AI-based estimation of the duration of reliability demonstration tests for components

Philipp Mell[a*], Martin Dazer[a]
[a]Institute of Machine Components, University of Stuttgart, Stuttgart, Germany

Abstract: In the product development process, demonstrating product reliability is crucial. While failure-based tests, or end-of-life tests, provide optimal results when testing components, predicting their duration is challenging since it depends on randomly occurring failures. This unpredictability contrasts with failure-free tests, where the duration appears to be predictable.

Efforts have been made to estimate the duration of failure-based reliability tests using prior knowledge about the failure mechanism. If information like the time-to-failure distribution is available, the test duration can be estimated through simulation. An earlier study showed that simulation can often be avoided by using an analytic equation or simple numerical convolution. This drastically lowers the bar for applying such test duration estimations, since the effort for both programming and running the simulation is avoided. However, these simlified approaches only perform well in specific test scenarios, leaving many cases where a proper substitute for simulations is missing. For these cases, a machine learning model is suggested in this paper.

An artificial neural network (ANN) model is trained and tested using simulated data for various products and failure modes. First, the best ways to describe the duration of failure-based tests with minimal parameters are identified. Second, the relationship between test duration parameters and boundary conditions is modeled through simulation. The data is then analyzed and divided into subsets for training individual ANNs. Lastly, the performance of the ANNs is assessed and compared, with special emphasis on the competence region of each network. The paper concludes with recommendations on when to use analytical or machine-learning-based estimation methods for specific reliability test scenarios.

Keywords: reliability demonstration test, test duration estimation, test planning, artificial neural network.

1. INTRODUCTION

Ensuring the reliability and longevity of technical products is essential. Therefore, reliability demonstration tests (RDTs) are carried out with components in the product development phase. The process of planning an RDT involves choosing from an extensive collection of test types, variants, and configurations. The selection of the most suitable RDT is directed by the trade-off between the quality of the test result, the duration of the test and its cost. The topic of test quality – i.e., the likelihood that the test will meet particular lifetime and reliability requirements – is studied most extensively [1-4]. Recent publications on this topic suggest using the probability of test success (P_{ts}) to specify how likely a particular test plan is to succeed [1-4]. This concept allows the evaluation and comparison of various RDTs, significantly advancing the generalization of test result quality. In particular, it boosts the ability to compare failure-free and failure-based tests [4-5]. Failure-based component tests generally give better results as long as the product is not drastically oversized. Excessive oversizing, however, violates sustainability goals, which are becoming increasingly relevant.

Research specifically focusing on the expected duration or cost of failure-based RDTs is scarce, with only few recent publications [6-7]. However, both quantities play an important role in everyday management decisions. In this context, a key role is attributed to the test duration, as many time-dependent cost factors are linked to the test duration [6]. This problem apparently does not exist for failure-free tests, for which the test duration is fixed from the beginning. For this reason, failure-free tests are still being favored in many companies – despite the significant disadvantages mentioned above. In addition, the better test duration predictability of failure-free tests is only apparent: unintended failures are quite likely and will compromise both the test result and the assumed test duration.

Predicting the duration of a failure-based RDT is challenging: First of all, the test duration is stochastically distributed, just like the single failure times of the test specimens. Secondly, several influencing factors must be considered [7]. The failure behavior of the product is of utmost relevance. As for all test planning, some prior knowledge about the failure behavior must exist. This knowledge is usually not exact, but subject to uncertainty. Operational boundary conditions, such as the number of available test rigs, available prototypes, or maximum test duration, also influence the choice of the optimal test plan. Finally, monetary constraints, often represented as the total budget available for the test, must be taken into account.

An introductory proposal to enable a prediction of the duration of failure-based RDTs in the face of all these challenges was made in [7]. With the approaches described there, a simple solution for the test duration prediction in serial tests (only a single test rig) and parallel tests (one test rig for each test specimen) is available. The presented approaches allow a highly efficient prediction with negligible error (RMSE < 1.5 %) without simulation. All other tests, referred to as “mixed tests”, are only predictable with higher error and in a limited parameter range. Therefore, a special approach is required for mixed tests.

As a possible remedy, the training and application of artificial neural networks (ANNs) is considered in this paper. The theoretical fundamentals of the considered failure-based RDTs, their test duration distributions and the terms used in this paper are presented in the next section. A short summary of the current state of research is given in Section 3, thereby defining the boundaries for this paper. Section 4 outlines the methodical framework by which the test duration distribution is to be estimated, including a subdivision of the considered RDTs. Section 5 gives a detailed look upon the used ANNs, the training process and the training and test data used. The results in terms of the performance of the generated ANNs are presented & analyzed in Section 6, leading to a list of key findings. Section 7 contains a conclusion of the outcomes. Based on that, an exemplary application is shown in Section 8, outlining the accuracy and efficiency of the presented test duration estimation approach. The paper closes with a summary and an outlook.

