An AI-based Assistance System for Determining the Risk of Disease and for Preventive Measures

Samira Maleki[1], *Nasser Jazdi-Motlagh*[1]

[1] *Institute of Industrial Automation and Software Engineering (IAS), University of Stuttgart, Pfaffenwaldring 47, 70550 Stuttgart, Germany*

ABSTRACT

Prevention of widespread diseases can make an important contribution to improving the quality of human life. Furthermore, disease prevention can serve to avoid future demands for medical rehabilitation due to demographic change. In this paper, a literature review on the state of the art in disease prevention through machine learning will be presented first. Subsequently, it was concluded that no previous applications have focused on determining the extent of the influencing factors on the risk of disease and thus identifying preventive measures to reduce the risk of disease. To address this research gap, this paper presents a concept for generating a personalized prediction model for a given disease, using machine learning algorithms for the automated analysis of a wide range of input data. To realize this concept, an assistance system is implemented be presented, which includes prediction models for the three diseases cold, hypertension and hypercholesterolemia to determine disease risks and preventive measures. After entering the user's health data, the assistance system determines the risk for each disease and the preventive measures to reduce the disease risks. Thereafter, the evaluation of the assistance system is presented by testing it on 5 people who used it daily for 4 months.

Keywords: AI-based software system, prediction model, preventive measures

INTRODUCTION AND STATE OF THE ART

Europe is undergoing demographic change for various reasons, including low birth rates in recent decades and rising life expectancy (Stula and Linz, 2010; Frevel, 2013). A characteristic of this phenomenon is the increase in the number of old people. Demographic change, on the one hand, and the expansion of demanding service activities, on the other, are leading to an ever-increasing proportion of elderly workers in the workforce (Allmendinger and Ebner, 2006). As people get older, the desire for a better quality of life also increases. Another expected effect of demographic ageing is the increase in the number of cases in rehabilitation (Nowossadeck, 2019). The increased desire for a better life quality and the use of the medical rehabilitation clearly illustrates the importance of the prevention of widespread diseases as well as the activities to achieve this goal. Thereby, today's progress in digitization and networking in the context of health data collection and data analysis can help to support people in the prevention of widespread diseases. Machine learning (ML) is a common approach to the analysis of large and diverse data sets and the extraction of information from them. From this, the aim of this paper is derived: applying machine learning in the context of a personalized software system for the prevention of diseases. Large amounts of health data, rapid data analysis and strong networking of information systems can be considered as basic prerequisites for the use of machine learning for the prevention of widespread diseases. In this context, it is first necessary to investigate how far these basic prerequisites are fulfilled with current technologies. From the first digital computer in 1941 to the computers developed in recent decades, computing power has increased by a factor of about 1017 (Deckert, 2019). This high performance of today's computers enables the fast processing of large amounts of data. About 47% of the world's population is online (Deckert, 2019). As reported (*Smartwatches*, 2019), the number of networked portable devices worldwide will be 835 million in 2020. Smartwatches and smart bracelets belong to this group. Even if only devices that can be used for medical purposes are considered, the number is still very high. The health data - such as body temperature, resting heart rate, blood pressure, etc. - of people is monitored by 3.7 million active medical devices (AMR, 2020). The greatest challenge is to extract useful information from the data. This volume of data contains valuable information that could significantly improve people's life quality. Through machine learning, predictive information can be obtained from this data. How important this topic is today is shown by the worldwide market promotion of the Internet of Medical Things for 136.8 billion dollars in 2021 (AMR, 2020). To gain an overview of the current state of the art, the following study examined the framework in which machine learning has been applied in medical technology up to now. Applications of machine learning in medicine can be divided into three groups according to their intended use: prevention, diagnostics, and therapy. Table *1* gives an overview of previous applications of machine learning in the field of disease prevention. A survey on the literature revealed that “no previous applications have focused on determining the extent of the influencing factors on the risk of disease”. In medicine, certain causes of common diseases have been proven. If the extent of the individual influencing

factors could be determined, it would be possible to determine the risk of disease and preventive measures to reduce this risk and ideally prevent the occurrence of disease.

