INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Automatic Evaluation of Children Reading Aloud on Sentences and Pseudowords

Jorge Proença[1,2], *Carla Lopes*[1,3], *Michael Tjalve*[4], *Andreas Stolcke*[5], *Sara Candeias*[6], *Fernando Perdigão*[1,2]

[1] Instituto de Telecomunicações, Coimbra, Portugal
[2] Department of Electrical and Computer Engineering, University of Coimbra, Portugal
[3] Polytechnic Institute of Leiria, Leiria, Portugal
[4] Microsoft & University of Washington, Seattle, WA, USA
[5] Microsoft AI and Research, Mountain View, CA, USA
[6] Microsoft, Digital Advisory Services, Lisbon, Portugal

{jproenca,calopes,fp}@co.it.pt, {michael.tjalve,andreas.stolcke,v-sacand}@microsoft.com

Abstract

Reading aloud performance in children is typically assessed by teachers on an individual basis, manually marking reading time and incorrectly read words. A computational tool that assists with recording reading tasks, automatically analyzing them and providing performance metrics could be a significant help. Towards that goal, this work presents an approach to automatically predicting the overall reading aloud ability of primary school children (6-10 years old), based on the reading of sentences and pseudowords. The opinions of primary school teachers were gathered as ground truth of performance, who provided 0-5 scores closely related to the expectations at the end of each grade. To predict these scores automatically, features based on reading speed and number of disfluencies were extracted, after an automatic disfluency detection. Various regression models were trained, with Gaussian process regression giving best results for automatic features. Feature selection from both sentence and pseudoword reading tasks gave the closest predictions, with a correlation of 0.944. Compared to the use of manual annotation with the best correlation being 0.952, automatic annotation was only 0.8% worse. Furthermore, the error rate of predicted scores relative to ground truth was found to be smaller than the deviation of evaluators' opinion per child.

Index Terms: reading level assessment, child speech, Gaussian process regression

1. Introduction

Evaluating the reading aloud proficiency of primary school children is usually an effort done by teachers or tutors, where they provide a level-appropriate reading task to a child and manually take notes for accuracy and time. This task cannot be accomplished very often during a school year as it can be very time consuming and needs to be accomplished in a 1-on-1 setting. A system that can automatically perform these functions could be an important supplement for teachers and allow assessments to be more frequent, as well as less affected by evaluator bias. An overall reading performance score can provide a clear overview of a child's level and be used to track progress over time.

This work focuses on providing an overall reading aloud performance score automatically, targeting children in primary school at 6 to 10 years of age. It should be emphasized that the comprehension of what is being read is not measured, and the concern is solely for oral reading fluency. However, overall reading competence is linked to oral reading fluency [1], and the latter can be defined as the ability to read text quickly, accurately and with proper expression [2], [3]. Techniques to assess fluency are often connected to second language learning [4], [5], targeted at adults or young adults for whom speech technologies are significantly mature. In fact, the methods for automatic reading assessment may also be used to detect specific reading disorders or in automatic reading tutors where a child's reading is tracked in real time, falling in the area of computer assisted language learning (CALL). Examples of projects that aimed to create reading tutors include LISTEN [6], Tball [7], SPACE [8] and FLORA [9].

Reading aloud by children often exhibits disfluencies that affect reading speed and accuracy. A common metric for reading performance is Correct Words per Minute (CWPM) [10] that considers reading speed of only correctly pronounced words. We develop several more features based on reading speed, task difficulty and frequencies of pauses and disfluencies and verify how well they correlate with the opinion of school teachers for a score of overall reading ability of several children. Furthermore, these features can be combined (with multi-feature regression techniques) to provide even closer results, since the distinct information given by them may correspond to the separate cues that influence a teacher's score.

