**Comparison of Machine Learning Algorithms for Somatotype Classification**

dr. sc. Darko Katović 1 and Miljenko Cvjetko, dipl. ing.

*1 Faculty of Kinesiology, University of Zagreb, Croatia*

*2 Microsoft (Xamarin Inc.), Software Engineer, Zagreb, Croatia*

*darko.katovic@kif.hr, miljenko.cvjetko@microsoft.com*

Keywords**:** Machine learning, Multiclass classifiers, Supervised classification, Somatotype.

Abstract: System modeling (identification) in complex systems like kinesiological and biological in general, is extremely difficult due to the high dimensions of parameters and usually non-linear functional dependencies. Data Science and especially Machine Learning (Deep Learning) algorithms seem to be quite an excellent tool for analysis and problem-solving in sports today. Data Science (Machine or Deep Learning) algorithms rely on basic use of statistical algorithms, but extend those with models such as Decision tree, K-means clustering, Neural networks, and Reinforcement learning, creating new algorithms that handle input data by predicting outputs that describe correlation relations or predict future states at time points (regression). This study attempts to analyze and research applications of machine learning in Sport science - Kinanthropometry related problem of determining somatotype by using the Microsoft Azure Machine Learning platform and comparing several supervised classifier algorithms (Multiclass Neural Network, Multiclass Decision Forest, Multiclass Decision Jungle and Multiclass Logistic Regression) which were compared versus classical somatotype categorization algorithms with dataset based on the Heath-Carter method Somatotype determination to gain experience and expertise.

1. **INTRODUCTION**

Some 30-40 years ago, mathematicians and computer scientists formalized some methods that try to model principles of human thinking - the brain. This area is called Artificial Intelligence – AI which includes logical systems like Expert (Knowledge-Based) Systems and Fuzzy Logic, then Genetic Algorithms, Machine Learning (with Deep Learning), Vision with Pattern Recognition and Language Processing (written and native) and much more. The fundament for Neural Networks and parts of data science called Machine Learning, which involves Deep Learning (supervised and unsupervised) is the brain's physical structure. Logical systems that model human thinking are described through Expert Systems and Fuzzy Logic, while some other biological behaviors can be represented and modelled through Genetic Algorithms. Though theory and math behind this field are far from trivial, and this area is not new for mathematicians, data or computer scientist, for average user AI might look very sophisticated, scary, and repellent. One of the reasons is nomenclature, which might be confusing. For example, 20 years ago, this area was called soft computing, and today hype buzzwords like data science and machine learning are used interchangeably thus confusing end-users and non-scientists.

The advancement of the software industry has made it possible to use (Neural Networks, Machine Learning, Deep Learning, etc.) software tools that implement complex mathematical algorithms using easily accessible platforms and software (free, commercial), leaving users alone with tools in complex scientific fields. On the other hand, hardware has evolved to the point that complex computing is possible, even on PCs and some smartphones.

Today data acquisition can be done with almost every object (device) that is in some form of interaction with an athlete (or team), by passively following its movement and transmitting information on the subject's state, its position, change in speed over time or the force it transmits (on the background, to another object, or even another participant in interaction).

Many sensors, so-called "edge devices" are appropriately integrated into the subject's clothing, footwear, or are in contact with the subject's skin and communicate with surrounding systems, which additionally collects physiological and biomechanical data. IT protocols in real-time transmit information (data) from IoT (edge) devices to the cloud, and they are in general, rarely used.

System modeling (identification) in complex systems like kinesiological and biological in general, is challenging due to the high dimensions of parameters and usually non-linear functional dependencies. Data Science and especially Machine Learning (Deep Learning) algorithms seem to be quite a useful tool for analysis and problem-solving in sports today. To increase the accuracy of conclusions possible by application of data science and machine learning one needs large amounts of data that must be integrated with complex algorithms and processing.

Today, there is a global race in the area of implementing categories of algorithms known as Machine Learning (ML) and Deep Learning (DL). Hence, specialized applications are able to reach greater accuracy in classification, regression or prediction (target event forecast), and these tools are becoming available to every user.

Data Science (Machine or Deep Learning) algorithms rely on the basic use of statistical algorithms, but extend them with models such as Decision tree, K-means clustering, Neural networks, and Reinforcement learning, thus creating new algorithms that handle input data by predicting outputs describing correlation relations or predicting future states at points in time (regression).

