Computer Science and Application 计算机科学与应用, 2021, 11(6), 1747-1754
Published Online June 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.116180

Hans汉斯

基于卷积神经网络的子宫内膜癌分类问题

钟　滢，张　悦

东北大学理学院，辽宁　沈阳
Email: 541441298@qq.com, zhangyue@mail.neu.edu.cn

收稿日期：2021年5月20日；录用日期：2021年6月17日；发布日期：2021年6月24日

摘　要

在本文中，讨论了基于卷积神经网络(CNN)对87位女性子宫内膜基因表达样本的分类问题。首先，删除掉缺失数据对应的基因，计算信噪比来过滤不相关的基因。然后，将每个指标相应的数据放入CNN中求出分类准确率。之后对每个指标进行归一化处理，同样通过CNN得到4个指标组合的分类准确率。最后，应用下三角矩阵和上三角零元素处理来改进初始化卷积核。后者有效地提高了训练集以及测试集的分类准确率。

关键词

子宫内膜癌，基因表达，卷积神经网络，归一化，卷积核

Classification of Endometrial Carcinoma Based on Convolutional Neural Network

Ying Zhong, Yue Zhang

College of Science, Northeastern University, Shenyang Liaoning
Email: 541441298@qq.com, zhangyue@mail.neu.edu.cn

Received: May 20th, 2021; accepted: Jun. 17th, 2021; published: Jun. 24th, 2021

Abstract

In this paper, the classification of 87 female endometrial gene expression samples based on convolutional neural network (CNN) is discussed. First, the genes corresponding to the missing data were deleted and the signal-to-noise ratio was calculated to filter out the unrelated genes. Then, the corresponding data of each indicator is put into CNN to calculate the classification accuracy rate. Then each indicator was normalized, and the classification accuracy rate of the four indicators combined was also obtained through CNN. Finally, the lower triangular matrix and the upper

文章引用: 钟滢, 张悦. 基于卷积神经网络的子宫内膜癌分类问题[J]. 计算机科学与应用, 2021, 11(6): 1747-1754.
DOI: 10.12677/csa.2021.116180

triangular zero element processing are used to improve the initial convolution kernel. The latter can effectively improve the classification accuracy rate of training set and test set.

Keywords

Endometrial Cancer, Gene Expression, Convolutional Neural Network, Normalization, Convolution Kernel

1. 引言

肿瘤是当今威胁我们人类健康和生命的主要疾病原因之一，预防和处理肿瘤问题也是所有科学家和研究者密切关心的问题[1]。现如今，合理地利用计算机技术，对癌症进行准确的早期预测变得越来越有意义。在 1999 年，首次提出基因表达方法，用于急性髓系白血病和急性淋巴细胞白血病的癌症分类问题[2]。从那时起，基于基因表达的癌症分类开始越来越受到研究者的关注[3] [4] [5]。

Peterson 和 Ringner 讨论了各种监督和非监督数据挖掘方法来分析产生的高维数据，重点是肿瘤基因表达谱的分类和预测[6]。基因选择方法有很多，如 K-split lasso，针对肿瘤基因表达数据，是一种有效的特征选择方法，数据冗余得以减少，样本的分类准确率得以提高[7]。将粒子群算法与灰狼优化算法相结合，对 Elman 循环神经网络的参数进行优化[8]。在肿瘤分类和预测领域，神经网络的应用取得了良好的效果，作者讨论了利用基因表达和人工神经网络对前列腺癌的分类和诊断预测[9]。

严重威胁妇女健康的疾病之一的子宫内膜癌，是子宫癌的一种，对这种妇科最常见的恶性肿瘤的流行病学、病理生理学和管理策略的全面了解，会使得产科医生或妇科医生能够识别风险较高的妇女，这是有助于减少风险并促进早期诊断[10]。

卷积神经网络作为一种新的网络模型，已逐渐被引入癌症预测中。作者综述了近年来利用卷积神经网络的深度学习方法进行基因表达数据分析的研究工作[11]。该系统旨在提高三维 MRI 图像中不同类型肿瘤的分类精度[12]。该算法由一个具有改进的 softmax 损失函数和正则化的卷积神经网络组成。