Similar works in this field automatically detect correct words and common reading mistakes [7], [11]–[14], but reading performance assessments focus mostly on reading of isolated words. Black et al. [11] aimed to automatically evaluate reading ability and provide a high-level literacy score for individual word reading tasks. Using automatically extracted features and a selection of features based on pronunciation, fluency and speech rate, a Pearson correlation of 0.946 was

http://dx.doi.org/10.21437/Interspeech.2017-1541

achieved to predict mean evaluator's scores (from 1 to 7). Duchateau et al. [12] also target the reading of isolated words. They evaluate a child's reading ability by the number of correctly read words divided by time spent (equivalent to CWPM) and show agreement to human evaluation with Cohen's Kappa [15] above 0.6 when considering 5 performance classes. Our work will focus on the reading of sentences and pseudowords of varying difficulty, to hopefully encompass a wider set of reading competencies.

Section 2 will introduce the data of European Portuguese children reading aloud for reading evaluation. Section 3 describes the process of obtaining and parsing ground truth scores of reading performance from teachers. The considered features relating to reading speed, disfluencies and word difficulty are presented in Section 4. The regression models and feature selection stages to predict overall reading performance score are detailed in Section 5.

2. Dataset of children reading aloud

A subset of the LetsRead database [16] of European Portuguese children reading is used for this evaluation. We considered 150 children (6-10 years old), with the frequency distribution along the grades (1 to 4) being 43, 40, 35 and 32. From each child, utterances of 5 sentences and 5 pseudowords are extracted, to a total of 2h36m of audio. Originally, each child read 20 sentences and 10 pseudowords, but it was previously deemed that the shorter subset provides similar evaluations of performance, allowing more children to be evaluated during the ground truth effort (or more evaluations per child).

The sentence prompts were extracted from children's tales and school books for the target group. Pseudowords (such as <traba> [tɾˈabɐ], <impemba> [ĩpˈẽbɐ] or <culenes> [kulˈɛnəʃ]) represent non-existing or nonsense words in the native language, used to evaluate morphological and phonemic awareness. Pseudowords of 2 to 4 syllables were created by shuffling the most common syllables in a lexicon of European Portuguese, maintaining full pronounceability [16].

3. Ground truth of reading performance

We have gathered the ratings of primary school teachers as a ground truth for reading performance, which will be the target of the trained regression models. A crowd-sourcing like application was developed, distributed to primary school teachers throughout Portugal. Each evaluator, after listening to 5 sentences and 5 pseudowords of a child, gave a performance score between 0 and 5, closely related to the expected reading level of grades 1-4, with 0 and 5 for extreme cases. 10 groups of evaluators evaluated different sets of 15 children (for a total of 150 children), with each group having an average of 10 evaluators (7 minimum, 13 maximum), for a total of 100 evaluators. Therefore, calculations of evaluator performance will be done inside its group. Each child was assessed by at least 7 evaluators.

3.1. Evaluating evaluators

To measure agreement between evaluators, we can compute Pearson's correlation of the 15 scores given by an evaluator to the 15 scores of another evaluator who assessed the same children, repeating for all evaluators of the same group. This pairwise correlation reflects the agreement between evaluators and can be used to identify 'bad' evaluators. We found clear outliers that did not agree with other evaluators and decided to remove any evaluators below 0.65 correlation (iteratively computing new correlations after removing the worst outliers). Table 1 presents average results after 5 evaluators were removed. We also compute correlation to the mean, i.e., the correlation of each evaluator to the average of other evaluators of the same group, providing similar conclusions.

Table 1: *Overall mean and standard deviation of evaluator pairwise correlation and correlation to the mean of other evaluators.*

Correlation	Mean ± S.D.	Maximum	Minimum
Pairwise	0.796 ± 0.060	0.885	0.657
To the mean	0.874 ± 0.069	0.967	0.679

3.2. Normalized scores

An evaluator can have certain biases: i) constantly giving lower scores than the average ones; ii) constantly giving higher scores than the average ones; iii) constantly giving scores near the minimum and maximum; or iv) constantly giving scores near the middle. Applying a z-normalization (z-norm) is a method to remove these effects and, for each evaluator, their scores are changed by subtracting their mean and dividing by their standard deviation. All values are then reconstructed to the original scale by multiplying by the overall standard deviation of all 1500 scores and adding the overall mean. This is an alternative to scaling the minimum and maximum to 0 and 5 and can provide values slightly lower than 0 or higher than 5. The mean of the transformed scores per child is the value taken as ground truth of reading performance score.

