# Lactose Intolerance Prediction Using Artificial Neural Networks

Lemana Spahić, Emir Šehović, Alem Šećerović, Zerina Đozić, and Lejla Smajlović-Skenderagić

**Abstract**

An Artificial Neural Network for lactose intolerance prediction is presented in this paper. The system input information were symptom related questions and answers from a condition-oriented questionnaire, that was filled by one hundred individuals from Bosnia and Herzegovina. Participants were genotyped on LCT 13910 C/T and LCT 22018 G/A polymorphisms, which are reliable predictors of lactose tolerance/intolerance, and that information was the output of the neural network. The ANN consisted of 6 input parameters, that feed the Bayesian regulation training algorithm with information. ANN performance evaluation was performed with 10 samples out of 100 genotyped samples and the results predict whether a person is lactose tolerant or lactose intolerant. The aim of the artificial neural network presented in this paper is to assist specialists in lactose intolerance prediction, avoiding unnecessary further laboratory and genetic testing in clinical practice.

**Keywords**

Lactose intolerance • Artificial neural network • Genotype • Diagnosis • Prediction

## 1 Introduction

Lactose is a disaccharide that is found in all mammalian milks and is very important for nutrition of newborn and infants. In order to be digested, lactose has to be hydrolyzed by enzyme lactase (*lactase*-phlorizin hydrolase, or LPH) into simple sugars, glucose and galactose [1, 2]. Lactase is a trans-membrane glycoprotein of the small intestinal brush border membrane of enterocytes [3, 4] coded, in humans, with LCT gene located on the chromosome 2 [5], long (q) arm at position 21.3. This gene is 49.3 kb in length, consisted of 17 exons and is translated into a 6 kb transcript [6].

Lactase digestive activity reaches its peak in the first few months of life and decreases after the age of two years [7]. Deficient or absent lactase enzymatic activity in the small intestine results in inability of organism to digest lactose from milk and other dairy products. This condition is called lactose intolerance. Besides the congenital lactase deficiency, which is a very rare condition inherited in an autosomal recessive manner [8], identified by total lactose intolerance already at infant age, there are three other types of lactose intolerance: primary, secondary, and developmental lactase deficiency [9–12]. Developmental lactase deficiency and reduced lactase activity is found in infants born before 34 weeks of gestation [10]. Gray was among the first scientists to describe secondary lactase deficiency [13]. Secondary lactase deficiency could occur as a consequence of small intestinal injuries, caused by many different factors such as infections, surgery, chemotherapy, celiac disease, gastroenteritis, prolonged use of antibiotics and other [11]. The most common type of lactose intolerance, which appears in adulthood, is in most of cases characterized with low lactase activity (hypolactasia) leading to primary lactase deficiency [12].

Primary lactase deficiency prevalence in adults differs worldwide, varying from less than 5% to almost 100% of population [14]. Study that included data from 89 countries (approximately 84% of the world's population) found that lactose intolerance is present in 19–37% of western, southern, and northern European population and in 57–83% of Middle East population [15].

Primary adult lactose intolerance is related to the absence of lactase persistence alleles, producing "lactase non-persistence" phenotype [16]. On the contrary, certain number of individuals keep neonatal levels of lactase enzymatic

L. Spahić · E. Šehović · A. Šećerović · Z. Đozić · L. Smajlović-Skenderagić (✉)
International Burch University, Genetics and Bioengineering, Francuske revolucije bb, 71 210 Ilidža, Sarajevo, Bosnia and Herzegovina
e-mail: l.smajlovic.skenderagic@ibu.edu.ba

L. Spahić
e-mail: lemana.spahic@stu.ibu.edu.ba


A. Badnjevic et al. (eds.), *CMBEBIH 2019*, IFMBE Proceedings 73,
https://doi.org/10.1007/978-3-030-17971-7_75

activity throughout the adulthood due to the presence of lactase related alleles, producing "lactase-persistence" phenotype. Except for some allelic differences (silent mutations), lactase persistent and lactase non-persistent groups of individuals have identical coding sequences [6]. Lactase persistence/non-persistence phenotypes are connected with few single nucleotide polymorphisms, whose respective frequencies vary across different world regions and ethnic groups [16]. The most investigated polymorphism associated with lactase persistence is LCT 13910*C/T (rs4988235), that is found to be in almost full concordance with LCT 22018*G/A (rs182549) polymorphism [17]. Both LCT-13910 CC and LCT22018 GG genotypes are strong predictors of lactase non-persistence [18].