本文选取了小样本、高维的子宫内膜基因数据，类似于简单的图片信息。将数据输入到卷积神经网络并进行正则化处理。最后，将遗传数据中包含的 4 个指标进行合并，再次放入卷积神经网络中。并且，对初始化卷积核进行改进，发现癌症的分类准确率得以提高。

本文的其余部分组织如下：第二部分介绍了卷积神经网络的基本理论和归一化方法。第三部分介绍数据预处理。第四部分介绍了整个基因数据的实验过程。第五部分给出结束语。

2. 基本理论

2.1. 卷积神经网络

上世纪的八九十年代，卷积神经网络的研究刚刚兴起。最早的卷积神经网络有时延网络和 Lenet-5 等。Lecun 总结了卷积神经网络的特点并将其命名为卷积神经网络，这使他成为了卷积神经网络之父[13]。近年来，深度学习理论的逐步发展和数值计算设备的相应改进，卷积神经网络得到了快速的发展。它已成功地应用于计算机视觉、自然语言处理等诸多领域。卷积神经网络也可以用于时间序列分类[14] [15]。

卷积神经网络主要具有表征学习的能力，能够按其阶层结构对输入信息进行平移不变分类，因此又称“平移不变人工神经网络”[16]。卷积神经网络的基本结构由输入层、卷积层、池化层、全连接层以及输出层组成，是一种前馈神经网络，其中包含了卷积计算和深度结构，是深度学习算法中比较具有代表性的[17]。卷积层和池化层通常使用多组结合的方式，交替设置。卷积层连接到池化层，池化层连接到卷积层，类似重复。

卷积神经网络中最重要的模块当属卷积层，它有两个重要的思想，即局部连接和权值共享。它由多个特征映射组成，每个特征映射又由多个神经元组成。通过卷积核，每个神经元才得以连接到上层特征映射的局部区域。卷积核实质就是一个权重矩阵。这里介绍了几种改进卷积核的方法[18] [19]。在卷积层中，通过卷积运算提取输入的不同特征[20]。之后的多个特性映射组成池化层，每个特性映射都唯一地对应于其上层的一个，却不改变特性映射的数量。池化层的目的是通过降低特征图的分辨率来获得空间不变的特征[21]。常用的池化方法包括最大池化、平均池化等。前者取局部接受域中值最大的点。后者计算本地接受域内所有值的平均值。在多个卷积层和池化层之后是一个或多个全连接层。每个神经元都与前一层的所有神经元完全相连。全连接层起到整合卷积层或池化层中具有类别区分的局部信息的作用。

2.2. 归一化法

由于使用了各种各样的数据，有必要检查数据的取值范围是否接近。如果差异太大，将导致不准确的结果。Min-max 归一化和 Z-score 归一化是常用的两种归一化方法。其中，最小－最大归一化是对原始数据进行线性变换，使更新后的数据在 0 到 1 范围内。采用最小－最大归一化的方法来降低图像采集过程中由多种因素引起的不确定性噪声[22]。转换函数如下：

$$x' = \frac{x - Min}{Max - Min}$$

其中 Max 为样本数据的最大值， Min 为样本数据的最小值。在归一化过程中，由于该方法只涉及到变量的最大值和最小值，因此在变量更新时会过度依赖这两个极值。添加了新数据后，可能需要重新定义数据。

Z 分数归一化使用原始数据的均值和标准差[23]。处理后的数据服从标准正态分布，均值为 0，标准差为 1。这种归一化是将一个量纲表达式转换为一个无量纲表达式。转换函数如下：

$$x' = \frac{x - Mean}{SD}$$

其中 $Mean$ 是样本均值， SD 是样本标准差。该方法使更新后的数据的平均值和标准差相同。归一化还有效地消除了各变量之间差异的变化程度。

3. 数据预处理

87 例基于基因表达的女性子宫内膜样本采集自 http://www.ncbi.nlm.nih.gov/geo/。该数据于 2015 年 8 月 15 日公布。其中包括 64 个子宫内膜癌样本和 23 个无癌样本。每个样本含有 27,578 个基因。每个基因有 5 个指标，包括平均 beta 值、强度、未甲基化信号、甲基化信号和检测 P 值，构成所有基因的 27,578 * 5 矩阵。首先，试图计算各指标的样本分类准确率。可见，子宫内膜样本的基因数据具有小样本、高维度的特点。在这些基因中，与癌症相关的基因可能有 100 个左右。因此，有必要先进行降维处理。对于基因某些指标中的缺失数据，考虑了直接使用缺失数据的基因、删除缺失数据的基因和补全缺失数据的三种处理方法(均值插值、齐次均值插值、建模预测等)。考虑到缺失数据的基因数量较少，故选择删除缺失数据基因的方法。

