocrinology. *Pediatric Diabetes*, 14(4):231–238, 2013.

[13] S. Franc, S. Borot, O. Ronsin, J.-L. Quesada, D. Dardari, C. Fagour, E. Renard, A.-M. Leguerrier, C. Vigeral, F. Moreau, P. Winiszewski, A. Vambergue, H. Mosnier-Pudar, L. Kessler, S. Reffet, B. Guerci, L. Millot, S. Halimi, C. Thivolet, P.-Y. Benhamou, A. Penfornis, G. Charpentier, and H. Hanaire. Telemedicine and type 1 diabetes: Is technology per se sufficient to improve glycaemic control? *Diabetes & Metabolism*, 40(1):61 – 66, 2014.

[14] J. A. Galloway, C. T. Spradlin, R. L. Nelson, S. M. Wentworth, J. A. Davidson, and J. L. Swarner. Factors influencing the absorption, serum insulin concentration, and blood glucose responses after injections of regular insulin and various insulin mixtures. *Diabetes Care*, 4(3):366–376, 1981.

[15] Saul Genuth. Relationship between metabolic control and complications in diabetes: Therapeutic implications of the diabetes control and complications trial. In Mark A. Sperling, editor, *Type 1 Diabetes: Etiology and Treatment*, volume 32, pages 215–231. Humana Press, Totowa, New Jersey, USA, 2003.

[16] Eleni I Georga, Vasilios C Protopappas, and Dimitrios I Fotiadis. Glucose prediction in type 1 and type 2 diabetic patients using data driven techniques. *Knowledge-Oriented Applications in Data Mining*, 2011.

[17] Raeann Gifford. Continuous glucose monitoring: 40 years, what we've learned and what's next. *ChemPhysChem*, 14(10):2032–2044, 2013.

[18] Dennis E. Hinkle, William Wiersma, and Stephen G. Jurs. *Applied statistics for the behavioral sciences*. Houghton Mifflin Boston, 2003.

[19] Johanna Hortensius, Robbert J. Slingerland, Nanne Kleefstra, Susan J.J. Logtenberg, Klaas H. Groenier, Sebastiaan T. Houweling, and Henk J.G. Bilo. Self-monitoring of blood glucose: The use of the first or the second drop of blood. *Diabetes Care*, 34(3):556–560, 2011.

[20] Karsten Jungheim and Theodor Koschinsky. Glucose monitoring at the arm: Risky delays of hypoglycemia and hyperglycemia detection. *Diabetes Care*, 25(6):956–960, 2002.

[21] D. Kerr, E. Cheyne, P. Thomas, and R. Sherwin. Influence of acute alcohol ingestion on the hormonal responses to modest hypoglycaemia in patients with type 1 diabetes. *Diabetic Medicine*, 24(3):312–316, 2007.

[22] Morwenna Kirwan, Corneel Vandelanotte, Andrew Fenning, and Mitch J. Duncan. Diabetes self-management smartphone application for adults with type 1 diabetes: randomized controlled trial. *Journal of medical Internet research*, 15(11), 2013.

[23] Peter Kok. Predicting blood glucose levels of diabetics using artificial neural networks. Master's thesis, Delft University of Technology, Delft, NL, 2004.

[24] Anders Krogh, Jesper Vedelsby, et al. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, pages 231–238, 1995.

[25] E.D. Lehmann and T. Deutsch. A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. *Journal of Biomedical Engineering*, 14(3):235 – 242, 1992. Annual Scientific Meeting.

[26] Nutri Life. Nutrilife diabetes management – android-apps op google play. `https://play.google.com/store/apps/details?id=com.NutriLifeStyle.NutriLife`, April 2014.

[27] Patrick J. Lustman, Ray E. Clouse, Paul S. Ciechanowski, Irl B. Hirsch, and Kenneth E. Freedland. Depression-related hyperglycemia in type 1 diabetes: A mediational approach. *Psychosomatic Medicine*, 67(2):195–199, 2005.

[28] maroonfolder. Glycemic load lite – android-apps op google play. `https://play.google.com/store/apps/details?id=com.maroonfolder.glycemicloadlite`, April 2014.

[29] Marilyn Y. McNiven Temple, Oded Bar-Or, and Michael C. Riddell. The reliability and repeatability of the blood glucose response to prolonged exercise in adolescent boys with iddm. *Diabetes Care*, 18(3):326–332, 1995.

[30] P. Micossi, M. Cristallo, M. C. Librenti, G. Petrella, G. Galimberti, M. Melandri, L. Monti, P. Spotti, M. Scavini, V. Di Carlo, and G. Pozza. Free-insulin profiles after intraperitoneal, intramuscular, and subcutaneous insulin administration. *Diabetes Care*, 9(6):575–578, 1986.

[31] Tom Michael Mitchell. *Machine learning*. McGraw-Hill, 1997.

[32] F. Ovalle. Clinical approach to the patient with diabetes mellitus and very high insulin requirements. *Diabetes Research and Clinical Practice*, 90(3):231 – 242, 2010.

[33] Scott M. Pappada, Brent D. Cameron, and Paul M. Rosman. Development of a neural network for prediction of glucose concentration in type 1 diabetes patients. *Journal of diabetes science and technology*, 2(5):792–801, 2008.

[34] L. C. Perlmuter, B. P. Flanagan, P. H. Shah, and S. P. Singh. Glycemic control and hypoglycemia: is the loser the winner? *Diabetes Care*, 31(10):2072–2076, Oct 2008.