4. Features

There may be several sources of information that teachers consider when deciding about overall reading aloud performance. We explore different types of features, analyze their performance individually in predicting performance score, and use a combination of them to train regression models. Two distinct sources for feature extraction will be considered: manual annotation (where several types of disfluencies are marked) and automatic annotation (where mispronunciations, false starts, repetitions, intra-word pauses and extensions are marked). Although using manual annotation may provide the purest analysis of which features matter for evaluation, it is automatic feature annotation that is needed to build a practical system for performance evaluation.

4.1. Automatic annotation

The employed methods to automatically annotate an utterance while detecting reading disfluencies follow two stages, and are described in further detail in [17]. The first is segmentation, providing candidate segments for words. Task specific decoding lattices (per utterance) are used where the sequence of the original prompt's words is the most probable, with additional arcs that allow repetitions of words and false starts based on each word's syllables. A 4% word error rate is achieved, without considering whether a word is pronounced correctly. In the second stage, candidate segments for words are classified as mispronounced or not. The classification is based on log-likelihood ratios between the reference word and the output of a phonetic recognizer.

4.2. Feature description

Table 2 lists the set of features considered. The same features are calculated for the two reading tasks separately (sentences and pseudowords), doubling the number of features. Features can be split into four groups: reading speed (1-6), silence related features (7-9), rate of disfluencies (10-14) and task meta-information (15-18), where difficulty is based on a set of rules for doubtful or harder pronunciations [16]. Since deletions and insertions are not detected by the automatic annotation, FastR is not computed for that case. Sentence features will be addressed with the prefix 's' and pseudoword features with the prefix 'p' (e.g., s1, s17, p1, p17).

4.3. Individual feature performance

To fit a single feature to the ground truth ratings, a simple linear regression model can be trained to minimize the sum of squared errors, resulting in a linear transformations:

$$\hat{y} = aX + b \tag{1}$$

where $\hat{y}$ is the predicted output, X is the observation vector, a is the coefficient (weight) of the input and b is the intercept (bias) term. For a multi-feature regression, a is a vector of weights and X is a matrix of feature vectors. To evaluate the fit of predicted scores to the ground truth, two performance metrics will be considered: Pearson's correlation coefficient (ρ or Corr) and root mean squared error (RMSE). We will consider a leave-one-out cross-validation with 150 folds to train and test regression models, where 149 samples are used to train a model and 1 is left out for testing. The 150 resulting test values are aggregated to compute Corr and RMSE. Table 2 shows results using each feature individually for linear regression. Random performance leads to a correlation coefficient of 0 and RMSE of about 1.9.

The overall best feature for both manual and automatic methods is s6: correct characters per minute in sentences (CCPM). Other reading speed metrics that do not depend on disfluencies (s1, s3 and s5) perform well but slightly worse than their counterparts using correctly read units. The opposite is observed for pseudoword features, where the best one was p5: characters per minute of the original prompt (CharsPM). This may be due to the poor performances on pseudoword reading, where values of 0 correct words per minute could be found, and the time it took to read them may have more information. It should be noted that the results obtained with features 1, 3 and 5 could be dependent on the conditions of our data. They may only be relevant features because the reading tasks were almost always completed, with very few early stoppings. For a live application, features 2, 4, 6 should probably be preferred. Automatic features performed slightly worse, probably due to disfluency detection errors. Figure 1 displays the best manual features against the ground truth, where there is evidence of a linear fit.