2. THEORETICAL BACKGROUND, CLASSIFICATION AND TERMINOLOGY

2.1 General remarks

The purpose of an RDT is to demonstrate that a product or system will achieve a specified lifespan with a certain level of reliability and confidence. For this, specimens of the final product are tested to gather data empirically. This data may either be precise failure times or a time interval (censored information). An RDT can either aim for the failure of all specimens (failure-based), the survival of all specimens (failure-free) or a mixture of both (either failure-based testing with censoring, or failure-free testing with permitted failures). The tests may be accelerated, e.g. by increasing the load relevant to the failure mechanism under consideration (accelerated life testing, ALT), or through degradation testing.

2.2 Boundary conditions

Reliability engineers must select an appropriate RDT based on their specific test situation, which depends on several boundary conditions:

(a) Failure distribution: The failure behavior of the product generally has to be characterized by a probability distribution for the considered failure mechanism. In case of several failure mechanisms, it is suggested to focus on the primary one (i.e., the one which leads to the earlies or the most failures). This gives a conservative estimate of the test duration. As failure distribution, the Weibull distribution with shape parameter b and scale parameter T is assumed. A failure-free time t_0 can be included by simply adding it to the lifetime of each tested unit. Effective RDT planning requires at least some (uncertain) knowledge of these parameters, which can be derived from product experts, previous tests, predecessor products, or analogous products in the same company or in literature. Therefore, it is assumed that prior knowledge about the distribution of the primary failure mechanism exists.

(b) Physical resources: The RDT is performed with n_u test units on n_r parallel test rigs. These physical resources are limited by the maximum number of units available ($n_{u,max}$) and the maximum number of test rigs that can operate simultaneously ($n_{r,max}$).

(c) Financial and temporal resources: Usually, there is a limit on the test cost (c_{max}) and on the timeframe ($t_{dur,max}$) available for the RDT. It is the goal of this paper to allow a proper estimate of the test duration t_{dur} of different RDTs to enable reliability engineers to make better test planning decisions. This turns out to be challenging, as the test duration generally follows an unknown probability distribution. As the test cost c mainly depends upon the number of used test units, test rigs and the test duration, the test cost estimation is trivial once the test duration has been estimated.

The goal of this paper therefore is to present an approach to estimate the duration of failure-based RDTs in feasible time when prior knowledge about the failure behavior of a product is given (with uncertainty). This way, reliability engineers will be equipped to weigh the pros and cons of different possible RDTs in terms of test result, test duration and test cost for their individual testing scenario.

2.3 Failure-free vs. Failure-based RDTs

Although failure-free RDTs are popular in many development departments because of their seemingly simple planning, they often turn out to be a poor choice [7]. Their heaviest disadvantages are the susceptibility to unintended failures – rendering the test planning invalid – and the fact that they only give a minimum reliability, encouraging oversizing and undermining sustainability goals [7]. The intended duration of failure-free RDTs is easily calculated with a single equation [7], which is why this case is not further considered here. Failure-based RDTs, on the other hand, try to obtain the maximum statistical information by determining the exact failure time of the tested units. As these failure times are stochastically distributed, so is the test duration of failure-based RDTs. To limit the test duration, some of the tested units can be suspended before their failure, which is called censoring. Two basic censoring schemes can be applied:

(a) Type I / time censoring: Each unit is tested for a maximum duration of t_c. If it didn't fail up until then, it is suspended and the RDT continues with the next unit.

(b) Type II / unit censoring: The number of failing units is limited. After the n_f-th failure, the test is aborted and the remaining $n_c = n_u - n_f$ units are suspended.

Failure-free RDTs can be seen as failure-based RDTs with very early censoring time (type I) or all units being censored (type II). Failure-based RDTs are the thus the more general case and are focused on in this paper.

2.4 Testing process

The following process for a failure-based RDT is assumed: To each of the n_r test rigs, one test unit is assigned and the test is started. If a unit fails or reaches the censoring time t_c (in time censored tests), it is suspended and replaced with a new test unit (as long as untested units remain). With the failure or suspension of the n_u-th unit (n_f-th unit in unit censored tests), the RDT ends. An example without censoring is given in Fig. 1.