[35] José C. Principe, Neil R. Euliano, and W. Curt Lefebvre. *Neural and adaptive systems: fundamentals through simulations*. John Wiley & Sons, Inc., 1999.

[36] Tristan Richardson, Melanie Weiss, Peter Thomas, and David Kerr. Day after the night before: Influence of evening alcohol on risk of hypoglycemia in patients with type 1 diabetes. *Diabetes Care*, 28(7):1801–1802, 2005.

[37] Maria C.E. Rossi, Antonio Nicolucci, Paolo Di Bartolo, Daniela Bruttomesso, Angela Girelli, Francisco J. Ampudia, David Kerr, Antonio Ceriello, Carmen De La Questa Mayor, Fabio Pellegrini, David Horwitz, and Giacomo Vespasiani. Diabetes interactive diary: A new telemedicine system enabling flexible diet and insulin therapy while improving quality of life: An open-label, international, multicenter, randomized study. *Diabetes Care*, 33(1):109–115, 2010.

[38] W. A. Sandham, Dimitra Nikoletou, D. J. Hamilton, K. Patterson, Alan Japp, and Catriona MacGregor. Blood glucose prediction for diabetes therapy using a recurrent artificial neural network. In *Conf Proc EUSIPCO*, volume 11, pages 673–6, 1998.

[39] Baiju R. Shah and Bernard Zinman. Insulin regimens for type 1 diabetes. In Mark A. Sperling, editor, *Type 1 Diabetes: Etiology and Treatment*, volume 32, pages 199–213. Humana Press, Totowa, New Jersey, USA, 2003.

[40] Nancy F. Sheard, Nathaniel G. Clark, Janette C. Brand-Miller, Marion J. Franz, F. Xavier Pi-Sunyer, Elizabeth Mayer-Davis, Karmeen Kulkarni, and Patti Geil. Dietary carbohydrate (amount and type) in the prevention and management of diabetes: A statement by the american diabetes association. *Diabetes Care*, 27(9):2266–2271, 2004.

[41] Jennifer Sherr, Dongyuan Xing, Katrina J. Ruedy, Roy W. Beck, Craig Kollman, Bruce Buckingham, Neil H. White, Larry Fox, Eva Tsalikian, Stuart Weinzimer, Ana Maria Arbelaez, William V. Tamborlane, and for the Diabetes in Children Network (DirecNet). Lack of association between residual insulin production

and glucagon response to hypoglycemia in youth with short duration of type 1 diabetes. *Diabetes Care*, 36(6):1470–1476, 2013.

[42] Carmel E.M. Smart, Megan Evans, Susan M. O'Connell, Patrick McElduff, Prudence E. Lopez, Timothy W. Jones, Elizabeth A. Davis, and Bruce R. King. Both dietary protein and fat increase postprandial glucose excursions in children with type 1 diabetes, and the effect is additive. *Diabetes Care*, 36(12):3897–3902, 2013.

[43] S. S. Soedamah-Muthu, S. Abbring, and M. Toeller. Diet, lifestyle and chronic complications in type 1 diabetic patients. In Chich-Pin Liu, editor, *Type 1 Diabetes: Complications, Pathogenesis, and Alternative Treatments*, pages 21–44. InTech, Rijeka, Croatia, 2011.

[44] Cézar Roberto Souza. The accord.net framework. `http://accord.googlecode.com/`, April 2014.

[45] Kenneth Strauss, Heidi De Gols, Irene Hannet, Tuula-Maria Partanen, and Anders Frid. A pan-european epidemiologic study of insulin injection technique in patients with diabetes. *Practical Diabetes International*, 19(3):71–76, 2002.

[46] Cajsa Tonoli, Elsa Heyman, Bart Roelands, Luk Buyse, StephenS. Cheung, Serge Berthoin, and Romain Meeusen. Effects of different types of acute and chronic (training) exercise on glycaemic control in type 1 diabetes mellitus. *Sports Medicine*, 42(12):1059–1080, 2012.

[47] Joseph Tran, Rosanna Tran, and John R White. Smartphone-based glucose monitors and applications in the management of diabetes: An overview of 10 salient "apps" and a novel smartphone-connected blood glucose monitor. *Clinical Diabetes*, 30(4):173–178, 2012.

[48] Volker Tresp, Subutai Ahmad, Ralph Neuneier, et al. Training neural networks with deficient data. *Advances in neural information processing systems*, pages 128–128, 1994.

[49] Eve Van Cauter, Kenneth S. Polonsky, and André J. Scheen. Roles of circadian rhythmicity and sleep in human glucose regulation. *Endocrine Reviews*, 18(5):716–738, 1997. PMID: 9331550.

[50] Katie Weinger, Alan M. Jacobson, Matthew T. Draelos, Dianne M. Finkelstein, and Donald C. Simonson. Blood glucose estimation and symptoms during hyperglycemia and hypoglycemia in patients with insulin-dependent diabetes mellitus. *The American Journal of Medicine*, 98(1):22 – 31, 1995.

[51] D. Randall Wilson and Tony R. Martinez. The general inefficiency of batch training for gradient descent learning. *Neural Networks*, 16(10):1429–1451, 2003.

[52] Hannele Yki-Järvinen, Eero Helve, and Veikko A. Koivisto. Hyperglycemia decreases glucose uptake in type i diabetes. *Diabetes*, 36(8):892–896, 1987.

[53] Zarita Zainuddin, Ong Pauline, and Cemal Ardil. A neural network approach in predicting the blood glucose level for diabetic patients. *International Journal of Computational Intelligence*, 5(1), 2009.