然后在每个指标中，根据信噪比对每个基因进行排序，过滤出不相关的基因。信噪比的计算公式如下：

$$SNR(g_i)=\frac{\left|\mu_+(g_i)-\mu_-(g_i)\right|}{\sigma_+(g_i)+\sigma_-(g_i)}$$

其中 g_i 为样本的第 i 个基因，$\mu_+(g_i)$、$\mu_-(g_i)$ 分别为癌样和非癌样本中 g_i 基因的平均值。$\sigma_+(g_i)$ 和 $\sigma_-(g_i)$ 分别为 g_i 基因在癌样和非癌样中的标准差。获得 87 个样本中各指标的 100 个高信噪比基因。各指标选择的 100 个基因的信噪比范围见表 1。

Table 1. SNRs range of 100 genes selected
表 1. 选择的 100 个基因的信噪比范围

取值	特征 1	特征 2	特征 3	特征 4	特征 5
最小值	1.4821	0.6628	0.8059	1.2854	0.8107
最大值	2.4575	1.0651	1.7030	1.9515	0.8107

4. 实验

4.1. 卷积神经网络

本文选择卷积神经网络作为基本模型。由于卷积神经网络的输入数据通常为二维结构，因此需要将每个指标选取的 100 个基因数据转换为 10 * 10 的二维数据矩阵。

从数据预处理的结果可以看出，第 5 个指标的数据一直没有变化。这样的数据是不适合进行数据分析处理的。因此，在接下来的实验中，第五个指标的数据将被舍弃。在卷积神经网络的过程中，我们使用留一交叉验证(LOOCV)来计算每个指标的分类准确率。对卷积神经网络结构中的每个参数进行更新和训练，以确定值，使每个集合具有更高的准确率。在这个过程中，我们发现这类样本更适合相对简单的结构。

在上述测试的基础上，选择了单层卷积结构。卷积神经网络结构如表 2 所示。下面的公式表示用于初始化卷积神经网络中的卷积核定义。

$$K=(A-0.5)*2*\sqrt{\frac{6}{fan\_in+fan\_out}}$$

其中 K 为初始化的卷积核。A 是一个 5 * 5 的随机矩阵，即在 0 和 1 之间均匀分布的随机数数组。fan_in 和 fan_out 分别为各层输入部分和输出部分的卷积核内的参数个数之和。经过多次实验，相应的错误率转化为正确率。对于每个指标，训练集和测试集的分类准确率如表 3 所示。

Table 2. Convolutional neural network structure setting
表 2. 卷积神经网络结构设置

结构	特征图个数	卷积核大小	迭代次数	每批数量	激活函数
取值	6	5 * 5	30	2	Sigmoid 函数

Table 3. The classifification accuracy rate of each indicator
表 3. 各指标的分类准确率

准确率	特征 1	特征 2	特征 3	特征 4
训练集	98.9174%	73.5632%	73.5632%	73.5632%
测试集	100%	73.5632%	73.5632%	73.5632%

从表 3 可以看出，除第一个指标外，分类准确率都比较低。接下来，我们尝试结合这 4 个指标来提高整体的分类准确率。

4.2. 线性综合指标

以上分类为单一指标。事实上，仅用一个指标来判断一个样本是否癌变是不合适的。根据下式，将4个指标组合成一套新的遗传数据进行分类。组合公式如下：

$$x_{new} = \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_4 x_4$$

其中 $\omega_i\ (i=1,2,3,4)$ 是系数并且满足 $\sum_{i=1}^{4} \omega_i = 1$。$x_i\ (i=1,2,3,4)$ 为基因中第 i 个指标的值。$\omega_i\ (i=1,2,3,4)$ 是对测试集中每个指标的分类准确率进行归一化处理得到。由表 3 可知，测试集中 4 个指标的分类正确率分别为 100%、73.5632%、73.5632%、73.5632%。可以得到如下归一化：

$$x_{new} = 0.3118x_1 + 0.2294x_2 + 0.2294x_3 + 0.2294x_4$$

在生成新数据的过程中，存在一个问题。这 4 个指标在取值范围上存在很大的差异。以训练集为例，各指标下的基因范围如表 4 所示。如果直接进行数据生成，小范围的数据难以发挥作用，失去了生成新数据的重要意义。然后采用 Min-max 归一化方法对 4 个指标下的数据进行处理，确保每个指标的取值范围在 0~1 之间。