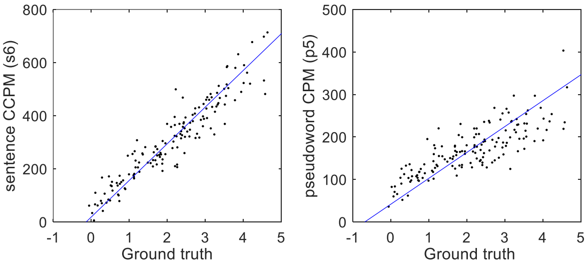

Figure 1: *Best sentence feature s6 (left) and best pseudoword feature p5 (right) from manual annotation versus ground truth, with respective linear fits.*

5. Models of overall performance score

Although some of the individual features already seem to be good predictors of reading aloud ability, it is expected that combining several of them will further approach the ratings by

Table 2: *Feature enumeration and performance of using each feature for a linear regression model to predict ground truth scores.*

		Sentences					Pseudowords				
Features			Manual		Auto			Manual		Auto	
Abbr.	Description	Feat.	Corr	RMSE	Corr	RMSE	Feat.	Corr	RMSE	Corr	RMSE
WPM	Words per minute (prompt)	s1	0.919	0.459	0.917	0.463	p1	0.744	0.776	0.744	0.776
CWPM	Correct words per minute	s2	0.928	0.434	0.923	0.447	p2	0.674	0.858	0.670	0.863
SyllsPM	Syllables per minute (prompt)	s3	0.927	0.435	0.926	0.439	p3	0.764	0.750	0.760	0.755
CSPM	Correct syllables per minute	s4	0.938	0.402	0.930	0.429	p4	0.684	0.848	0.684	0.848
CharsPM	Characters per minute (prompt)	s5	0.931	0.424	0.930	0.428	p5	**0.805**	**0.689**	**0.803**	**0.693**
CCPM	Correct characters per minute	s6	**0.940**	**0.397**	**0.931**	**0.425**	p6	0.703	0.827	0.691	0.840
SILrate	Total silence / Total time	s7	0.647	0.885	0.736	0.787	p7	0.324	1.099	0.397	1.067
SILini	Average initial silence time	s8	0.347	1.091	0.480	1.019	p8	-0.157	1.176	-0.130	1.176
SILiniRate	Initial silence time / Total time	s9	0.283	1.115	0.231	1.132	p9	-0.073	1.173	0.005	1.169
MispR	Rate of mispronunciations	s10	0.525	0.991	0.615	0.916	p10	0.389	1.070	0.494	1.011
ExtraR	Rate of repetitions + false-starts	s11	0.498	1.008	0.555	0.968	p11	0.139	1.154	0.236	1.130
SlowR	Rate of extensions + intra-word pauses	s12	0.540	0.978	0.328	1.098	p12	0.247	1.127	-0.172	1.192
FastR	Rate of insertions + deletions	s13	0.171	1.146	N/A	N/A	p13	0.092	1.160	N/A	N/A
DisfR	Rate of all disfluencies	s14	0.663	0.872	0.683	0.850	p14	0.490	1.013	0,530	0,985
nSylls	Total number of syllables	s15	0.456	1.034	0.456	1.034	p15	0.147	1.151	0.519	0.993
nChars	Total number of characters	s16	0.483	1.018	0.483	1.018	p16	0.361	1.084	0.147	1.151
Diff1	Difficulty 1 – Pronunciation rules	s17	0.493	1.011	0.493	1.011	p17	0.464	1.029	0.361	1.084
Diff2	Difficulty 2 – Rules and length	s18	0.491	1.012	0.491	1.0123	p18	0.462	1.031	0.464	1.029

teachers. For this purpose we explore several regression models: multi-feature linear regression (LR), Gaussian process regression (GPR), and neural networks with one linear output (NN). GPR trains kernel-based probabilistic models to predict continuous values and confidence intervals of a score can be calculated. It is especially useful to avoid overfitting [18]. The GPR models were trained using a squared exponential kernel as the covariance function. A simple neural network for regression was used with one hidden layer followed by a linear unit, providing linear outputs in the evaluation's range. Since the amount of data for training is low (149 samples maximum), we obtained the best results using only a single perceptron in the hidden layer. The networks were trained with Bayesian regularization to improve generalization [19].