Given the fact that data hype, big data and data science are not merely buzzwords but reflect the reality, the main idea behind this work is to experiment with AI algorithms, compare specific AI algorithms with other AI approaches, as well as known deterministic (exact) algorithms and prepare methodology to explain them to end-users in social sciences like kinesiology, practitioners like medical personnel in health sector, and coaches and trainers in sports and fitness.

The initial work with the existing software implementation showed some implementation flaws in various software systems. Obtained results led to a decision to implement several versions of Somatotype classification algorithms – Heath Carter and machine learning algorithms. Machine learning algorithms started as a parallel investigation during the implementation of the Heath Carter algorithm with data available, but it was insufficient, so data acquisition continued in the spring of 2019.

Based on the data available, the first step was to investigate classification algorithms for somatotype classification. Somatotype classification belongs to a group of multi-class classification, so a simple comparison of multi-class machine learning algorithms is given in this article without a more in-depth analysis of data for training and validation.

The investigation of other AI algorithms for mathematical system modelling or prediction and planning, such as regression is left as a goal set for the future.

1. **System theory**

One of the main goals in sports, fitness, and health is to bring a particular human being (an athlete or a patient) from one state, called initial state, into the other state, called final state. The principle is the same even if the final state is the same as the initial state; in this case, one can talk about maintaining a fitness state or health state. This can be done with the supervision of the coach, trainer, instructor, therapist, doctor, or team of persons.

If one would resort to abstraction and think of that very athlete or the patient as a system, it would become evident that these concepts represent fundamentals of control theory, with control theory being based on the systems theory, which is in turn based on mathematical (formal) modelling.

The system created by the transformation from one state into another is called a controlled system, and it represents or models the athlete or the patient, whereas the coach or the physician represent or model a controlling system. From the point of view of terminology, it is evident that scientific (formal) approach requires a mathematical model of both the controlled and the controlling system. Moreover, this is where historical development had tremendous obstacles. Namely, humans as biological systems have highly complex, often non-linear, transient mathematical models, which exhibit characteristics of several if not all system model types such as continuous, time discrete and even event discrete models.

The transforming system's state can be performed either through feed-forward control or feedback control. In a feed-forward control input signal is applied to the controlled system, and output is the function of the mathematical model of the system applied to the input signal (see Fig. 1).

Figure 1: Feed-forward control

In a feedback system, there is "interaction" of the controlled and controlling system. The output of the mathematical model function of the controlled system applied on the output is observed/measured by the controlling system as it inputs signal and applies its function on its input signal being the output signal of the controlled system. The resulting output signal of the controlling system is mixed or directly fed on the controlled system input (see Fig. 2).

Figure 2:Feedback control

In both cases, mathematical function or the model of the controlled system (f) is required for various reasons - such as accuracy, precision, stability, and proofing. This falls under the responsibility of system identification, being the part of system theory in general.

This work attempts to compare system models based on a deterministic and well-known Heat-Carter Anthropometric Somatotype formula to those based on machine learning algorithms.

1. **Literature review**

Looking at sports studies, the Machine learning topics can be summarized within few categories: the prediction of a game outcome (Bunker and Thabtah, 2019), (Panjan, Sarabon and Filipcic, 2010), (Sipko, 2015), (Torres and Hu, 2013), the prediction, developing and improving performances of teams or individual players (Gombolay, Jensen and Son, 2017), (Keim *et al.*, 2017), and the classification, modeling, planning and selection of competitive strategies (Meżyk and Unold, 2011), (Miller, 2016).

The present use of machine learning does not make kinanthroplogical studies an exception. Thus, typical steps for the determination of the body morphology and composition relate to mathematical formulas based on the Heath and Carter methodology as described by Carter (Carter and Heath, 2002). The mentioned technique also enables the determination of the body morphology and composition associated with specific health issues or sports activity.

Despite a limited number of applications, there are different types of approaches explaining the importance of the body structure determination dependency with either some aspects of sports result success (Houcine, Ahmed and Saddek, 2014), (Ramos-Jiménez *et al.*, 2016), (Tóth *et al.*, 2014), abilities to perform physical activity (Ryan-Stewart, Faulkner and Jobson, 2018), (Willgoose and Rogers, 1949), or health issues (Koleva, Nacheva and Boev, 2000), (Koleva, Nacheva and Boev, 2002), (Malina *et al.*, 1997).

1. **MATERIALS AND METHODS**

A comparison of algorithms in this study was made by using Microsoft Azure Machine Learning Studio[[1]](about:blank).