Table 1: Machine learning applications for disease prevention (Jin *et al.*, 2009; Peddinti *et al.*, 2017; Luboz *et al.*, 2018; Lundberg *et al.*, 2018; Conner-Simons and Gordon, 2019)

Application	Concept and its implementation
Prevention of hypoxemia	Prediction of hypoxemia risk during surgery using ML algorithms which are trained using surgical records
Prevention of type 2 diabetes	Identification and modeling of the effects of different combinations of influencing factors on type 2 diabetes using an ML-based system
Prevention of heart diseases	Use of ML algorithms to evaluate the electrocardiogram to provide more reliable information to cardiologists
Prevention of pressure ulcers	Support of a biomechanical model by ML-based approach to estimate internal tissue tension and risk of pressure ulcers
Prevention of breast cancer	A Deep-Learning based model was trained using labelled mammograms to recognize the subtle patterns in breast tissue.

A CONCEPT FOR INTELLIGENT DISEASE PREVENTION AND THE REALIZATION

To close the research gap mentioned above, the following four aspects must be considered: (1) Integration of the health data with information on health status - reporting on the presence or absence of diseases; (2) Application of Machine Learning to determine the association between influencing factors and health status; (3) Enabling personalized prevention to achieve more accurate prevention; (4) Consideration of data privacy during the collection and storage of personal data. As mentioned in the last section, a large amount of human health data can now be collected using medical devices, portable and implantable sensors. These include activity trackers such as smart watches, smart textiles or subcutaneous sensors. Some of these sensors can also collect information on health status. In addition, health information can also be collected through medical records, laboratory reports, medical examinations, etc. By combining health data and health status, it is possible to investigate the influence of health data on certain diseases. This data from different systems can be collected using a semantic description and made available within a software system for further analysis. To analyze this data and to identify a pattern between health data and health status, the supervised or semi-supervised learning algorithms of machine learning can be used. For this purpose, prediction models can be created for different diseases. Furthermore, machine learning allows the extension and optimization of these prediction models under continuous usage of the above-mentioned software system. A prediction model is an abstract representation of health data and its influence on a specific disease. Different immune strengths, genetic background and lifestyle of each person lead to the fact that the prediction model of each disease must be personalized for each person. This means that the prediction models for each person must be created on the basis of their existing data and

developed further over time through use. The personalized prediction model not only performs better, but also respects basic privacy policies, as each person's data is stored locally and used only for their own prediction models. To integrate all prediction models into a software system and to personalize them, an assistance system can be used, which is a flexible and adaptable system for information and decision support and can be installed, for example, on a smartphone. Figure 1 illustrates the concept of an assistance system considering the four important aspects mentioned above for intelligent determination of disease risks and preventive measures.

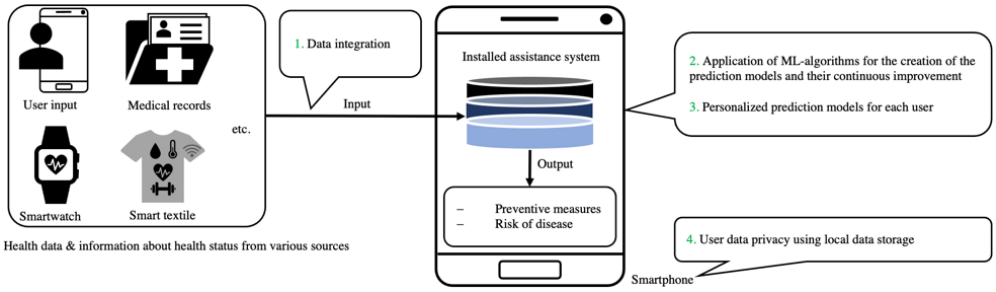

Figure 1:An assistance system for determination of disease risks and preventive measures

To realize the four conceptual aspects mentioned above, an assistance system was developed that enables intelligent and personalized disease prevention. The assistance system must determine the disease risks and preventive measures based on the personal data of the users. For this purpose, a prediction model for each disease using machine learning algorithms and training data - combination of health data and information about the state of health - is integrated into the assistance system. Each prediction model compares the similarity between health data and training data after the user's health data has been entered. The prediction model then determines the risk for each disease. Furthermore, the prediction models suggest prioritized preventive measures based on the extent to which individual health data influence the given risks.