To tackle overfitting to the training data, feature selection methods are also employed. Stepwise regression can decide which features to include or remove iteratively for a regression model [20]. A feature is added to a linear model if the change in sum of squared errors (SSE) from its inclusion is statistically significant (in this case, a *p*-value of an *F*-statistic test lower than 0.05). No features were ever removed after inclusion. We consider the trained linear regression and also take the selected features to train GPR and NN models. Another regression method that can be used for feature selection is the least absolute shrinkage and selection operator (LASSO), a regularization technique applied here for a regularized least-squares regression [21]. SSE is minimized but with constraints on the sum of absolute values of coefficients of features, where many features will have weights of 0 and the others can be selected to apply LR, GPR and NN.

Table 3 summarizes the test results after training the considered regression models with leave-one-out cross-validation using features obtained either from manual or automatic annotations. The first models use the best sentence feature (s6) and the best pseudoword feature (p5). For manual features, NN proved to be the best model after a stepwise feature selection and the improvement to only using the best features is clearer in RMSE. Results were slightly worse with automatic features and, contrarily, the best results are without feature selection and the best model was GPR.

Table 3: *Performance of multi-feature regression models using manual and automatic features.*

	Manual		Auto	
Model	**Corr**	**RMSE**	**Corr**	**RMSE**
LR (s6)	0.940	0.397	0.931	0.425
LR (s6,p5)	0.943	0.386	0.933	0.419
GPR (s6,p5)	0.948	0.371	0.938	0.403
LR all	0.926	0.442	0.931	0.426
GPR all	0.947	0.375	**0.944**	**0.388**
NN all	0.938	0.403	0.939	0.399
Stepwise add + LR	0.947	0.373	0.919	0.458
Stepwise add + GPR	0.949	0.367	0.932	0.422
Stepwise add + NN	**0.952**	**0.357**	0.925	0.442
LASSO	0.944	0.387	0.932	0.423
LASSO + LR	0.942	0.392	0.932	0.421
LASSO + GPR	0.942	0.392	0.939	0.400
LASSO + NN	0.948	0.368	0.935	0.411

In stepwise selection with automatic features different selections were made across the folds, while in the manual case, the same features were consistently selected. This may explain why feature selection was especially helpful in the manual case. When using all features, GPR seems to be the approach generalizing best. Nevertheless, analyzing which features were most often selected can provide an important insight. For both manual and automatic approaches, s6 (CCPM) and p17 (Diff1) are consistently selected. For manual only, s13 (FastR) and p1 (CWPM) are also selected. For automatic only, s3 (SyllsPM) and p14 (DisfR) often appear. This shows that the reading of both sentences and pseudowords was relevant to predict the ground truth ratings. Although reading speed of pseudowords did not appear for the automatic approach, the rate of all detected disfluencies (p14) does. Figure 2 plots the predictions from GPR using all automatic features along with the ground truth scores, with the 95% confidence interval of GPR and standard deviation of GT. Apart from a couple of outliers, most of the data and predictions fall inside those intervals.

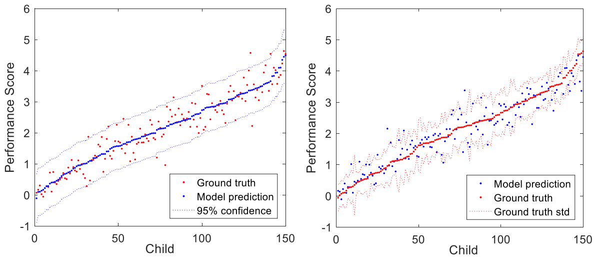

Figure 2: *Ground truth and predictions with 95% confidence of GPR model (left, prediction sorted) and ground truth standard deviation (right, GT sorted).*

6. Conclusions

Aiming to predict the ratings given by primary school teachers for overall reading aloud performance of children, we developed and combined features automatically obtained from automatic annotation. Although a good correlation is already achieved with typical metrics such as correct words per minute, a significant improvement was obtained with additional features. Gaussian process regression was the best performing regression approach when using automatic features, and seems to be especially suitable to avoid overfitting when the entire set of features is used.