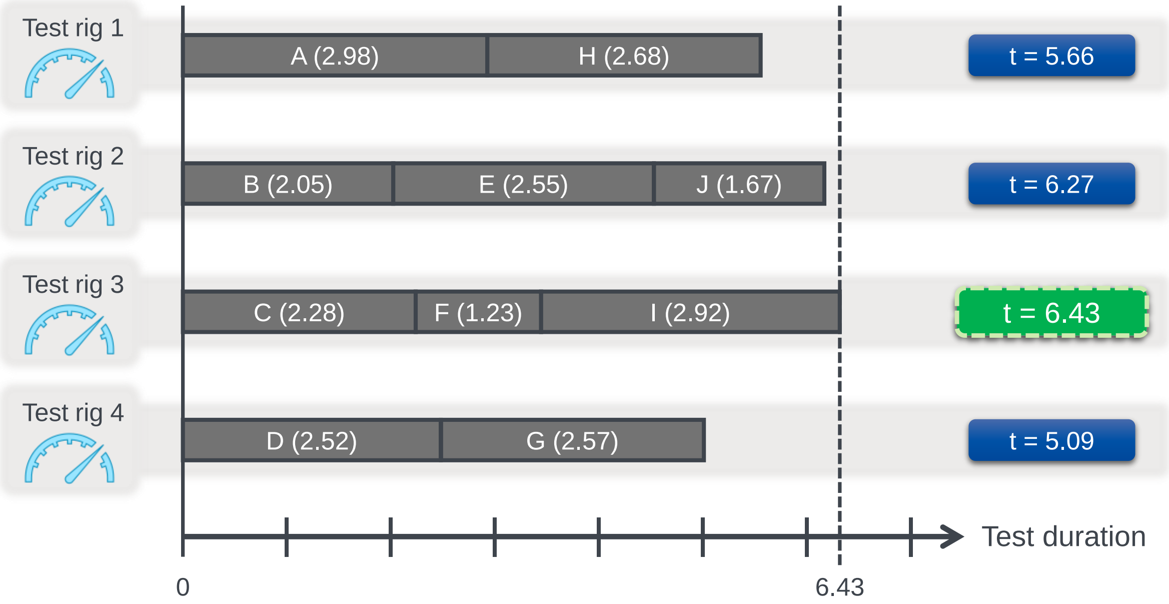

Figure 1. Exemplary uncensored reliability demonstration test (RDT) with $n_u = 10$ units (labeld A-J) on $n_r = 4$ parallel test rigs. Time (in arbitrary units) advances from left to right. Each time a unit fails, it is immediately replaced with a new unit. The lifetime of each unit is given in parentheses. The total test duration (6.43) is defined by the test rig which finished last – in this case test rig 3.

2.5 Assumptions and simplifications

It is assumed that prior knowledge about the failure distribution parameters T and b exists. This prior knowledge may either be an exact value, or the distribution moments of an estimate. For example, b might be given in terms of its mean estimate $\hat{b}$ and its variance $\mathrm{Var}(b)$ from a maximum likelihood estimation (ML) estimation. The characteristic lifetime T scales all lifetime quantities, including the test duration. Normalizing all time quanitites by T thus allows to ignore T in the upcoming considerations without loss of generality.

As ALT is the standard in many test scenarios, a load dependency of the product lifetime – and hence the estimated test duration – can be included through its effect on T. A typical assumption in ALT is that elevated load does not affect the distribution shape, making this approach permissible [2-4].

Regarding the test plan, it is assumed that either time censoring or unit censoring is applied, but not both simultaneously. In unit censored tests, only $n_r > 1$ and $n_c < n_r$ has to be considered [7].

2.6 Test case classification

To systematically structure all possible RDTs, they are divided into three disjunct cases. Perfectly serial tests on a single test rig ($n_r = 1$) and perfectly parallel tests with as many test rigs as units ($n_u = n_r$) can be considered analytically [7]. All other cases (i.e., $n_u > n_r > 1$) are a mixture of both and are hereafter referred to as mixed tests.

3. STATE OF RESEARCH

The first approach for the estimation of the duration of a failure-based RDT that comes to mind is a simulation. By application of the Monte Carlo method with an appropriate number of iterations – 1e5 have found to be sufficient [7] –, the test duration distribution can be determined empirically. However, these simulations tend to be lengthy, especially when uncertain prior knowledge and different possible test plans have to be considered. Furthermore, they are prone to errors in the programming process. After all, a practical solution for everyday use should be straightforward. It has thus been studied if analytical or numerical methods can be used to replace simulations without relevant error [7].

As performance measures, the median (50 % quantile) and higher quantiles of the test duration distribution come to mind. While the median characterizes the general location of the distribution, higher quantiles give an application focused conservative estimate of the test duration. In the following, the 90 % quantile of the test duration distribution will be used, which gives the test duration which is not exceeded in 9 out of 10 cases.

Observing the percentual error in the 50 % and the 90 % quantile, it has been found that non-simulative approaches are well applicable in many relevant cases of failure-based RDTs. Specifically, perfectly serial and perfectly parallel tests give an RMSE of ≤ 0.1 % for all cases but time censored serial tests, where the RMSE is 1.3 %. Virtually all considered parameter combinations lie within an error range of $\pm$ 10 % [7].

For mixed tests, however, the found analytical methods are only applicable in a limited parameter range with acceptable error. Only when restricting the failure mode shape parameter to $b \geq 1.5$ and the ratio between failing units and test rigs to values ≥ 2.5 can the RMSE be limited to 5 %. However, in around 5 % percent of the considered parameter combinations, the error is still greater than 10 %. Another approach is therefore necessary for mixed tests.

