From Limited Data to Meaningful Insights

Two studies on chronic stress prediction using machine learning

Doctoral thesis

to obtain a doctorate (PhD)

from the Faculty of Medicine

of the University of Bonn

Arezoo Bozorgmehr

from Shahreza, Iran

2024

Written with authorization of

the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. med. Birgitta Weltermann MPH(USA)

Second reviewer: Dr. rer. nat. Javad Ghofrani, University of Luebeck (Germany)

Day of oral examination: 06/11/2023

From the Institute of General Practice and Family Medicine

Director: Prof. Dr. med. Birgitta Weltermann MPH(USA)

Table of Contents

List of abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
DEGS1	German Health Interview and Examination Survey
DNN	Deep Neural Networks
DT	Decision Tree
EHR	Electronic Health Records
GPs	General Practitioners
KFZA	Short questionnaire for Workplace Analysis
KNN	K-Nearest Neighbors
LR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perceptron
MSIMI	Mental Stress-Induced Myocardial Ischemia
PPV	Positive Predictive Value
PrAs	Practice Assistants
PSS	Perceived Stress Scale
RF	Random Forest
ROC	Receiver Operating Characteristic
SES	Socioeconomic Status

SES	Socioeconomic Status
SHAP	Shapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TICS-SSCS	Trier Inventory for Chronic Stress
XGBoost	eXtreme Gradient Boosting

1 Introduction

1.1 Overview on Machine Learning

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on developing algorithms and models capable of automatically learning patterns from data. The goal is to able predictions or decisions without the need for explicit programming. This process trains a model using a provided dataset and then uses this trained model to make predictions or provide actions when presented with new, unseen data. ML has revolutionized various fields by enabling automated data analysis and decision-making (Sen et al., 2021). One of the critical factors that significantly affects the performance and reliability of ML algorithms is the quality of the data used for training and inference. While the quantity of data has traditionally been emphasized, it is increasingly recognized that the quality of the data is equally, if not more, important. The quality of data refers to its accuracy, completeness, consistency, and relevance to the problem addressed. In the context of ML, high-quality data serves as the foundation for building robust and accurate predictive models. High-quality data ensures accurate and reliable results, enhances the generalizability of models, improves robustness to variability, enables informed decision-making, and addresses ethical considerations. Therefore, rigorous efforts should be made to ensure data quality throughout the data collection, the preprocessing, and the analysis stages in order to build effective and trustworthy ML models (Liu et al., 2016). ML algorithms can be broadly categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning (Bonaccorso, 2017; Sen et al., 2021; Jain et al., 2020). Within each category, different learning scenarios and tasks are addressed. Supervised learning is the most common and well-studied category of ML. In this type of learning, the algorithm learns from labeled examples, where corresponding target labels or outcomes accompany the input data. The goal is to train a model that can accurately predict the labels for unseen or future inputs. These algorithms perform tasks such as classification, regression, and sequence labelling. Several supervised learning algorithms were employed in both studies conducted during this dissertation research, including random forests (RF), support vector machines (SVM), and artificial neural networks (ANN) (Nasteski, 2017).

Unsupervised learning involves learning patterns and structures from unlabeled data. Unlike supervised learning, there are no target labels provided. The algorithm's objective is to discover hidden patterns, relationships, or groupings within the data. It is often used for exploratory data analysis, dimensionality reduction, and clustering. Common unsupervised learning algorithms include k-means clustering, hierarchical clustering, principal Component analysis (PCA), and auto-encoders (Ghahramani, 2004).

In reinforcement learning, an agent is trained to make sequential decisions in an environment to maximize a reward signal. The agent learns through trial and error, receiving feedback in the form of rewards or penalties based on its actions. The goal is to learn an optimal policy that maximizes the cumulative reward over time. Key concepts in reinforcement learning include states, actions, rewards, and the exploration-exploitation trade-off. Reinforcement learning is widely used in areas such as robotics, game playing, and autonomous systems. Well-known reinforcement learning algorithms include Q-learning, policy gradients, and deep Q-networks (DQN) (Sutton und Barto, 2018).