Since metrics based on both sentence reading and pseudowords reading tasks were usually chosen during feature selection, we may conclude that teachers gave their overall ratings based on both tasks. Although reading speed metrics are the most important overall, detecting disfluencies proves to be relevant as well, even of specific types of disfluencies. Difficulty also seems to be an important normalizing factor, especially for pseudowords, since their difficulty is consistently chosen during feature selection. Further work should focus on investigating which unconsidered factors lead to some of the outliers that fell outside confidence intervals or ground truth standard deviation.

7. Acknowledgments

This work was supported in part by Fundação para a Ciência e Tecnologia under the project UID/EEA/50008/2013 (pluri-annual funding in the scope of the LETSREAD project at Instituto de Telecomunicações). The authors acknowledge the support given by Microsoft to this project. Jorge Proença is supported by the SFRH/BD/97204/2013 FCT Grant.

8. References

[1] L. S. Fuchs, D. Fuchs, M. K. Hosp, and J. R. Jenkins, “Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis,” *Scientific Studies of Reading*, vol. 5, no. 3, pp. 239–56, 2001.

[2] H. C. Buescu, J. Morais, M. R. Rocha, and V. F. Magalhães, “Programa e Metas Curriculares de Portugês do Ensino Básico,” Ministério da Educação e Ciência, May 2015.

[3] National Reading Panel, “Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction,” National Institute of Child Health and Human Development, USA, 2000.

[4] S. M. Abdou *et al.*, “Computer aided pronunciation learning system using speech recognition techniques,” in *Proc. Interspeech 2006*, Pittsburgh, USA, 2006, pp. 849–852.

[5] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, “Automatic pronunciation scoring of words and sentences independent from the non-native’s first language,” *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, Jan. 2009.

[6] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, “A Prototype Reading Coach That Listens,” in *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, Menlo Park, CA, USA, 1994, pp. 785–792.

[7] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, “Automatic detection and classification of disfluent reading miscues in young children’s speech for the purpose of assessment,” in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007, pp. 206–209.

[8] J. Duchateau *et al.*, “Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules,” *Speech Communication*, vol. 51, no. 10, pp. 985–994, Oct. 2009.

[9] D. Bolaños, R. A. Cole, W. Ward, E. Borts, and E. Svirsky, “FLORA: Fluent Oral Reading Assessment of Children’s Speech,” *ACM Trans. Speech Lang. Process.*, vol. 7, no. 4, p. 16:1–16:19, Aug. 2011.

[10] J. Hasbrouck and G. A. Tindal, “Oral reading fluency norms: A valuable assessment tool for reading teachers,” *The Reading Teacher*, vol. 59, no. 7, pp. 636–644, 2006.

[11] M. P. Black, J. Tepperman, and S. S. Narayanan, “Automatic Prediction of Children’s Reading Ability for High-Level Literacy Assessment,” *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 4, pp. 1015–1028, May 2011.

[12] J. Duchateau, L. Cleuren, H. V. hamme, and P. Ghesquière, “Automatic assessment of children’s reading level,” in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1210–1213.

[13] E. Yilmaz, J. Pelemans, and H. V. hamme, “Automatic assessment of children’s reading with the FLaVoR decoding using a phone confusion model,” in *Proc. Interspeech 2014*, Singapore, 2014, pp. 969–972.

[14] X. Li, Y.-C. Ju, L. Deng, and A. Acero, “Efficient and Robust Language Modeling in an Automatic Children’s Reading Tutor System,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 4, pp. 193–196.

[15] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[16] J. Proenca, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, “The LetsRead Corpus of Portuguese Children Reading Aloud for Performance Evaluation,” in *Proc of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož, Slovenia, 2016.

[17] J. Proença, D. Celorico, C. Lopes, S. Candeias, and F. Perdigão, “Automatic Annotation of Disfluent Speech in Children’s Reading Tasks,” in *Proc. IX Jornadas en Tecnologías del Habla and V Iberian SLTech Workshop - IberSPEECH’2016*, Lisbon, Portugal, 2016, pp. 172–181.

[18] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press, 2006.

[19] D. J. MacKay, “A practical Bayesian framework for backpropagation networks,” *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992.

[20] N. R. Draper and H. Smith, *Applied regression analysis*, 3rd ed. Wiley, 1998.

[21] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.