Disease Selection, Influencing Factors and Training Data

For each disease, a prediction model can be created. To implement the concept in a prototype, the focus was on creating prediction models for three diseases. Two criteria were considered in the selection of the diseases: (1) the frequency of the diseases and (2) if possible common main influencing factors exist. The frequency of the diseases in Germany as a representative of the European Union was examined. According to (Thefeld, 2000), almost half of the respondents had a cold once or twice in the last six months of 2018, which led to the cold being selected as the first disease. According to (Kushimoto, Gando and Saitoh, 2014), 24.5% of men and 21.5% of women in Germany suffer from hypertension. For this reason, it was chosen as the second disease. Hypercholesterolemia was chosen as the third disease because according to (Kushimoto, Gando and Saitoh, 2014), 32.2% of men and 34.9% of women in Germany have high cholesterol levels. In order to limit the factors

influencing the selected diseases in the prototype to be developed, ten identical health factors (see Table *2*), were considered to have the greatest impact on these diseases. To create the basis of the prediction models, first simulated data were used. For each disease, the relevant health data and their normal and abnormal limits were determined from the medical literature. These were summarized in Table *2*. Based on the information in Table *2*, two groups of health data are simulated for each prediction model, one with health labels and another with disease labels. It should be considered that the labels represent the corresponding health status of the health data. Hereby "zero" is used as the value for all health labels and "one" as the value for all disease labels. The simulation of health data is performed randomly within the defined limits. Subsequently, irrelevant health data of a disease are simulated in such a way that they do not show abnormal limit values for other diseases.

Table 2: Training data (Middeke, Pospisil and Völker, 2000; Bassenge, Schneider and Daiber, 2005; Volkert, 2006; Vgontzas *et al.*, 2009; Gangwisch *et al.*, 2010; Günster, Klose and Schmacke, 2011; Heidbreder and Young, 2012; Kushimoto, Gando and Saitoh, 2014; Hirshkowitz *et al.*, 2015; Schmidt, 2016; Siedentopp, 2016; Böcker *et al.*, 2019)

Health factors	**Cold**	**Hypertension**	**Hypercholesterolemia**
Gender	irrelevant	↑ M ≥ 55 -↑ F ≥ 65 y.o.	↑ M ≥ 45 - ↑ F ≥ 55 y.o.
Body temp.	↑ 35°C-36,5°C	irrelevant	irrelevant
Resting heart rate	↑ >100 bpm	↑>100 bpm	↑>100 bpm
Activity	↓ >30 min/day	↓ >30 min/day	↓ >30 min/day
Age	↑>60 y.o	↑ M ≥ 55 -↑ F ≥ 65 y.o	↑ M ≥ 45 -↑ F ≥ 55 y.o
BMI	↑ >25 kg/m2	↑ >25 kg/m2	↑ >25 kg/m2
Sleep duration	↑ <6 hours/day	↑ <6 hours/day	↑ <6 hours/day
Water cons.	↑ <1500ml/day	irrelevant	irrelevant
Nicotine cons.	↑ ≥ 1 Ciga./day	↑ ≥ 1 Ciga./day	↑ ≥ 1 Ciga./day
Alcohol cons.	↑ >25 ml/day	↑ >50 ml/day	↑ >50 ml/day

Health factor increases (↑) -decreases (↓)- the risk of the disease within the established limits

Criteria for Selecting the Learning Algorithm and its Functionality

The focus in the development of prediction models is on the influence of health data on specific diseases. Therefore, supervised, or semi-supervised learning can be used as a learning mechanism. Since in this paper all training data are simulated with labels, supervised learning is used as a learning style. For the selection of the learning algorithm, the desired format of the results must first be defined. By using classification, the disease risks can be categorized into different classes. Alternatively, the disease risks can be determined by using the regression in percent.