The choice of ML algorithm depends on the nature of the data, the learning task, and the desired outcomes.

1.1.1 Application of machine learning techniques in public health

In the field of medicine, ML has gained significant attention and has the potential to revolutionize healthcare. Some applications of ML in medicine include disease diagnoses, treatment planning, prognostic and predictive analyses, electronic health records (EHR) analyses, medical imaging analyses, and drug discovery or development (Katsis et al., 2017; Kumari und Bhatia, 2022).

For disease diagnoses, ML models train on medical data such as patient records, lab results, and medical images to support in disease diagnosis. For instance, models can be developed to classify medical images (e.g., X-rays, MRIs) to detect diseases like cancer or identify patterns in patient data to assist in diagnosing rare conditions (Siddiq, 2020; Ahsan et al., 2022; Iqbal et al., 2021).

ML algorithms can assist healthcare professionals in treatment planning by creating personalized treatment plans for patients. By analyzing patient data including medical history, genetics, and treatment outcomes, models can suggest optimal treatment options and dosage recommendations tailored to individual patients.

For prognostic and predictive analysis, ML can predict patient outcomes and provide insights into disease progression. Models can identify risk factors, predict disease progression, estimate patient survival rates, and guide treatment decisions (Zhenzhen et al., 2020).

EHR analyses use ML to extract information from large-scale patient data. This includes identifying disease patterns, predicting patient readmission rates, detecting adverse drug reactions, and optimizing healthcare resource allocation (Samad et al., 2019; Yuan et al., 2021).

ML algorithms train on medical images to assist in image interpretation and diagnosis. For instance, models can detect anomalies in radiology images, segment specific organs or tissues, and aid in the early detection of diseases (Apostolopoulos et al., 2020).

In drug discovery and development, ML can accelerate the process. Models can analyze large volumes of biological and chemical data to identify potential drug candidates, predict their efficacy, and optimize drug formulation and dosage (Patel et al., 2020).

These are just a few examples of how ML is applied in the field of medicine. The use of ML techniques has the potential to improve medical decision-making, enhance patient outcomes, and advance medical research and treatment strategies. It is important to note that ML in medicine requires careful consideration of ethical, legal, and privacy concerns. The use of sensitive patient data must adhere to strict regulations to ensure patient confidentiality and data security.

1.1.2 Impact of sample-size and quality of data

The performance of ML models is influenced by the sample-size of the dataset used for training. The size of dataset plays a crucial role in ML. The importance of dataset size stems from its impact on various aspects of model training, performance, and generalization. Here are some key reasons highlighting the significance of dataset size in ML (Ratner, 2017; Dulhare et al., 2020):

- *Model training:* Large datasets provide more examples for the model to learn from, enabling it to capture a broader range of patterns and relationships in the data. This improves the model's ability to make accurate predictions and enhance its overall performance. (Marr, 2016).

- *Generalization:* A well-trained ML model should be able to generalize well to unseen data. Having a larger dataset can build a more robust and generalized model by exposing it to a wider variety of instances, reducing the risk of overfitting to specific examples or noise in the training data (Vemuri, 2020).
- *Improved performance:* Generally, as the dataset size increases, ML models tend to achieve better performance. With more data, models can better estimate the true underlying patterns and parameters, leading to improved accuracy, precision, recall, or other relevant performance metrics.
- *Complex model training:* Complex models, such as deep neural networks (DNN), often require large datasets to learn the intricate features and representations effectively. These models tend to have a high number of parameters, and training them on small datasets can result in overfitting due to the model's capacity to memorize the limited examples (Goodfellow et al., 2016).
- *Rare events and imbalanced classes:* In scenarios where rare events or imbalanced class distributions exist, having a larger dataset helps in capturing sufficient instances of those events or classes. This allows the model to learn their characteristics more effectively and make more accurate predictions (Krawczyk, 2016).
- *Feature exploration and selection:* When working with a large dataset, there is more flexibility for feature exploration and selection. Larger datasets often have more diverse features, providing a broader scope to identify relevant and informative features that contribute significantly to the model's performance (Guyon und Elisseeff, 2003).