Table 4. The value range of genes under each indicator (training set)
表 4. 各指标(训练集)下基因的取值范围

取值	特征 1	特征 2	特征 3	特征 4
最小值	0.0162	243	83	105
最大值	0.9559	26938	17644	17806

随后，将归一化后新生成的数据放入卷积神经网络，进行训练和测试，还选择留一交叉验证(LOOCV)。卷积神经网络中的初始化卷积核与前一节的处理方法相同。在实验过程中，我们记录了图 1 所示特征图数目从 1 到 10 时不同数据集的错误率。我们可以找到最合适的特征图数目。当特征图个数为 5 时，训练集和测试集的误差都较小。通过转换，最优的训练集和测试集的分类准确率分别为 96.67%和 95.4%。

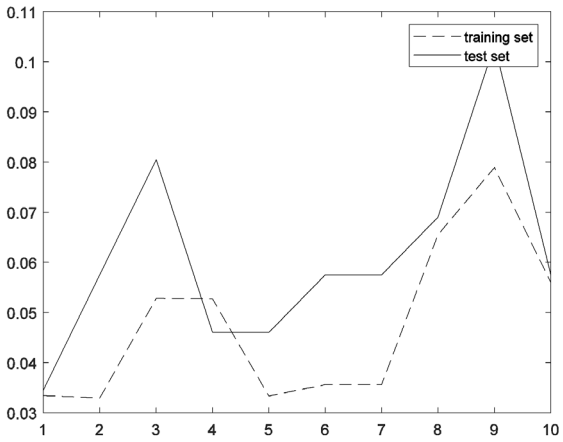

Figure 1. The error rate of different data sets when the number of feature graphs is from 1 to 10
图 1. 当特征图数目为 1~10 时，不同数据集的错误率

4.3. 卷积核改进

在卷积神经网络的改进中，我们采用了不同的卷积核初始化方法。根据样本中遗传数据的排列规则，

来选择卷积核初始化规则。每个样本的 100 个遗传数据在矩阵中是从第一列到最后一列排列的。由于从每个样本中过滤出的 100 个遗传数据按照信噪比由大到小进行排序，因此有用的信息被排列在矩阵的左侧。我们试图生成下三角矩阵而不是随机生成这个矩阵。初始化卷积核的公式如下所示。

$$K=(B-0.5)*2*\sqrt{\frac{6}{fan\_in+fan\_out}}$$

其中 B 为下三角矩阵，其余保持不变。同样，我们记录了图 2 所示特征图数目为 1~10 时不同数据集的错误率。我们发现最合适的特征图数目。当特征图个数为 7 时，训练集和测试集的误差都较小。通过转换，最优训练集和测试集的分类准确率分别为 96.99%和 97.7%。与之前的结果相比，准确率有了提高。

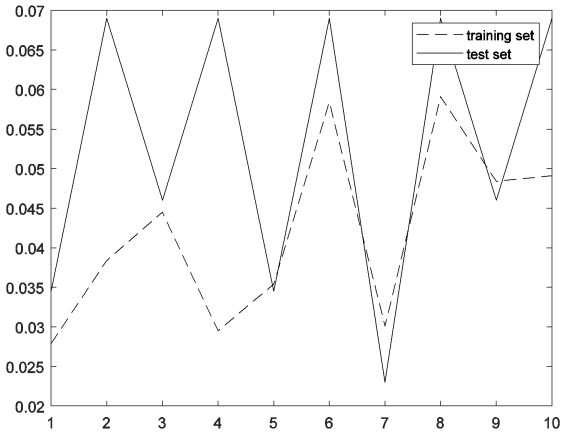

Figure 2. The error rate of different data sets when the number of feature graphs is from 1 to 10
图 2. 当特征图数目为 1~10 时，不同数据集的错误率