In lactose intolerant individuals, non-digested lactose passes from the intestines to the colon, were it serves as a bacterial substrate. Depending on the amount of lactose ingested, people with lactose intolerance can, shortly after consumption of milk and dairy products, experience discomfort and pain as a manifestation of different gastrointestinal symptoms. The most common symptoms of lactose intolerance are diarrhea, bloating, flatulence, nausea, gut distension, and abdominal pain [19].

Lactose intolerance can be distinguished from other disorders by different diagnostic tests such as: lactose tolerance test, hydrogen breath test, stool acidity test (children) or by genetic testing. First three types of diagnostic tests require ingestion of certain amounts of lactose, which can cause discomfort and could be painful for the patients. Genetic testing of lactose intolerance associated lactase polymorphisms is not widely available [20].

Machine learning is a field in artificial intelligence and is one of the most rapidly developing subfields of artificial intelligence research. Machine learning enables highly proficient intelligent data analysis. The inexpensive and relatively easy methods developed within the last two decades for collecting and storing data also contributed to making machine learning procedures easier and more consistent. Since the beginning, machine learning was used and implemented within the medical field [21]. Many hospitals and clinics worldwide are monitoring and collecting data which can later be used for machine learning purposes. The machine learning methodology is most convenient for very specific diagnostic problems [22].

Approximations of explanations of certain processes can be considered as the essence of machine learning. Approximations generally do not and cannot explain the whole process, therefore usage of other algorithms would be more convenient. The machine learning process takes into account that the patterns observed within the existing dataset will not change within the future datasets regarding the same problem. In medicine, machine learning programs used for predictions of medical diagnosis are mostly based on concrete biological and physical parameters [23, 24]. However, sometimes, as is the case here, a target condition and symptom-oriented questionnaire can be used for creating the machine learning system [25].

The fundamental basis of machine learning is the optimization of prediction performance by utilizing previously collected data or previously gained experience. The machine learning models can be classified into two groups: predictive and descriptive. The predictive model makes future estimates based on the collected data, while the descriptive model obtains knowledge from the data. Sometimes, both of the models can be implemented into a single model [22].

ANNs are trained in such a way that the optimal weighting and bias values are acquired in order to obtain the desired mapping or clustering of data. In this manner, ANNs can find relationship within and between datasets without defining the exact mathematical principle behind it. The connection between the neurons in a neural network is what defines its architecture. There are two types of architectures: feedforward and feedback [26–28].

This paper presents the development and feasibility of an ANN for Lactose intolerance prediction. This diagnostic tool can assist specialists in clinical practice to make the diagnostic process significantly faster by avoiding unnecessary lactose tolerance and genetic testing.

## 2 Materials and Methods

### 2.1 Dataset

The dataset used in the development of this neural network was based on symptoms reported in lactose intolerance related questionnaire and obtained LCT 13910 C/T and 22018 G/A genotypes. Study included 100 unrelated participants from Bosnia and Herzegovina. Genetic analysis was done using PCR-RFLP methodology proposed by Bulchoes et al. [29]. The restriction digestion products were analyzed using agarose gel electrophoresis. LCT 13910 and 22018 related genotypes were determined according to the size of the digestion products.

The specific questionnaire was designed in order to investigate the occurrence and severity of main lactose intolerance symptoms, and to analyze symptoms with respect to obtained LCT 13910 and 22018 genotypes and self-reported lactose tolerance.

The questions that bared most correlation to the genotypes were:

1. Do you have close family members who experience health problems after consuming milk or dairy products?
2. Do you feel discomfort after consuming milk or dairy products?

3. Do you feel nausea after consuming milk or dairy products?
4. Do you feel flatulence after consuming milk or dairy products?
5. Do you feel pain in your stomach after consuming milk or dairy products?
6. Do you have diarrhea after consuming milk or dairy products?

The abovementioned questions and the answers in form of symptom intensity were the only parameters used as inputs of ANN. Their inputs are defined in Fig. 1, with Q1–Q6 each indicating a question respectively.

The dataset consisted of 100 samples whose distribution is presented in Table 1.

## 2.2 Development of Artificial Neural Network

Feedforward neural network architecture was constructed as it is best suited for solving problems related to classification.

The data division, for the purposes of artificial neural network training, was done in a 90/10 ratio, as confirmed by various trials. In order to prevent overfitting and due to its usefulness in pattern recognition, Bayesian regulation training algorithm was used. For each training iteration the train/test performance was calculated as Mean Square Error between the actual and predicted values (MSE).

Most prominently used training functions were used to test the performance of ANN in order to choose the appropriate architecture for further development. As it can be seen from Table 2, best performance was observed using Bayesian regularization training algorithm (Trainbr) with 20 neurons in the hidden layer. Bayesian regularization is an algorithm most prominently used with datasets consisting of small number of samples and therefore it was expected to be the most suitable algorithm for this particular dataset [30].