Although large datasets offer potential benefits for model learning, generalization, and improved performance, it is crucial to acknowledge the importance of data quality, diversity, and representativeness. Simply having a large dataset does not guarantee superior results. The quality and relevance of the data, including issues of noise, bias, and diversity, play a significant role in model performance. Therefore, careful consideration should be given to data quality in addition to its size (Saha und Srivastava, 2014; Cai und Zhu, 2015).

1.1.3 Challenges of small datasets

In contrast, working with small datasets in ML poses several challenges that can affect the model's performance and generalization ability. Here are some key challenges associated with small datasets (Barnard et al., 2019):

- *Limited sample size:* It is a significant challenge associated with small datasets, as it restricts the number of available training samples. With a small number of examples, the model faces difficulty in learning complex patterns and relationships present in the data.
- *High variability:* Small datasets often exhibit higher variability, meaning that each individual data point can have a significant impact on the model's training and performance. A single outlier or noise in the data can have a disproportionate influence on the model, leading to overfitting or poor generalization.
- *Limited exploration of feature space:* Feature engineering and selection play a crucial role in model performance. However, small datasets often have limited variability in feature values, making it challenging to explore and identify relevant features that can effectively discriminate between different classes or patterns (Cai et al., 2018).
- *Insufficient training signal:* Small datasets may not provide enough training signal for the model to capture complex relationships accurately. This results in reduced predictive power and lower performance compared to models trained on larger datasets.
- *Overfitting:* It occurs when a model learns to fit the training data too closely, capturing noise or idiosyncrasies that are specific to the limited samples. With small datasets, the risk of overfitting increases, as the model may try to memorize the training samples instead of learning the underlying patterns (Power et al., 2022).
- *Lack of data representation*: Small datasets may lack sufficient representation of all possible variations or classes present in the real-world scenario. This can lead to biased or incomplete learning, limiting the model's ability to generalize to unseen instances accurately.
- *Validation and evaluation challenges:* With limited samples, it becomes challenging to divide the data into separate training, validation, and testing sets while maintaining representative distributions. This can lead to less reliable estimates of model

performance and make it harder to assess the model's ability to generalize to unseen data (Raschka, 2018).

- *Class imbalance:* Imbalanced class distributions, where certain classes have significantly fewer samples than others, are more likely to occur in small datasets. This can lead to biased models that favor the majority class and struggle to accurately predict the minority classes (Elrahman und Abraham, 2013).

Addressing these challenges demands careful consideration and the application of specific techniques tailored to the data size. First of all, the ML method selection plays a crucial role in dealing with limited data (Bonaccorso, 2017; Zhang und Ling, 2018). Ensemble methods like RF, AdaBoost (adaptive Boosting), and extreme gradient boosting (XGBoost) are particularly effective as they combine multiple models (weak learners) to enhance performance and handle small datasets more robustly (Zhang und Ma, 2012; Sagi und Rokach, 2018). By combining predictions from multiple models, ensemble methods outperform single models, leading to improved generalization and prediction accuracy. There are two types of ensemble methods as follows:

- Bagging (Bootstrap aaggregating): It involves training multiple instances of the same learning algorithm on different subsets of the training data. These subsets are created through random sampling with replacement. Each individual model is trained independently, and during prediction, the outputs are combined, typically through averaging or majority voting. Bagging reduces variance and enhances model stability. A well-known example is the RF algorithm.
- Boosting: In this process, the models are trained sequentially, with each subsequent model targeting to correct the errors of the previous models. Data points are weighted based on their difficulty in prediction, with challenging examples receiving higher weights. This iterative process focuses on improving the model's performance on difficult examples, leading to higher accuracy and better generalization. The popular boosting algorithms are the gradient boosting machine (GBM) and XGBoost.