Furthermore, uncertainty in the prior knowledge regarding the failure distribution parameters T and b. This is straightforward in a simulation through Bootstrapping, however it also increases the effort exponentially. For non-simulative test duration estimation, a different approach needs to be used.

4. METHODICAL FRAMEWORK

The ultimate goal is to estimate the duration t_x of an RDT in the worst x % of cases. To achieve this goal for mixed tests, two approaches are considered:

(a) General approach: the stochastical distribution of the test duration is estimated, allowing to obtain the test duration quantile for any percentage x. This approach is to be preferred due to its universality.

(b) Specific approach: only specific, pre-defined quantiles x of the test duration distribution are estimated. This approach is necessary if no smooth test duration distribution emerges, or if the distribution cannot be properly fitted with a known distribution. As predefined test distribution quantiles, the longest test out of two (t_{50}) and the longest test out of ten (t_{90}) are chosen.

Both approaches are carried out by training ANNs to estimate the distribution parameters (general approach) or the test duration quantiles (specific approach). In the first case, the 50 % and 90 % quantiles, t_{50} and t_{90}, are subsequently calculated from the inverse cumulative distribution function (icdf) of the fitted distribution.

As is known from previous studies, an insufficiently smooth test duration distribution is imminent in the case of time censored tests [7]. Therefore, the following considerations are separated by censoring scheme in uncensored, time censored and unit censored tests. While the general approach is applied for every censoring scheme, the specific approach is additionally used for time censored tests. Some exemplary test duration distributions from simulation and their distribution fits and quantiles are shown in Figure 2. It can be seen there that in different cases, different distribution fits are superior. Depending on the chosen parameters, the data is either right-skewed or close to normal. For time censored tests, distinct spikes at multiples of the censoring times can be observed. However, as long as the distribution doesn't degenerate into a single value, the distribution fits still give surprisingly exact estimates for time censored tests.

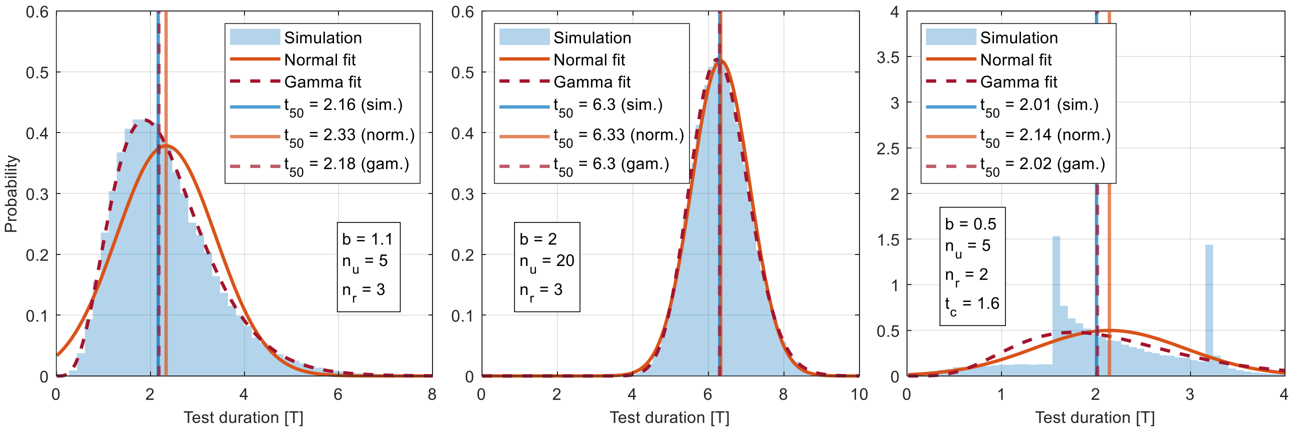

Figure 2. Test duration distribution of 3 exemplary uncensored (left, center) and time censored (right) RDTs with given parameters. The simulation data (blue) stems from a Monte Carlo simulation with 1E6 repetitions, sorted into 60 bins and acts as reference. A normal fit (orange) and a gamma fit (red, dashed) to the data is given, together with values for the 50 % quantile t_{50} (median).

5. NEURAL NETWORK TRAINING

For both approaches presented in Section 4, several ANNs are trained and compared. Some parameters are fixed for all ANNs. Firstly, only shallow neural networks with 2 or 3 hidden layers are considered, since previous studies have shown that these are most efficient for estimations connected to Weibull distributed failure data [2]. As training function, the Bayesian Regularization (BR) algorithm is used, which does not require a validation data set, but instead penalizes additional weights and biases in the ANN to prevent overfitting [8-9]. Each dataset used is divided into training data and test data 70:30 randomly. Training stops after a maximum of 800 epochs, or if the variable weighting factor of the BR algorithm indicates that an optimum has been reached. Each ANN is trained 5 times to avoid effects of the data division on the performance. To avoid unwanted weighting of the inputs and outputs due to different value ranges, they are all mapped to the interval [-1, +1]. As performance function during training, the mean square error (MSE) is used. As training data, simulations of the total test duration for 181,484 combinations of failure distribution and RDT are used. The covered parameter range per censoring scheme is given in Table 1. Since only mixed tests are to be estimated, parameter combinations with $n_r = n_u$ (parallel tests) have been filtered out.