After determining the most suitable training algorithm, the network was further tested with different combinations of transfer functions in the hidden layer (Table 3). As it can be inferred from Table 3, the best performance was achieved with 20 neurons in the hidden layer with Tansig transfer function in the hidden layer and Logsig transfer function in the output layer, which are the defaults in Bayesian regularization.

## 3 Results and Discussion

The final result of the evaluation suggests that the most suitable architecture for Artificial Neural Network for Lactose Intolerance prediction is the one with Bayesian regularization training algorithm, default transfer functions and 20 neurons in the hidden layer.

The final model consists of 6 neurons in the input layer of the network, one for each input parameter in form of 6 questions from the questionnaire. The architecture of the network continues with 20 neurons and Tansig transfer function in the hidden layer, ending with Logsig transfer function in the output layer. The output layer has only one neuron, with the final output of either 0 or 1, lactose tolerant or lactose intolerant respectively.

Subsequent validation was performed using 10 samples from the initial dataset, which makes 10% of the overall dataset. Evaluation of ANN performance through specificity, sensitivity and accuracy parameters is displayed in Table 4. Specificity is calculated as a number of correctly classified

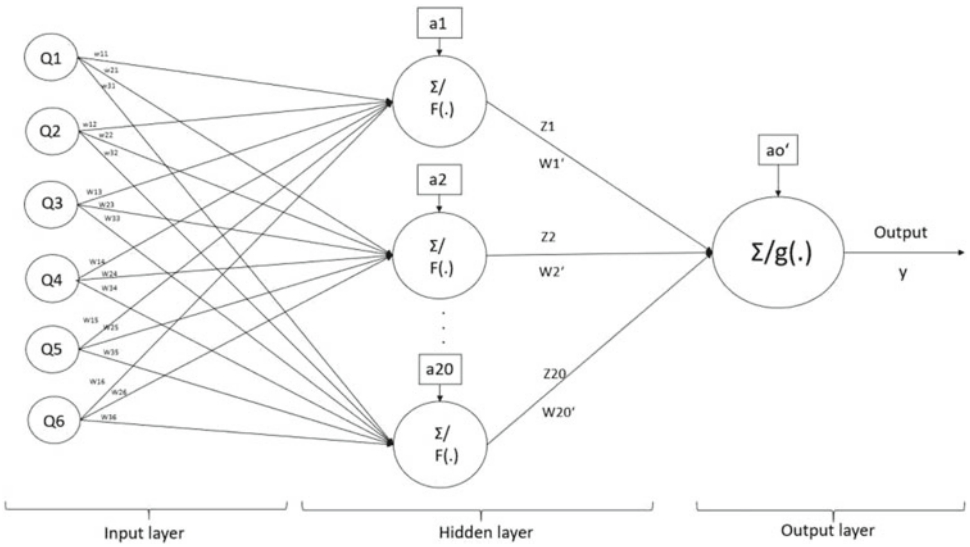


**Fig. 1** Architecture of ANN for lactose intolerance prediction

**Table 1** Lactose tolerance dataset distribution

| Training dataset | | Subsequent validation dataset | |
|---|---|---|---|
| Sample group | Number of samples | Sample group | Number of samples |
| Lactose tolerant | 50 | Lactose tolerant | 5 |
| Lactose intolerant | 40 | Lactose intolerant | 5 |
| ■ | 90 | ■ | 10 |

**Table 2** ANN performance evaluation with different combinations of training algorithms and neuron numbers

| Training algorithm | Number of neurons in hidden layer | ANN performance |
|---|---|---|
| Trainbr | 5 | 3.9536 e-09 |
| Trainbr | 10 | 8.2886 e-09 |
| Trainbr | 20 | **9.9055 e-10** |
| Trainbr | 50 | 2.3298 e-09 |
| Trainlm | 5 | 0.0701 |
| Trainlm | 10 | 0.0719 |
| Trainlm | 20 | 0.0667 |
| Trainlm | 50 | 0.3767 |
| Trainbfg | 5 | 0.5468 |
| Trainbfg | 10 | 1.2571 |
| Trainbfg | 20 | 0.2979 |
| Trainbfg | 50 | 11.2636 |