The research covered the following algorithms being the part of Microsoft Azure Machine learning Studio (Barnes, 2015): Multiclass Neural Network, Multiclass Decision Forest, Multiclass Decision Jungle and Multiclass Logistic Regression (Barnes, 2015). These machine learning comparison algorithms have been compared versus simplified classical somatotype categorization (central, endomorph, ecto-endomorphic, mesomorphic, meso-ectomorph, endo-mesomorphic, and ectomorphic) algorithms based on the Heath-Carter method of Somatotype determination (Carter and Heath, 2002).

1. **Data**

Due to the lack of the required amount of data needed to test the model ratings, the somatotype categorization used has been generated based on the parameters (Table 1) from a somatotype study on adolescents (Subramanian et al., 2016).

Table 1: Random generated sample

Somatotype

Mean

St.dev.

Max. scale value

endomorph

2.72

1.21

16

mesomorph

2.97

1.21

12

ectomorph

3.33

1.13

9

The size of such randomly generated, standardly distributed sample is n=1000 (round=0.5).

Due to a later evaluation of the model, a given dataset was split for analysis into a training (75%) and testing (25%) subsets of data (Microsoft Azure Machine Learning Studio (MAMLS) parameters: Splitting mode = Split Rows, Fraction of rows in the first output dataset = 0.75, Randomized split).

1. **Algorithms**

Classification in Machine Learning, in general, is a technique of learning, where an instance is mapped to one of many labels. In multiclass classification, the goal is to archive classification in more than two classes. By using selected classification algorithms, the machine learns patterns from data in such a way that the learned representation successfully maps the original dimension to the suggested class without any intervention by a human expert.

Multiclass Neural Network (*Multiclass Neural Network - Azure Machine Learning Studio | Microsoft Docs*, no date) node is used to build a multiclass model based on a feedforward artificial neural network. The feedforward artificial neural network adopts a unidirectional multi-layer structure. Each layer contains several neurons, and the neurons of the same layer are not interconnected. Inter-layer information transmission is unidirectional.

Multiclass Decision Forest (*Multiclass Decision Forest - Azure Machine Learning Studio | Microsoft Docs*, no date) works by building multiple decision trees and then voting on the most popular output class. Voting is a form of aggregation, in which each tree in a classification decision forest outputs a non-normalized frequency histogram of labels. The aggregation process sums these histograms and normalizes the result to get the "probabilities" for each label. The trees that have high prediction confidence have a greater weight in the final decision of the ensemble.

Multiclass Decision Jungles (*Multiclass Decision Jungle - Azure Machine Learning Studio | Microsoft Docs*, no date) (Shotton *et al.*, 2013) represent a recent extension to decision forests. Their advantages include a lower memory footprint and a better generalization performance than that of a decision tree (which results in somewhat higher training time). It should also be mentioned that Decision Jungles are non-parametric models that can represent non-linear decision boundaries; they perform integrated feature selection and classification and are resilient in the presence of noisy features.

Multiclass Logistic Regression (*Multiclass Logistic Regression - Azure Machine Learning Studio Microsoft Docs*, no date) use classifiers that can be used to predict multiple outcomes. The multiclass classification problem can be solved by naturally extending the binary classification technique for some algorithms. These include neural networks, decision trees, k-Nearest Neighbor, Naive Bayes, and Support Vector Machines (Aly, 2005). While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms, which can, however, be turned into multinomial classifiers by a variety of strategies.

To understand the results obtained better, it is necessary to explain the terms through which they are expressed: precision, recall, and accuracy. All three are metrics for evaluating classification models.

It is commonly thought how both precision and recall indicate the accuracy of the model. For a more lucid interpretation, it must be emphasized that precision (1) expresses the proportion of the data points for a given model and their actual relevance, and recall (2) expresses the ability to find all relevant instances in a dataset. Accuracy (3), of course, explains the correctness of the classification model (*Precision vs. Recall - Towards Data Science*, no date).

(1)

(2)

(3)

1. **RESULTS AND DISCUSSION**

With a view to gaining experience and expertise by using the Microsoft Azure Machine Learning platform, several supervised classifier algorithms have been compared. Machine Learning Algorithms combined with modern tools that implement them offer quite a simple problem-solving framework, but without deeper understanding and inadequate datasets can lead to wrong conclusions.

To understand better, let us recall the goal: Classification of seven somatotypes (﻿simplified classification) based on sampled data (sampled data size n=1000). The evaluation of multiclass classifiers has been made by using precision, recall, and accuracy metrics. A new understanding of data requires further analysis of micro-precision and micro recall.