Table 1. Parameter ranges for the considered combinations of failure distribution and the RDT configuration, separated by censoring scheme. The last row gives the available data used for the ANN training and test.

	EoL	**Time censored**	**Unit censored**
Weibull shape b	0.5 – 5		
Total units n_u	3 – 60		
Parallel test rigs n_r	2 – 15		
Censoring time t_c	-	0.2 T – 0.8 T	-
Censored units n_c	-	-	1 – 15 $(n_r > n_c, n_u - n_c \geq 3)$
# of simulated test cases	14,420	63,744	103,320

A hyperparameter tuning is performed with the ANN parameters shown in Table 2. While rectified linear units (ReLU) as activation function led to quick convergence and low training times, the results were badly generalized. This could be an effect of the stepwise covered parameter space of the training data. Therefore, the better generalizing sigmoid activation function (hyperbolic tangent) is used. The usage of only two hidden layers has led to better results than three layers. A high number of neurons per layer proved to be more effective. Thus, two hidden layers with 50 and 20 neurons are being used.

Table 2. ANN parameters which were considered in the hyperparameter tuning and chosen as optimal

<table>
<tr><th colspan="2">Hyperparameters</th><th colspan="6">Considered values</th><th>Chosen values</th></tr>
<tr><td rowspan="3">ANN parameters</td><td># of hidden layers</td><td colspan="6">1, 2, 3</td><td>2</td></tr>
<tr><td># of neurons per layer</td><td colspan="6">4, 8, 10, 16, 20, 30, 32, 50</td><td>50 (1st), 20 (2nd)</td></tr>
<tr><td>Activation function</td><td colspan="3">ReLU
(Rectified linear unit)</td><td colspan="3">Sigmoid
(hyperbolic tangent)</td><td>Sigmoid
(hyperbolic tangent)</td></tr>
<tr><td rowspan="3">Inputs / Outputs</td><td>Input quantities for the number of tested units</td><td colspan="2">n_u</td><td colspan="2">n_u, m_2</td><td colspan="2">m, m_2</td><td>m, m_2 (relative)</td></tr>
<tr><td>Additional input F_f for time censored RDTs</td><td colspan="3">Yes</td><td colspan="3">No</td><td>No</td></tr>
<tr><td>Distributions</td><td colspan="6">Normal, Lognormal, Gamma, None</td><td>various, see Sec. 6</td></tr>
</table>

Apart from the typical ANN hyperparameters, different input and output data have also been considered, as is shown in Table 2. Regarding the inputs, the number of test rigs n_r is undisputed. Regarding the number of tested units, however, theoretical approaches suggest that instead of the actual number n_u, the relative number

$$m = n_u / n_r \tag{1}$$

might be of higher influence [7]. If m is no integer, the number of remaining units,

$$m_2 = \mathrm{mod}(n_u, n_r) \cdot n_r, \tag{2}$$

is necessary as well. While ANNs are able to derive these quantities themselves, it requires unnecessary computation and therefore additional hidden layers, neurons and training. Therefore, it was checked in the hyperparameter training if it is beneficial to use m and m_2 instead of n_u as input quantities. The results proved that this is the case (not shown here). Combinations of n_u and m_2, however, performed worse. For time censored RDTs, it was studied if an additional input with the probability of a unit actually failing,

$$F_f = 1 - \exp\left(-\left(\frac{t_c}{T}\right)^b\right), \tag{3}$$

is beneficial. The results were ambiguous, but the effect was low. Therefore, the smaller input parametrization is used and F_f is not handed over as input to the ANN. As ANN outputs, the parameters of the considered distribution (in case of the general approach) or the 50 % and 90 % distribution quantiles (in case of the specific approach) are used. Because of their wide range of values, using the natural logarithm of the distribution quantiles and the gamma distribution parameters as ANN output showed to give better results.

6. RESULTS

By training ANNs with the different hyperparameters presented in the previous section, dozens of neural networks capable of estimating the test duration have been obtained. In the following, the performance of the best ANNs for each censoring scheme is presented. For this selection, only ANNs with minimum test data error and with comparable performance in their 5 training repetitions were considered. The results are given in Table 3 and are visualized – for the exemplary case of uncensored tests – in Figure 3.