接下来，对上述矩阵 B 再次进行改进。把矩阵 B 右上角所有 0 的元素都换成−1，称它为 C。初始化卷积核的公式如下所示。

$$K=(C-0.5)*2*\sqrt{\frac{6}{fan\_in+fan\_out}}$$

同样，我们将特征图数目从 1 到 10 的不同数据集的错误率绘制如图 3 所示。我们找到了同时满足训练集和测试集误差都较小的特征图数目，即当特征图个数为 7 时，训练集和测试集的误差都较小。通过转换，最优训练集和测试集的分类准确率分别为 97.43%和 97.7%。

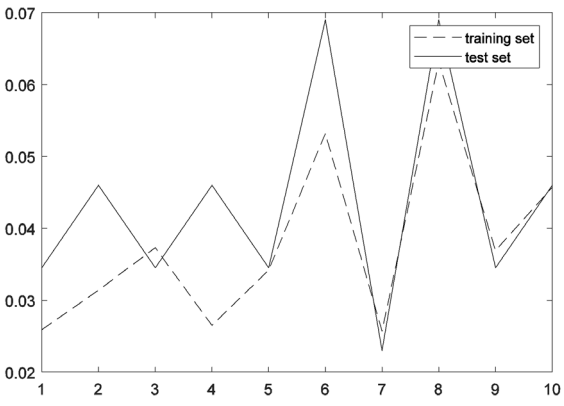

Figure 3. The error rate of different data sets when the number of feature graphs is from 1 to 10
图 3. 当特征图数目为 1~10 时，不同数据集的错误率

4.4. 比较

上述初始卷积核变化结果列于表 5 中进行比较，结果显示，训练集的分类准确率在逐步提升，测试集的准确率可以提高到 97.7%。显然，最后一种卷积核初始化方法是最好的，它有效地增大了信噪比较高的基因所占的权重，抑制了信噪比较低的基因的权重，从而提高了卷积神经网络的分类准确率，该方法是适用于此数据集的分类问题。

Table 5. The classification accuracy rate of different initial convolution kernel methods
表 5. 不同初始卷积核方法的分类准确率

初始核方法	特征图个数	训练集	测试集
1	5	96.67%	95.4%
2	7	96.99%	97.7%
3	7	97.43%	97.7%

5. 结论

本文对遗传数据进行常规处理，并将新的遗传数据放入卷积神经网络中。然后，我们改变初始化的卷积核来提高神经网络的分类准确率。它们分别是下三角矩阵处理和上三角 0 元素处理。从结果来看，上三角 0 元素处理是最好的。测试集的准确率达到了 97.7%，训练集的准确率也非常高。此外，还可以通过改变卷积神经网络的其他结构来试图提高分类准确率，并且，对于其他癌症基因数据的分类预测，同时可以利用这类方法进行探究。

基金项目

本文受国家自然科学基金(61703083 和 61673100)和国家留学基金委(201706085041)资助。

参考文献

[1] 黄磊. 癌症分类中基因选择的收缩特征选择算法研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2015.

[2] Golub, T.R., Slonim, D.K., Tamayo, P., *et al.* (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**, 531-537. https://doi.org/10.1126/science.286.5439.531

[3] Alon, U., Barkai, N., Notterman, D.A., *et al.* (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colontissues Probed by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745-6750. https://doi.org/10.1073/pnas.96.12.6745

[4] Armstrong, S.A., Staunton, J.E., Silverman, L.B., *et al.* (2002) *MLL* Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia. *Nature Genetics*, **30**, 41-47. https://doi.org/10.1038/ng765

[5] Hu, H.P., Niu, Z.J., Bai, Y.P. and Tan, X.H. (2015) Cancer Classification Based on Gene Expression Using Neural Networks. *Genetics and Molecular Research*, **14**, 17605-17611. https://doi.org/10.4238/2015.December.21.33

[6] Peterson, C. and Ringner, M. (2003) Analyzing Tumor Gene Expression Profiles. *Artificial Intelligence in Medicine*, **28**, 59-74. https://doi.org/10.1016/S0933-3657(03)00035-6

[7] Zhang, J., Hu, X.G. and Zhang, Y.H. (2012) K-Split Lasso: An Effective Feature Selection Method for Tumor Gene Expression Data. *Journal of Frontiers of Computer Science and Technology*, **6**, 1136-1143.