Table 2: Classification algorithm comparisons

Algorithm

Precision

Recall

Accuracy

Multiclass Neural Network

0.848399

0.724194

0.981714

Multiclass Decision Jungle

0.744827

0.765457

0.985143

Multiclass Logistic Regression

0.200942

0.283972

0.915429

Multiclass Decision Forest

0.765841

0.732045

0.977143

The results in Table 2. and their parallel comparisons indicate that the Multiclass Decision Jungle algorithm has the highest accuracy of all algorithms (for this type of data).

Furthermore, it can also be seen that the model created by using Multiclass Neural Network has the best (macro) precision, whereas the same model accuracy is marginally lower than the Multiclass Decision Jungle model.

An additional technique for summarizing the performance of a classification algorithm includes the analysis of Confusion matrix (error matrix). It enabled a better understanding of what types of errors (Type I or Type II) the algorithm is making and it can be used for describing the performance of a classification model on a set of test data for which there are known correct values.

Figure 3: Multiclass Neural Network Confusion Matrix

The main diagonals of the Multiclass Neural Network Confusion Matrix and Multiclass Decision Jungle Confusion Matrix (Figure 3 and Figure 4) follow the conclusions (about the model choice) presented earlier and additionally assist in selecting a model.

Figure 4: Multiclass Decision Jungle Confusion Matrix

Figure 5: Multiclass Logistic Regression Conf. Matrix

Figure 6: Multiclass Decision Forest Confusion Matrix

Multiclass Decision Forest Confusion Matrix (Figure 6) points to a somewhat weaker model precision, while the Multiclass Decision Forest Confusion Matrix (Figure 5) additionally confirms unacceptable deviations in the classification.

The research itself has not gone further into optimizing and tweaking machine learning algorithms in order to achieve better performance (precision and speed). This can be partly seen as a limiting factor of this study and is to be overcome in the future.

1. **CONCLUSION**

As Machine and Deep Learning are entirely new and complex fields both in science and technology, the intention of the paper was to start small, with available data, and compare four somatotype data classification models.

The data for models obtained by machine learning have been compared with software implementation of deterministic Heath-Carter formula for anthropometric somatotype.

Study results show that even with their default settings, some of the classification models are already close to the desired accuracy.

Optimizations and comparison with deterministic somatotype classification algorithms like Heath-Carter remains however a topic of further research, together with new applications like prediction and regression.

It may be concluded that machine learning algorithms and other algorithms used in data science could make modeling of complex biological systems, like humans in sports and fitness, more easy. However, experts performing modeling should be aware that machine learning algorithms depend on input data, and in numerous cases "garbage in" will lead to "garbage out". Translated to sports, the said might imply that improper input (training stimuli) in cases of incorrect model can lead to wrong conclusions.

The implementation of the Heath Carter algorithm with its non-linear functional dependencies proved that machine learning could provide more insights into the Heath Carter algorithm itself.

The morphologic somatotype classification module currently has two implementations – exact Heath Carter implementation (three algorithms) and ML implementation. Both variations in their first step map ten anthropometric variables mapped into 3-dimensional numeric representation, and in a subsequent step a 3-dimensional vector is mapped into a somatotype class. The second step is similar to the HelloWorld sample of machine learning – Iris classification.

The step of mapping anthropometric data to numeric vector revealed issues with some of the current implementations.

The morphological somatotype classification software module is just one of the modules of a more extensive software system implementing other more extensive areas of kinesiology and sports theory, such as data acquisition, modelling, analysis, as well as planning and programming. Current efforts are focused on adding components for data acquisition that opens a way to more tests and further research.

**REFERENCES**

Aly, M. (2005) ‘Survey on Multiclass Classification Methods,' *Technical Report, Caltech*, pp. 1–9.

Barnes, J. (2015). *Microsoft Azure Machine Learning*, *Microsoft Azure*.

Bunker, R. P., and Thabtah, F. (2019) ‘Applied Computing and Informatics A machine learning framework for sport result prediction,' *Applied Computing and Informatics*. King Saud University, 15(1), pp. 27–33. doi: 10.1016/j.aci.2017.09.005.

Carter, J. E. L., and Heath, B. H. (2002) *The Heath-Carter Anthropometric Somatotype - Instruction Manual*, *The Heath-Carter Anthropometric Somatotype - Instruction Manual*.

Gombolay, M. C., Jensen, R. and Son, S. H. (2017) ‘Machine Learning Techniques for Analyzing Training Behavior in Serious Gaming,' *IEEE Transactions on Computational Intelligence and AI in Games*, pp. 1–12. doi: 10.1109/TCIAIG.2017.2754375.