As performance metric, the relative error ε on all test data in the relevant range of $\beta \geq 1$ is used. The percentage of test data falling below $\pm$ 5 %, 2 %, and 1 % is calculated. Also, the root mean square of the relative error (RMSRE) is calculated to give an impression which average percentual error is to be expected. It is desired that the RMSRE be below 1 %. The presented performance data belongs to the best of these repetitions. As $R^2 \geq 99$ % in all cases, this quantity is not informative and therefore not shown.

As RDTs are unreasonable for $b < 1$, the according test data have been omitted for obtaining the performance metrics shown in Table 3 and Figure 3. The training data, however, included such cases when the distribution based general approach was used. One might argue that removing this data from the training set would lead to more specialized ANNs with higher performance in the relevant range of $b \geq 1$. This hypothesis was examined and could be confirmed for the specific approach, which was only used for time censored tests. For the general approach, however, the performance changes are negligible or even negative. A possible explanation is that the additional training data for $b < 1$ helped the ANNs to learn the highly nonlinear influence of the shape parameter over its whole range. Therefore, for $b \leq 1$ in, separate ANNs have been trained for the specific approach of time censored tests. Their resulting performance is slightly lower (RMSRE of 0.297 and 0.340 for t_{50} and t_{90}).

Table 3. Performance on all test data with $b \geq 1$ of the optimum ANNs per censoring scheme and test duration quantile. The RMSRE gives the mean percentual deviation from the true value. "Log-transformed" indicates that the ANN outputs are the log of the distribution parameters or the test duration quantiles, respectively. For uncensored RDTs, t_{90} can be estimated slightly better assuming a gamma distribution instead of a lognormal distribution (t_{50} by the gamma distribution is inferior and thus omitted). For time censored RDTs, estimating t_{50} and t_{90} directly gives better results than assuming a normal distribution.

Censoring scheme & test duration quantile		**Share of test data with… [%]**			**RMSRE [%]**	**Assumed distribution (if applicable)**
		$\varepsilon \leq 5$ %	$\varepsilon \leq 2$ %	$\varepsilon \leq 1$ %		
No censoring	t_{50}	100	99.64	97.22	0.358	Lognormal
	t_{90}	100	99.74	98.97	0.261	
	t_{90}	100	100	99.18	0.244	Gamma, log-transformed
Time censoring	t_{50}	100	99.94	98.63	0.274	None (Direct quantile estimation, log-transformed)
	t_{90}	100	99.95	99.13	0.237	
	t_{50}	99.12	93.78	86.20	1.107	Normal
	t_{90}	99.24	93.40	83.93	1.071	
Unit censoring	t_{50}	100	99.96	98.80	0.274	Gamma, log-transformed
	t_{90}	100	99.96	99.53	0.227	

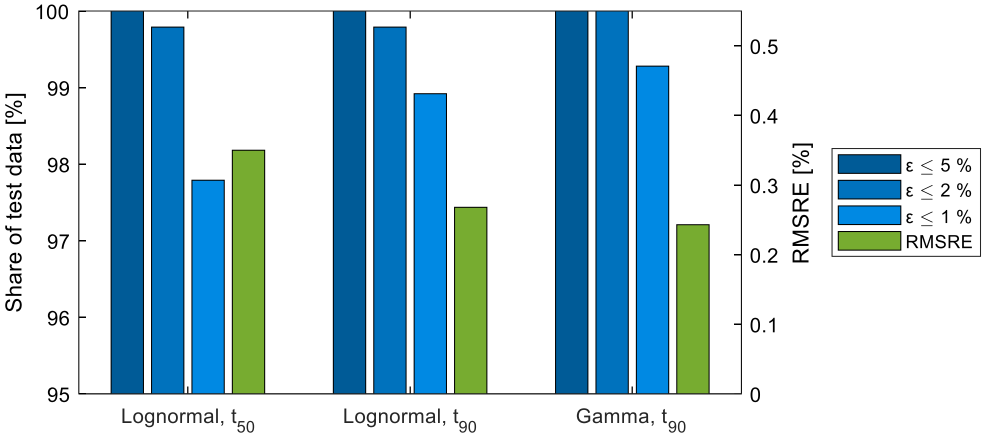

Figure 3. Performance comparison of the two best ANNs for estimating the duration quantiles of uncensored tests, visualized by means of the share of test data falling below some error limit ε (blue, left axis) and the rms value of ε, RMSRE (green, right axis). The shown ANNs assume a lognormal and a gamma distribution of the test duration, respectively. As the estimates for the 50 % quantile, t_{50}, are inferior when assuming a gamma distribution, only the results for t_{90} are shown here.