[8] Hu, H.P., Wang, H.Y., Bai, Y.P. and Liu, M.X. (2019) Determination of Endometrial Carcinoma with Gene Expression Based on Optimized Elman Neural Network. *Applied Mathematics and Computation*, **341**, 204-214. https://doi.org/10.1016/j.amc.2018.09.005

[9] Tirumala, S.S. and Narayanan, A. (2019) Classification and Diagnostic Prediction of Prostate Cancer Using Gene Expression and Artificial Neural Networks. *Neural Computing and Applications*, **31**, 7539-7548. https://doi.org/10.1007/s00521-018-3589-8

[10] 胡红萍, 高帅, 孙强, 白艳萍. 基于基因表达子宫内膜癌的分类[J]. 数学的实践与认识, 2017, 47(18): 111-115.

[11] Gunavathi, C., Sivasubramanian, K., Keerthika, P. and Paramasivam, C. (2020) A Review on Convolutional Neural Network Based Deep Learning Methods in Gene Expression Data for Disease Diagnosis. *Materials Today*: *Proceedings*, **45**, 2282-2285. https://doi.org/10.1016/j.matpr.2020.10.263

[12] Maharjan, S., Alsadoon, A., Prasad, P.W.C., *et al.* (2020) A Novel Enhanced Softmax Loss Function for Brain Tumour Detection Using Deep Learning. *Journal of Neuroscience Methods*, **330**, 108520. https://doi.org/10.1016/j.jneumeth.2019.108520

[13] Lecun, Y., Boser, B., Denker, J.S., *et al.* (1989) Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, **1**, 541-551. https://doi.org/10.1162/neco.1989.1.4.541

[14] Zou, X.W., Wang, Z.D., Li, Q. and Sheng, W.G. (2019) Integration of Residual Network and Convolutional Neural Network along with Various Activation Functions and Global Pooling for Time Series Classification. *Neurocomputing*, **367**, 39-45. https://doi.org/10.1016/j.neucom.2019.08.023

[15] Sezer, O.B. and Ozbayoglu, A.M. (2018) Algorithmic Financial Trading with Deep Convolutional Neural Networks: Time Series to Image Conversion Approach. *Applied Soft Computing*, **70**, 525-538. https://doi.org/10.1016/j.asoc.2018.04.024

[16] 刘萌萌. 基于深度学习的微小元件姿态识别算法研究[D]: [硕士学位论文]. 郑州: 郑州大学, 2020.

[17] 吴新建. 基于卷积神经网络的图像标注算法研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2019.

[18] Liu, J.L., Chao, F., Lin, C.-M., *et al.* (2021) DK-CNNs: Dynamic Kernel Convolutional Neural Networks. *Neurocomputing*, **422**, 95-108. https://doi.org/10.1016/j.neucom.2020.09.005

[19] Li, G.Q., Shen, X.Z., Li, J.J. and Wang, J.Y. (2021) Diagonal-Kernel Convolutional Neural Networks for Image Classification. *Digital Signal Processing*, **108**, 102898. https://doi.org/10.1016/j.dsp.2020.102898

[20] 陈文祺. 基于深度学习的手机移动端视线跟踪算法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2019.

[21] Gu, J.X., Wang, Z.H., Jason, K., *et al.* (2018) Recent Advances in Convolution Neural Networks. *Pattern Recognition*, **77**, 354-377. https://doi.org/10.1016/j.patcog.2017.10.013

[22] Jha, S., Son, L.H., Kumar, R., *et al.* (2019) Neutrosophic Image Segmentation with Dice Coefficients. *Measurement*, **134**, 762-772. https://doi.org/10.1016/j.measurement.2018.11.006

[23] Finch, B.K. and Beck, A.N. (2011) Socio-Economic Status and z-Score Standardized Height-for-Age of U.S.-Born Children (Ages 2 - 6). *Economics and Human Biology*, **9**, 272-276. https://doi.org/10.1016/j.ehb.2011.02.005