Houcine, A., Ahmed, A. and Saddek, Z. (2014) ‘Designing a Software to Count the Body Composition and Somatotype and its Role in Pursing the Morphological State of Sportsmen,' *AASRI Procedia*, 8, pp. 38–43. doi: 10.1016/j.aasri.2014.08.007.

Keim, D. *et al.* (2017) ‘How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects,' *Data*, 2(1), p. 2. doi: 10.3390/data2010002.

Koleva, M., Nacheva, A. and Boev, M. (2000) ‘Somatotype, nutrition, and obesity.', *Reviews on environmental health*. Germany, 15(4), pp. 389–398.

Koleva, M., Nacheva, A. and Boev, M. (2002) ‘Somatotype and disease prevalence in adults.', *Reviews on environmental health*. Germany, 17(1), pp. 65–84.

Malina, R. M. *et al.* (1997) ‘Somatotype and cardiovascular risk factors in healthy adults,' *American Journal of Human Biology*. John Wiley & Sons, Ltd, 9(1), pp. 11–19. doi: 10.1002/(SICI)1520-6300(1997)9:1<11: AID-AJHB3>3.0.CO;2-T.

Meżyk, E. and Unold, O. (2011) ‘Machine learning approach to model sport training,' *Computers in Human Behavior*. Pergamon, 27(5), pp. 1499–1506. doi: 10.1016/J.CHB.2010.10.014.

Miller, T. W. (2016) *Sports analytics and data science: winning the game with methods and models*. Available at: http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1601557.

*Multiclass Decision Forest - Azure Machine Learning Studio | Microsoft Docs* (no date). Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-forest (Accessed: 29 July 2019).

*Multiclass Decision Jungle - Azure Machine Learning Studio | Microsoft Docs* (no date). Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle (Accessed: 29 July 2019).

*Multiclass Logistic Regression - Azure Machine Learning Studio | Microsoft Docs* (no date). Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-logistic-regression (Accessed: 29 July 2019).

*Multiclass Neural Network - Azure Machine Learning Studio | Microsoft Docs* (no date). Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-neural-network (Accessed: 29 July 2019).

Panjan, A., Sarabon, N. and Filipcic, A. (2010) ‘Prediction of the successfulness of tennis players with machine learning methods,' *Kinesiology*, 42, pp. 98–106. Available at: http://hrcak.srce.hr/file/82650?origin=publication\_detail.

*Precision vs. Recall - Towards Data Science* (no date). Available at: https://towardsdatascience.com/precision-vs-recall-386cf9f89488 (Accessed: 29 July 2019).

Ramos-Jiménez, A. *et al.* (2016) ‘Body Shape, Image, and Composition as Predictors of Athlete's Performance,' in *Fitness Medicine*. InTech, pp. 1–18. doi: 10.5772/65034.

Ryan-Stewart, H., Faulkner, J. and Jobson, S. (2018) ‘The influence of somatotype on anaerobic performance,' *PLoS ONE*, 13(5), pp. 1–11. doi: 10.1371/journal.pone.0197761.

Shotton, J. *et al.* (2013) ‘Decision Jungles: Compact and Rich Models for Classification,' in Burges, C. J. C. et al. (eds) *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pp. 234–242. Available at: http://papers.nips.cc/paper/5199-decision-jungles-compact-and-rich-models-for-classification.pdf.

Sipko, M. (2015) ‘Machine Learning for the Prediction of Professional Tennis Matches,' *MEng Thesis*, pp. 1–64. doi: 10.1016/B978-008044570-0/50134-3.

Subramanian, S. K. *et al.* (2016) ‘Somatotyping in Adolescents: Stratified by Sex and Physical Activity,' *International Journal of Anatomy & Applied Physiology*, 2(301), pp. 32-38. doi: 10.19070/2572-7451-160005.

Torres, R. A., and Hu, Y. H. (2013) *Prediction of NBA games based on Machine Learning Methods*.

Tóth, T. *et al.* (2014) ‘Somatotypes in sport', *Acta Mechanica et Automatica*, 8(1), pp. 27–32. doi: 10.2478/ama-2014-0005.

Willgoose, C. E. and Rogers, M. L. (1949) ‘Relationship of Somatotype to Physical Fitness', *The Journal of Educational Research*. Routledge, 42(9), pp. 704–712. doi: 10.1080/00220671.1949.10881739.

[[1]](about:blank)https://studio.azureml.net