Analyzing the performance metrics in Table 3, the following general statements can be made:

- The duration distribution of uncensored and unit censored tests can be estimated with higher accuracy than that of time censored tests with the general approach. This agrees with earlier studies, which have shown that the discontinuous characteristics of time censored tests make them harder to estimated [7].
- The 90 % quantile of the test duration, t_{90}, can generally be estimated with higher accuracy than the 50 % quantile, t_{50}. A possible explanation is that the shape of the test duration pdf in the central area varies more, and is hence more difficult to estimate than the right tail. As higher quantiles like t_{90} is more conservative and therefore more relevant, this is assessed beneficial from a practical perspective.
- It is possible to estimate the test duration distribution with an expected error of < 1 % in every censoring scheme by application of the optimal estimation approach.
- A maximum error of 5 % is maintained in every censoring scheme when the specific approach is used for time censored tests.

Apart from the results shown here, the error of the trained ANNs has also been studied for correlation with any of the parameters. A weak correlation is found for uncensored and unit censored tests. There, t_{50} is estimated with higher error when the shape parameter b is low, and that t_{90} is estimated with higher error when the shape parameter is high. A weaker correlation with opposite direction is found for the number of test rigs, n_r.

7. CONCLUSION

The estimation of the duration of uncensored, time censored and unit censored RDTs is possible with high accuracy using properly trained ANNs. Such ANNs feature no more than two hidden layers, but need sufficiently many neurons in each layer. A combination of 50 and 20 neurons in the two layers gave best results. As activation function, the hyperbolic tangent was able to give much smoother (i.e., better generalized) estimates than the ReLU activation function. The rms value of the estimation error is well below 1 % for all censoring schemes and can be limited to 5 % in the worst case if the optimal approach is used.

Additionally, the best probability distribution to approximate the test duration distribution has been found:

- Without censoring, the test duration is best described with a lognormal distribution. If high test duration quantiles like t_{90} are of particular interest, a gamma distribution fits even better.
- For time censoring, a normal distribution gives the best fit if the general approach is desired. However, the specific approach gives significantly better results.
- The duration of unit censored RDTs is best approximated by a gamma distribution.

With the optimal distributions and ANNs, it is possible to assess how uncertain prior knowledge affects the estimated test duration, which is shown in the next section. and then estimating the desired test duration distribution with the chosen ANN.

8. APPLICATION WITH UNCERTAIN PRIOR KNOWLEDGE

The test duration estimation with the obtained ANNs is applied to a realistic problem to assess its accuracy and efficiency. A product with a single failure mechanism is considered. The failure distribution is Weibull, and uncertain prior knowledge about its parameters exists from prototype tests. According to this prior knowledge, the characteristic lifetime T and the shape parameter b have a mean and a standard deviation of

$$\begin{aligned} \mu_T = 30\ h, \qquad \sigma_T = 5\ h; \\ \mu_b = 1.9, \qquad \sigma_b = 0.3. \end{aligned} \tag{4}$$

As both parameters must be positive, a lognormal distribution of both parameters is assumed, which is calculated based on the moments in eq. (4) and is depicted in Figure 4 (left and center).

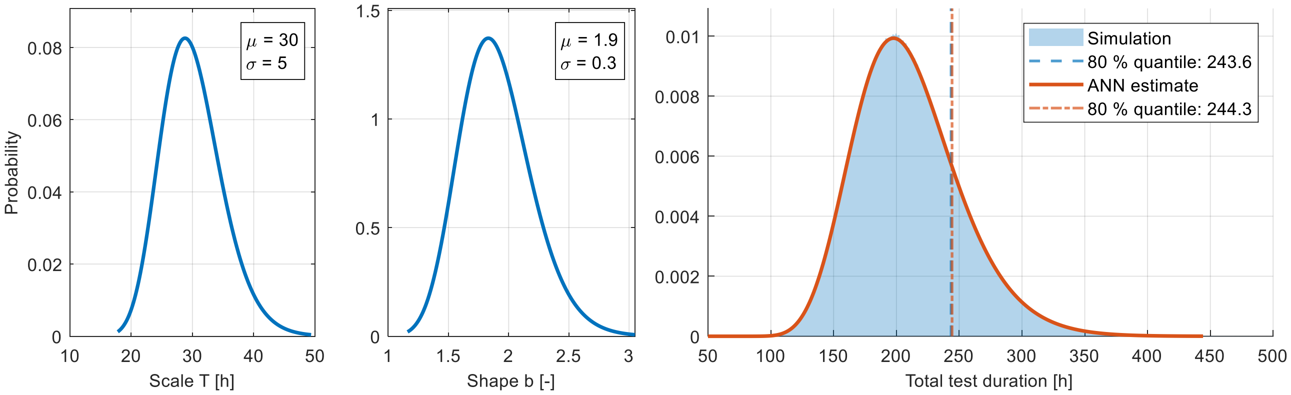

Figure 4. For given uncertain knowledge of the failure parameters T (left) and b (center) with given mean μ and standard deviation σ from prototype tests, the test duration distribution on the right arises. The test configuration was $n_u = 29$ units on $n_r = 4$ parallel test rigs without censoring. The blue values are obtained via simulation with the Monte Carlo method, the red line is estimated with the obtained ANNs.