**Table 3** ANN performance evaluation with different transfer function combinations and different neuron numbers

| Training algorithm | Number of neurons | Transfer functions | ANN performance |
|---|---|---|---|
| Trainbr | 5 | Tansig; logsig | 3.9536 e-09 |
| Trainbr | 10 | Tansig; logsig | 8.2886 e-09 |
| Trainbr | 20 | Tansig; logsig | **9.9055 e-10** |
| Trainbr | 50 | Tansig; logsig | 2.3298 e-09 |
| Trainbr | 5 | Purelin; purelin | 0.0686 |
| Trainbr | 10 | Purelin; purelin | 0.0664 |
| Trainbr | 20 | Purelin; purelin | 0.0664 |
| Trainbr | 50 | Purelin; purelin | 0.0665 |
| Trainbr | 5 | Logsig; logsig | 0.0654 |
| Trainbr | 10 | Logsig; logsig | 0.0654 |
| Trainbr | 20 | Logsig; logsig | 0.0643 |
| Trainbr | 50 | Logsig; logsig | 0.0642 |

**Table 4** Confusion matrix of subsequent validation dataset

| | ANN | | |
|---|---|---|---|
| | Lactose tolerant | Lactose intolerant | |
| Lactose tolerant | 5 | 0 | 5 |
| Lactose intolerant | 0 | 5 | 5 |
| | Specificity 100% | Sensitivity 100% | |

**Table 5** Confusion matrix for most of the k-folds of cross validation

| | ANN | | |
|---|---|---|---|
| | Lactose tolerant | Lactose intolerant | |
| Lactose tolerant | 36 | 3 | 4 |
| Lactose intolerant | 4 | 47 | 6 |
| | Specificity 92.3% | Sensitivity 92.2% | |

samples of lactose tolerant group divided by the total number of lactose tolerant samples. Sensitivity is calculated as the number of correctly classified samples of lactose intolerant group divided by the total number of lactose intolerant samples. Accuracy is determined by the number of correctly classified samples divided by the total number of samples. All three analyzed parameters resulted with 100% for subsequent validation dataset, meaning that this neural network can correctly differentiate lactose tolerance and lactose intolerance.

k-fold cross validation method was implemented as an additional step in order to test the performance of ANN more thoroughly, according to the code presented in appendix. The dataset was subdivided into 10 classes for training and testing and the resulting accuracy varied slightly in multiple runs. The results obtained over trials of k-fold cross validation average to an accuracy level of 92.2% which is expected when taking the overall sample size into consideration. The most prominent resulting confusion matrix after cross validation is presented in Table 5.

## 4 Conclusion

An Artificial Neural Network for lactose intolerance prediction was presented in this paper. Training was done using 90 samples from a 100 samples dataset, 10 of which were used for subsequent validation. The ANN demonstrated very high specificity and sensitivity which indicates that successful and reliable ANNs based on Lactase non- persistence symptoms can be created.

Diagnosis of lactose intolerance is not a straightforward procedure and it usually involves analysis of initial symptoms and medical history combined with results of related biochemical and genetic testing. This ANN is an automatic diagnostic tool that is based solely on self-reported symptoms related to digestion of lactose. The final result is a tool, that if clinically optimized, is able to predict lactose intolerance without any laboratory testing.

Future perspectives of this work will include gathering more samples and performing LCT SNP related genotyping, which will further improve the scope of training parameters and enable the efficiency of the network when unexpected symptoms are reported. This work has the potential to be used in healthcare and provide medical professionals with both time and cost-effective lactose intolerance diagnosis procedure.