Since a continuous value in this study can provide a more accurate insight for users, a regression algorithm is used. There are numerous algorithms that can be used for regression tasks. Four algorithms or model types were considered: (1) Random Forest, (2) Neural Network, (3) Linear Regression and (4) Decision Tree. All these algorithms can theoretically be used for the use case of this contribution. Thus, the algorithm was selected according to two criteria: (1) the speed of the prediction model generation and (2) the higher performance. Since health data are simulated with high accuracy, speed was the top priority when selecting the algorithm. In accordance with (Azure, 2020), models with linear regression and decision tree algorithms train faster than those with neural networks and random forest algorithms. To choose between linear regression and decision tree, it is necessary to check whether a linear approximation is acceptable. In (Panesar, 2019) linear approximation is acceptable if y can be calculated from a linear combination of input variables (x_1 to x_n). Formula (1) serves to provide a general overview of this condition. In Formula (1), h(x) represents estimated value of y; (θ_0 to θ_n) are the parameters of the model; x_1 to x_n representing the features or input variables and n finally the number of features. The hypothesis function, uses parameters of the model to estimate how each feature affects y. This means that with a trained model based on linear regression, the extent of the influence of health data on each disease can be determined, which is exactly the desired outcome of this study. Furthermore, (Kim, 2008) confirmed a better performance of linear regression compared to the decision tree for continuous variables. To generate each prediction model a mathematical function is created based on the training data using linear regression with multiple variables (see Formula (2)). In Formula (2), $h_1(x)$ represents the risk of disease; $\theta_{m,n}$ Linear regression parameters and finally x_1 to x_{10} health factors (see Table 2). To simplify presentation and calculation, the linear regression parameters (ϑ_1) of the prediction model are presented as vector. ϑ_1 were determined based on training data. The value of the individual elements of $\vartheta_1$illustrates the extent of the influence of individual health data (x_1 to x_{10}) on the disease risk. Using linear regression parameters can the risks of the diseases for new health data be calculated using Formula (3). X_{new} stands for the vector representation of new health data. To adjust the matrix dimensions, X_{new} was extended by an element (x_0=1).

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \quad (1)$$

$$h_1(x) = \theta_{1,0} + \theta_{1,1} x_1 + \theta_{1,2} x_2 + \ldots + \theta_{1,10} x_{10} \quad , \quad \vartheta_1 = [\theta_{1,0}\ \theta_{1,1}\ \theta_{1,2} \ldots\ \theta_{1,10}] \quad (2)$$

$$h_1(x) = \vartheta_1 \times X_{new} = [\theta_{1,0}\ \theta_{1,1}\ \theta_{1,2}\ \ldots \theta_{1,10}] \times [x_0\ x_{1,new} \ldots\ x_{10,new}]^{-1} \quad (3)$$

The prediction models of all three diseases were calculated in the same way as $h_1(x)$. If the elements of the disease risk formula ($\theta_{n,0}x_1$ to $\theta_{n,10}x_{10}$) are ordered by value, the influencing factors of the disease risk can be presented hierarchically. Consequently, preventive measures will be defined as measures to reduce the values of this ordered parameter. Another aspect that can increase the accuracy of prediction models is the transformation of health data (x_1 to x_{10}) into a comparable magnitude.

For this purpose, standardization was carried out according to the principle described in (Angelov, 2017).

The simulated training data are the basis for the creation of the prediction models. The training data of the associated prediction models are constantly being expanded using the assistance system by the users. Concerning the extension of the training data, it must be considered that due to the selection of supervised learning, the prediction models all training data must have labels. For this reason, the health data that contains information about health status can be used as additional training data. To further develop the prediction models, the regression parameters are updated when the training data is enhanced. Figure 2 gives an overview of the software components of the assistance system developed in MATLAB and Figure 3 illustrates the GUI of the assistance system. This assistance system has two windows: (1) user definition window and (2) continuous prediction and further development window. To use the assistance system, the user must first be defined in the first window. The second window allows the user to use the assistance system permanently.