In this example, the aim is to plan an uncensored RDT for this product. 29 test units and 4 parallel test rigs are available. The question is how long the RDT will take with 80 % confidence when considering different numbers of units and rigs from the available maximum numbers. Previously, a two-level simulation using the Monte Carlo method would have been used for this task, sampling parameters T and β in an outer loop and repeatedly performing the chosen RDT in an inner loop. Now, the test duration distribution can be estimated with ANNs in each of these cases. When all 29 test units and 4 test rigs are used, the resulting test duration is given in Figure 4 (right). The ANN approximation of the maximum test duration deviates by less than 0.3 % from the simulation. To estimate all other relevant numbers of used test rigs and test units, 80 different RDTs must be considered. By simulation, this takes 74.4 min on the reference system. The ANN approximation takes 42.2 s. This is an acceleration by 2 orders of magnitude. It should be kept in mind here that in practical application, drastically more than only 80 RDTs might have to be considered, which makes the simulation approach poorly suited for efficient and user-friendly use.

9. SUMMARY & OUTLOOK

This paper studied the accuracy and efficiency of an ML based approach to estimate the test duration of failure-based RDTs given uncertain prior knowledge. It has been shown that by using ANNs to approximate the test duration distribution, this task can be fulfilled with an expected error of less than 1 %. The optimal distribution fits for each censoring scheme have been found. A hyperparameter training allowed to narrow down the range of reasonable ANN parameters. In contrast to the previous state of the art, a repeated simulation of the considered RDTs, an acceleration by 2 orders of magnitude can be achieved.

In summary, it is now possible to estimate the duration of failure-based RDTs with manageable effort. This evens out the major disadvantage over failure-free RDTs. The new method now has to demonstrate its applicability in practical use cases. On the one hand, this requires interlinking with existing methods for the test duration estimation, e.g. those presented in [7]. On the other hand, the test duration estimation has to be embedded in a holistic methodology for the planning of failure-based RDTs. To this end, a combination with estimation methods for the demonstrated lifetime has to take place. The appropriate application will show whether further improvements to the ANNs presented in this paper are necessary. If so, more specialized parameter regions according to the found error correlation are an obvious starting point.

Acknowledgments

The authors gratefully acknowledge the assistance of Maximilian Schrempf, who provided preparatory material for this paper as part of his research thesis at the Institute of Machine Components, University of Stuttgart.

References

[1] Dazer, M.; Stohrer, M.; Kemmler, S.; Bertsche, B.: Planning of reliability life tests within the accuracy, time and cost triangle, *2016 IEEE Accelerated Stress Testing & Reliability Conference (ASTR)*, 28.-30.09.2016, Pensacola Beach, FL, USA.

[2] Mell, P.; Karle, F.; Herzig, T.; Dazer, M.; Bertsche, B.: Accelerating Optimal Test Planning With Artificial Neural Networks, *2022 Annual Reliability and Maintainability Symposium (RAMS)*, 24.-27.01.2022, Tucson, AZ, USA, DOI: 10.1109/RAMS51457.2022.9893938.

[3] Bertsche, B.: Reliability in Automotive and Mechanical Engineering, Springer, Berlin, Heidelberg, Germany, 2008, DOI: 10.1007/978-3-540-34282-3.

[4] Grundler, A.; Dazer, M.; Herzig, T.: Statistical Power Analysis in Reliability Demonstration Testing: The Probability of Test Success, *Applied Sciences, 12, no. 12, 6190*, 2022. DOI: 10.3390/app12126190.

[5] Dazer, M.; Grundler, A.; Benz, A.; Arndt, M.; Mell, P.: Pitfalls of Zero Failure Testing for Reliability Demonstration, *Proceedings of the 32nd European Safety and Reliability Conference (ESREL), 1639–1646*, Research Publishing, Singapore, 2022.

[6] Arndt, M.; Dazer, M.: Analysis of Efficiency in Response Surface Designs Considering Orthogonality Deviations and Cost Models, *Proceedings of the 33rd European Safety and Reliability Conference (ESREL), 1167-1174,* Research Publishing Singapore, 2023.

[7] Mell, P.; Dazer, M.: Estimating the duration of failure-based reliability demonstration tests, *2024 Annual Reliability and Maintainability Symposium (RAMS)*, 22.-25.01.2024, Albuquerque, NM, USA, DOI: 10.1109/RAMS51492.2024.10457820.

[8] Foresee, F. D..; Hagan, M. T.: Gauss-Newton Approximation to Bayesian Learning, *Proceedings of International Conference on Neural Networks (ICNN'97),* Houston, TX, USA, vo. 3, pp. 1930-1935, 1997, DOI: 10.1109/ICNN.1997.614194

[9] MacKay, D. J. C.: Bayesian Interpolation, *Neural Computation*, vol. 4 (3), 415–447, 1992. DOI: 10.1162/neco.1992.4.3.415