In addition, regularization techniques such as Lasso or Ridge regression can prevent overfitting and improve generalization. Moreover, to overcome the validation and evaluation challenges, cross-validation techniques like k-fold cross-validation or leave-

one-out is employed. These techniques provide more reliable estimates of model performance on small datasets by systematically partitioning the data and evaluating the model multiple times (Berrar, 2019). Techniques like bootstrapping and Monte Carlo simulations offer information on performance variability and uncertainty.

Another challenge is the class imbalance, which is tackled using various strategies. Data augmentation generates synthetic samples, while domain adaptation adapts models trained on larger datasets to the target small dataset. Techniques like synthetic minority oversampling technique (SMOTE), oversampling and random undersampling help balance class distributions (Elreedy und Atiya, 2019; Chawla et al., 2002). Ensemble-based techniques such as cost-sensitive learning and boosting algorithms effectively handle class imbalance. By considering these factors, researchers can overcome the limitations of small datasets in development of prediction models using ML.

1.1.4 Machine learning model interpretation

Model interpretation refers to the process of understanding and explaining how a trained model makes predictions or decisions. It involves uncovering the relationships between input features and the model's output, identifying the factors or variables that are most influential in driving the predictions, and gaining insights into the decision-making process of the model. It aims to provide human-understandable explanations for the model's behavior, especially in complex and black box models such as DNN or ensemble methods like RF. Christoph Molnar developed a guide for making black box models explainable (Molnar, 2020).

Model interpretation in ML plays an essential role in medicine, driven by the need for transparency, accountability, and trust in healthcare decision-making. In medical applications, where predictions can significantly impact patient outcomes, understanding the factors and reasoning behind model predictions is crucial. It enables healthcare professionals to validate decisions, identify biases or errors, and provide meaningful explanations to patients and stakeholders. This promotes clinical understanding, facilitates effective communication, and empowers practitioners to make informed decisions based on model outputs. In the medical field, where lives are at stake, confidence in ML models is paramount. Interpretability builds trust in predictions, allowing for verification and validation of decisions. It ensures compliance with regulations requiring

explanations for AI system decisions and enables providers to justify actions while adhering to ethical standards. It also detects the error and bias leading to improved accuracy and fairness in healthcare applications.

Interpretable models serve as valuable clinical decision-support tools, assisting healthcare professionals in understanding the reasoning behind diagnoses or treatment recommendations. By providing explanations, models improve the decision-making process, increase confidence in treatment plans, and potentially reduce medical errors. Model interpretation uncovers underlying disease mechanisms and factors by identifying important features or biomarkers, shedding light on complex variable interactions. This knowledge contributes to medical research advancements and the development of targeted interventions (Molnar, 2020; Rudin et al., 2022)

Various techniques and methods can be employed for model interpretation, such as feature importance analysis, partial dependence plots, shapley additive explanations values (SHAP), local interpretable model-agnostic explanations (LIME), and rule extraction (Lundberg und Lee, 2017; Lundberg et al., 2020; Ribeiro et al., 2016; Zafar und Khan, 2021).These techniques aim to provide transparency, explainability, and interpretability to ML models, enabling users to understand and interpret the underlying factors driving the model's predictions or decisions.

1.2 Chronic stress and effects on various diseases

Chronic stress is a long-term state of psychological and physiological arousal caused by ongoing stressors, including financial difficulties, caregiving responsibilities, work pressures, relationship problems, major life changes, and exposure to environmental factors. The cumulative effect of these stressors can overwhelm an individual's coping abilities, leading to chronic stress (McEwen, 2022). When the body experiences chronic stress, the stress response system, which involves the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system, remains activated on a higher or more persisting level than usual. These results in the release of stress hormones like cortisol and adrenaline, increased heart rate, elevated blood pressure, and heightened levels of inflammation in the body. Over time, these physiological changes can contribute to the development of various health problems, including cardiovascular diseases, weakened

immune function, mental health disorders, and impaired cognitive function (Stephens und Wand, 2012).