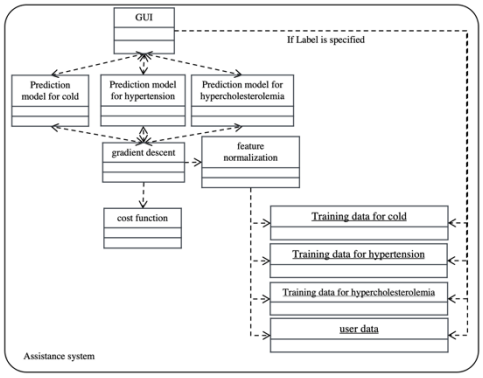

Figure 2: Software components of the system

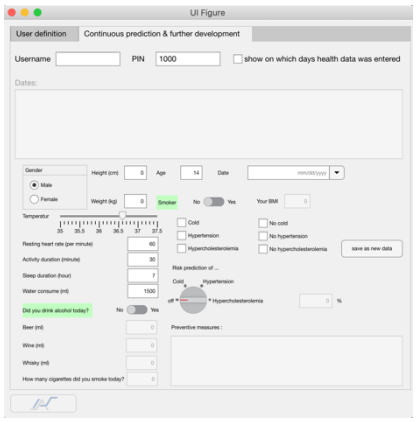

Figure 3: GUI of the assistance system

EVALUATION

To test the assistance system, it was used continuously by five people for four months. During this period the users entered the health data (x_1 to x_{10}) into the assistance system daily. The following points were considered during entering the diseases that occurred: (1) If there were no symptoms of a cold, the user confirmed that he did not have a cold. According to (Horting, 2020), headache, aching limbs, neck pain, runny nose, cough and fever are the main symptoms of the cold; (2) If hypertension or no hypertension was confirmed by a medical examination or a portable blood pressure monitor, the user entered in the assistance system whether or not hypertension was present; (3) If a blood test confirmed hypercholesterolemia or no hypercholesterolemia, the user entered in the assistance system whether or not the disease was present. The regression parameters were constantly updated by the anonymized, registered health data of the users and accordingly the prediction models were constantly improved. The identified disease risks and preventive measures of

the assistance system have become more accurate over time and have reacted better to changes in the input data.

SUMMARY AND OUTLOOK

Based on an overview of the state of the art, a lack of a software system was identified that can automatically determine the extent of the influencing factors on a disease and preventive measures to reduce the risk of disease. To fill this research gap in the state of the art, a concept was presented which allows intelligent disease prevention and the identification of hierarchical preventive measures, under consideration of the following four conceptual aspects: (1) The combination of health data with information on the state of health; (2) Application of machine learning to automatically determine the association between influencing factors and health status. (3) Enabling personalized prevention to achieve more accurate predictions; (4) Consideration of data privacy in the acquisition and storage of personal data. This concept was implemented using an assistance system. Regarding the evaluation, this assistance system was used by five users continuously for four months. The results show that the assistance system became more and more accurate and reacted more efficiently to changes in the input data by extending the training data based on the anonymously stored health data of the users. As part of the extension of the proposed assistance system, it would be possible to automatically collect health data and information on the state of health in the assistance system, which come from multiple sources, without manual input from the user. To achieve this goal, these heterogeneous data must be described semantically and made available for application in Machine Learning.

REFERENCES

Allmendinger, J. and Ebner, C. (2006) 'Arbeitsmarkt und demografischer Wandel'. Hogrefe-Verlag Göttingen, 50(4), pp. 227–239.

AMR (2020) *Allied Market Research*. Available at: https://www.alliedmarketresearch.com/.

Angelov, P. P. (2017) *Empirical Approach to Machine Learning*, *IEEE Transactions on Cybernetics*. doi: 10.1109/TCYB.2017.2753880.

Azure (2020) *Azure*. Available at: https://docs.microsoft.com/de-de/azure/machine-learning/algorithm-cheat-sheet.

Bassenge, E., Schneider, H. T. and Daiber, A. (2005) 'Oxidativer Stress und kardiovaskuläre Erkrankungen', *DMW-Deutsche Medizinische Wochenschrift*. © Georg Thieme Verlag Stuttgart· New York, 130(50), pp. 2904–2909.

Böcker, W. *et al.* (2019) *Lehrbuch Pathologie*. Elsevier Health Sciences.