## References


1. Troelsen, J.T.: Adult-type hypolactasia and regulation of lactase expression. Biochim. Biophys. Acta **1723**, 19–32 (2005)
2. Rasinperä, H., Savilahti, E., Enattah, N.S., et al.: A genetic test which can be used to diagnose adult-type hypolactasia in children. Gut **53**, 1571–1576 (2004)
3. Danielsen, E.M., Skovbjerg, H., Norén, O., Sjöström, H.: Biosynthesis of intestinal microvillar proteins intracellular processing of lactase-phlorizin hydrolase. Biochem. Biophys. Res. Commun. **122**(1), 82–90 (1984)
4. Naim, H.Y., Sterchi, E.E., Lentze, M.J.: Biosynthesis and maturation of lactasephlorizin hydrolase in the human small intestinal epithelial cells. Biochem. J. **241**(2), 427–434 (1987)
5. Kruse, T.A., Bolund, L., Grzeschik, K.H., Ropers, H.H., Sjöström, H., Noren, O., et al.: The human lactase-phlorizin hydrolase gene is located on chromosome 2. FEBS Lett. **240**(1–2), 123–126 (1988)
6. Boll, W., Wagner, P., Mantei, N.: Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. Am. J. Hum. Genet. **48**(5), 889 (1991)
7. Troelsen, J.T., Olsen, J., Møller, J., SjÖstrÖm, H.: An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology **125**(6), 1686–1694 (2003)
8. Savilahti, E., Launiala, K., Kuitunen, P.: Congenital lactase deficiency. A clinical study on 16 patients. Arch. Dis. Child. **58** (4), 246–252 (1983)
9. Simoons, F.J.: Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. Am. J. Dig. Dis. **15**(8), 695710 (1970)
10. Antonowicz, I., Lebenthal, E.: Developmental pattern of small intestinal enterokinase and disaccharidase activities in the human fetus. Gastroenterology **72**(6), 1299–1303 (1977)
11. Heyman, M.B.: Lactose intolerance in infants, children, and adolescents. Pediatrics **118**(3), 1279–1286 (2006)
12. Vesa, T.H., Marteau, P., Korpela, R.: Lactose intolerance. J. Am. Coll. Nutr. **19**(sup2), 165S–175S (2000)
13. Gray, G.M., Walter, W.M., Colver, E.H.: Persistent deficiency of intestinal lactase in apparently cured tropical sprue. Gastroenterology **54**(4), 552–558 (1968)
14. Sahi, T.: Hypolactasia and lactase persistence historical review and the terminology. Scand. J. Gastroenterol. **29**(sup202), 1–6 (1994)
15. Storhaug, C.L., Fosse, S.K., Fadnes, L.T.: Country, regional, and global estimates for lactose malabsorption in adults: a systematic review and meta-analysis. Lancet Gastroenterol. Hepato. **2**(10), 738–746 (2017)
16. Kuokkanen, M., Enattah, N.S., Oksanen, A., Savilahti, E., Orpana, A., Järvelä, I.: Transcriptional regulation of the lactase-phlorizin

hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut **52**(5), 647–652 (2003)
17. Mattar, R., Mazo, Carrilho.: Lactose intolerance: diagnosis, genetic, and clinical factors. Clin. Exp. Gastroenterol **113** (2012)
18. Adler, G., Adler, M., Valjevac, A., Mackic-Djurovic, M., Kiseljakovic, E.: First Bosnian study of LCT −13910C > T and −22018G > A single nucleotide polymorphisms associated with adult-type lactose intolerance. Folia Med. Fac. Med. Univ. Saraeviensis **52**(1), 3–8 (2017)
19. Wilt, T.J., Shaukat, A., Shamliyan, T., Taylor, B.C., MacDonald, R., Tacklind, J., et al.: Lactose intolerance and health. Evid. Rep. Technol. Assess. (Full Rep.) **192**(1), 410 (2010)
20. Newcomer, A.D., McGill, D.B., Thomas, P.J., Hofmann, A.F.: Prospective comparison of indirect methods for detecting lactase deficiency. N. Engl. J. Med. **293**(24), 1232–1236 (1975)
21. Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds.): Machine Learning: An Artificial Intelligence Approach. Springer Science & Business Media (2013)
22. Alpaydin, E.: Introduction to Machine Learning/Ethem Alpaydin (2010)
23. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7**(6), 673 (2001)
24. Baxt, W.G.: Use of an artificial neural network for data analysis in clinical decision making: the diagnosis of acute coronary occlusion. Neural Comput. **2**(4), 480489 (1990)
25. Kononenko, I.: Machine learning for medical diagnosis: history, state of the art and perspective. Artif. Intell. Med. **23**(1), 89–109 (2001)
26. Badnjevic A., Cifrek M., Koruga D.: Classification of chronic obstructive pulmonary disease (COPD) using integrated software suite. In: IFMBE XIII Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON), pp. 25–28. Sevilla, Spain (Sept 2013)
27. Alic, B., Sejdinovic, D., Gurbeta, L., Badnjevic, A.: Classification of stress recognition using artificial neural network. In: IEEE MECO (2016)
28. Alic, B., Gurbeta, L., Badnjevic, A., et.al.: Classification of metabolic syndrome patients using implemented expert system. In: CMBEBIH 2017. IFMBE Proceedings, vol 62, pp 601–607. Springer
29. Bulhões, A., Goldani, H., Oliveira, F., Matte, U., Mazzuca, R., Silveira, T.: Correlation between lactose absorption and the C/T-13910 and G/A-22018 mutations of the lactase-phlorizin hydrolase (LCT) gene in adult-type hypolactasia. Braz. J. Med. Biol. Res. **40**(11), 1441–1446 (2007)
30. Burden, F., Winkler, D.: Bayesian regularization of neural networks. In: Artificial Neural Networks, pp. 23–42. Humana Press (2008)