In addition to physical effects, chronic stress also affects mental and emotional well-being. It can lead to symptoms of anxiety, depression, cognition, sleep patterns, and difficulty concentrating (Marin et al., 2011; Dreher et al., 2019; Datta und Arnsten, 2019; Sanford et al., 2015; Hu et al., 2020). Chronic stress has profound effects on various diseases. It can contribute to the development and exacerbation of conditions across multiple systems in the body. It is associated with various diseases like cardiovascular diseases, diabetes, and cancer (Cohen et al., 2007; Kivimäki und Steptoe, 2018; Eizirik et al., 2008; Dai et al., 2020).

Regarding cardiovascular health, mental stress has emerged as a notable risk factor for coronary artery disease and stroke (Everson-Rose et al., 2014). Acute mental stress can result from various sources such as anger, fear, and job strain, while chronic stress can arise from long-term exposure to work-related stress, low socioeconomic status, and other factors. Both acute and chronic stress can lead to physiological changes that increase the risk of cardiovascular events. The brain has a key role in processing emotional stimuli and triggering the fight or flight response, which can induce myocardial ischemia even without significant coronary obstruction. This condition, known as mental stress-induced myocardial ischemia (MSIMI), can have clinical consequences such as angina, myocardial infarction, arrhythmias, and left ventricular dysfunction. However, MSIMI is often underestimated as it may occur without pain and at lower levels of cardiac work compared to exercise-induced ischemia, primarily due to coronary vasoconstriction and microvascular dysfunction (Vancheri et al., 2022).

Henein et al. investigated the impact of mental stress on cardiovascular health, focusing on endothelial dysfunction as a key factor in atherosclerosis. They found that mental stress disrupts endothelial function through various mechanisms, including increased sympathetic activity and inflammation. The study highlights the need to consider psychosocial factors in preventing coronary artery disease and suggests potential interventions (Henein et al., 2022).

Chronic stress also has become a highly prevalent concern in modern society due to its detrimental effects on individuals' physical and mental well-being, which relates to the health of population and society. Accurate measurement tools are essential for assessing

and addressing this issue effectively. These tools enable researchers and healthcare professionals to identify at-risk individuals, develop interventions, and evaluate the effectiveness of stress management strategies.

1.2.1 Self-assessment questionnaires for chronic stress

Measuring chronic stress is challenging due to its subjective and diverse nature, as well as multiple potential sources. Researchers have developed various approaches to capture the different dimensions of chronic stress. Here are some commonly used instruments:

1. Trier inventory for chronic stress (TICS-SSCS): This tool is a comprehensive instrument developed by Schulz et al. based on the systemic-requirement-resource model of health. TICS-SSCS is designed to assess chronic stress and has been deemed to have high content validity (Schulz et al., 2004).
2. Perceived stress scale (PSS): This instrument assesses an individual's perception of stress. It measures the degree to which situations in one's life are appraised as stressful (Cohen et al., 1983; Klein et al., 2016) .
3. Job content questionnaire (JCQ): The JCQ focuses specifically on work-related stress and assesses various aspects of job demands, control, and support (Karasek et al., 1998).
4. Effort-reward imbalance (ERI) model: This model examines the imbalance between efforts spent at work and the rewards received, which can contribute to chronic stress (Siegrist et al., 2014).
5. Work-related quality of life (WRQoL) scale: This scale assesses the impact of work-related stress on an individual's quality of life, including physical health, psychological well-being, and work-related satisfaction (Simon und Darren, 2007).
6. Life events and difficulties schedule (LEDS): This tool is used to assess the occurrence and impact of life events and difficulties, which can contribute to chronic stress (Brown und Tirril, 1978).
7. Perceived control scale: This scale measures an individual's perception of control over stressful situations, which can influence their experience of chronic stress (Thompson und Schlehofer, 2020)

The Instruments are chosen depending on the study objectives, target population, and the aspects of stress to measure, including its frequency, duration, intensity, and subjective appraisal.