Conner-Simons, A. and Gordon, R. (2019) 'Using AI to predict breast cancer and personalize care'. MIT News.

Deckert, R. (2019) *Digitalisierung und Industrie 4.0*. Springer.

Frevel, B. (2013) *Herausforderung demografischer Wandel*. Springer-Verlag.

Gangwisch, J. E. *et al.* (2010) 'Short sleep duration as a risk factor for hypercholesterolemia', *Sleep*. Oxford University Press, 33(7), pp. 956–961.
Günster, C., Klose, J. and Schmacke, N. (2011) *Chronische Erkrankungen*. Schattauer. Available at: www.versorgungs-report-online.de.
Heidbreder, A. and Young, P. (2012) 'Auch tagsüber immer schläfrig'. Springer, 13(10), pp. 67–74.
Hirshkowitz, M. *et al.* (2015) 'National Sleep Foundation's sleep time duration recommendations: methodology and results summary', *Sleep health*. Elsevier, 1(1), pp. 40–43.
Horting, M. (2020) *Die Symptome einer Erkältung*. Available at: https://www.erkaeltung-online.de/symptome/.
Jin, Z. *et al.* (2009) 'HeartToGo: a personalized medicine technology for cardiovascular disease prevention and detection', in. IEEE, pp. 80–83.
Kim, Y. S. (2008) 'Comparison of the decision tree, artificial neural network, and linear regression methods', *Expert Systems with Applications*. Elsevier, 34(2), pp. 1227–1234.
Kushimoto, S., Gando, S. and Saitoh, D. (2014) 'Körpertemperatur', *Journal Club AINS*. © Georg Thieme Verlag, 3(01), pp. 31–33.
Luboz, V. *et al.* (2018) 'Personalized modeling for real-time pressure ulcer prevention in sitting posture', *Journal of tissue viability*. Elsevier, 27(1), pp. 54–58.
Lundberg, S. M. *et al.* (2018) 'Explainable machine-learning predictions for the prevention of hypoxaemia during surgery', *Nature biomedical engineering*. Nature Publishing Group, 2(10), p. 749.
Middeke, M., Pospisil, E. and Völker, K. (2000) 'Bluthochdruck senken ohne Medikamente'. Thieme, Stuttgart.
Nowossadeck, E. (2019) 'Einfluss der demografischen Alterung auf die Inanspruchnahme der medizinischen Rehabilitation in Deutschland', *Die Rehabilitation*. © Georg Thieme Verlag KG, 58(02), pp. 96–103.
Panesar, A. (2019) *Machine Learning and AI for Healthcare*, *Machine Learning and AI for Healthcare*. doi: 10.1007/978-1-4842-3799-1.
Peddinti, G. *et al.* (2017) 'Early metabolic markers identify potential targets for the prevention of type 2 diabetes', *Diabetologia*. Springer, 60(9), pp. 1740–1750.
Schmidt, M. (2016) 'Bewegungstherapie und Rehabilitation', *Manuelle Medizin*. Springer, 54(1), pp. 46–49.
Siedentopp, U. (2016) 'Wasser', *Deutsche Zeitschrift für Akupunktur*. Springer, 59(4), pp. 45–49.
Smartwatches (2019). Available at: https://www.statista.com/study/36038/smartwatches-statista-dossier/.
Stula, S. and Linz, K. (2010) 'Demografischer Wandel in Europa'. DEU.
Thefeld, W. (2000) 'Verbreitung der Herz-Kreislauf-Risikofaktoren Hypercholesterinämie, Übergewicht, Hypertonie und Rauchen in der Bevölkerung'. Springer, 43(6), pp. 415–423.
Vgontzas, A. N. *et al.* (2009) 'Insomnia with objective short sleep duration is associated with a high risk for hypertension', *Sleep*. Oxford University Press, 32(4), pp. 491–497.
Volkert, D. (2006) 'Der Body-Mass-Index (BMI)', *Aktuelle Ernährungsmedizin*. © Georg Thieme Verlag KG Stuttgart· New York, 31(03), pp. 126–132.