In Germany, the TICS is used in several populations. It provides a structured and reliable approach to measure chronic stress across multiple domains. Using 57 items, this tool assesses nine specific domains including work overload, social isolation, pressure to perform, social overload, excessive demands from work, work discontent, lack of social recognition, social tensions, and chronic worrying of chronic stress. Additionally, a short screening scale called the short screening scale for chronic stress (TICS-SSCS) was developed based on a representative sample of the German population (N = 604) (Petrowski et al., 2012). It consists of 12 items derived from five of the nine stress areas: chronic worrying, work overload, social overload, excessive demands of work, and lack of social recognition. The response format for the TICS-SSCS is a 5-point Likert scale, where on which participants report the frequency of experiencing each stress-related item. This scale ranges from 0 (‘never’) to 4 (‘very often’). This instrument assesses chronic stress experienced over the past three months, allowing for a temporal focus on recent stressors. The TICS-SSCS is a standardized and validated tool. It exhibits strong internal consistency, with a Cronbach's Alpha coefficient = .91 for the 12-item TICS-SSCS for chronic stress, indicating high reliability. Moreover, further analyses demonstrated that individual item reliabilities ranged from .84 to .91, with a mean alpha of .87, reinforcing the instrument's robustness and consistency (Schulz et al., 2004; Schulz und Schlotz, 1999).

1.2.2 Chronic stress in healthcare professionals

Chronic stress in healthcare professionals is a significant concern. Demanding job requirements, long hours, heavy workloads, and intense emotional situations contribute to this stress. It has detrimental effects on their well-being, leading to burnout, depression, anxiety, decreased job satisfaction, and impaired performance. This impacts both their quality of life and patient care. Like other healthcare professionals, general practitioners face unique stressors in their roles, such as critical patients and high patient volumes. If working in their own practice, they also experience business-related stressors like administrative burdens, personnel, and financial responsibilities.

The cross-sectional study by Viehmann et al. explored chronic stress of German general practitioners (GPs) and practice assistants (PrAs) at individual and practice levels. Stress levels of personnel from 136 German general practices were measured with the TICS-SSCS questionnaire. Results showed that female GPs reported the highest stress levels, followed by PAs and male GPs. Approximately 26.3 % of personnel at the practice level reported high stress. More working hours per week were linked to high chronic stress for both GPs and PAs. Importantly, stress levels were higher in these primary care professionals compared to the general population in Germany. The study emphasizes the need for stress reduction strategies at both individual and practice levels, supported by an intra-class correlation coefficient of 0.25. Personal and practice characteristics contribute to chronic stress in GPs and PAs, highlighting the need for targeted interventions in healthcare (Viehmann et al., 2017).

Addressing chronic stress in healthcare professionals is essential for their well-being and the quality of healthcare. Organizations are implementing strategies like support programs and improved work environments to reduce stress, benefiting both individuals and the healthcare system through improved employee satisfaction, reduced burnout, enhanced patient safety, and better healthcare outcomes.

1.2.3 Chronic stress in the German population

Chronic stress is a prevalent issue in the German population, and its impact on mental health was extensively studied. In the first wave of the German health interview and examination survey (DEGS1), which was conducted from 2008 to 2011, self-perceived chronic stress was measured using TICS-SSCS in a large sample of participants aged 18-64 years. The prevalence of chronic stress and its association with various factors were examined (Gößwald et al., 2013). The results showed that women reported significantly higher stress levels compared to men, with 13.9 % of women and 8.2 % of men experiencing high stress levels. There were no significant differences in stress levels among different age groups. However, the prevalence of high stress levels decreased as socioeconomic status (SES) increased, with 17.3 % of individuals with low SES reporting high stress levels compared to 7.6 % of those with high SES. Notably, the difference in stress levels between medium and high SES among women was not significant. The study

also found that individuals with poor social support had a higher prevalence of high stress levels (26.2 %), while those with strong social support had significantly lower stress levels (7 %). Moreover, high stress levels were associated with a higher occurrence of mental health problems such as diagnosed burnout syndrome, sleep disturbances, and depressive symptoms in both men and women. More than half of adults with current depressive symptoms are affected by chronic stress (53.7 %), along with a significant proportion of individuals diagnosed with burnout syndrome (45.9 %) and experiencing sleep disturbances (22.1 %). The study further demonstrated that the prevalence of mental health problems increased with higher stress levels, and the presence of multiple mental health problems was more common among women than men in the context of high stress levels (Hapke et al., 2013; Gößwald et al., 2013).

1.2.4 Prediction of chronic stress using machine learning approaches

ML techniques offer new opportunities for predicting psychological diseases such as chronic stress by analyzing complex patterns and relationships in datasets, surpassing traditional statistical methods. The ML models analyze patterns and relationships in datasets to identify the specific factors contributing to chronic stress, such as workplace conditions, demographic variables, personal characteristics, and lifestyle factors. This knowledge develops targeted prevention and intervention strategies to address chronic stress effectively. Standardized scales like TICS-SCSS have provided a key role in the evaluation of chronic stress in different populations, providing a consistent framework for researchers to assess and compare stress levels.

This dissertation focuses on analyzing chronic stress data measured by the TICS-SCSS to explore the predictive capabilities of ML algorithms in two distinct studies. The first study developed prediction models for chronic stress in PrAs using ML classifiers and compared these with a classical statistical approach. The second study focused on developing an interpretable multiclass ML model to predict chronic stress. These studies identified the impact of various protective and risk factors associated with chronic stress, which is necessary to develop effective protective measures.

2 Material and methods

In this section, I provide a comprehensive overview of the datasets used and the specific methodologies employed in my both research studies. The primary objective is to ensure transparency and replicability of the studies by offering detailed descriptions of the data sources, collection methods, and relevant characteristics of the datasets.

2.1 Study 1: Chronic stress in general practice assistants

2.1.1 Dataset

For the study 1 on chronic stress among PrAs, the dataset was obtained from a cross-sectional study conducted in 2014 among general practices associated with the teaching practice network of the Institute for general medicine, university hospital Essen, Essen, Germany. The study included 764 professionals from 136 practices. The published study specifically focused on two main groups: 214 GPs, including practice owners and employed physicians, and 550 PrAs (Viehmann et al., 2017). The study 1 focused on analyzing chronic stress among 550 PrAs with a predominant representation of females (99.3 %). Consequently, all analytical approaches were applied exclusively to the dataset comprising female subjects (n = 546). PrAs but not GPs are addressed in the analyses because they are the largest professional group within general practices. A very small sample size may present challenges, particularly when dividing the data into training and testing sets. The final dataset with 546 PrAs included information on sociodemographic variables, work characteristics, and chronic stress levels measured by TICS as outcome. The German short questionnaire for workplace analysis (KFZA) was used to assess perceived workload. It is a widely accepted screening tool for psychological stress in the workplace. The KFZA employs closed-ended questions and encompasses dimensions such as work content, resources, stressors, and organizational culture. Positive dimensions yield high scores, indicating favorable aspects, while the stressors dimension yields high scores for negative work aspects. Each dimension comprises various factors and single items, and responses are measured on a Likert scale ranging from 1 (does not apply at all) to 5 (is completely true) (Prümper et al., 1995).

The RF classifier was utilized for feature selection, enabling the identification of the most relevant features for training the data across various ML methods (Saeys et al., 2008). By leveraging this algorithm, which combines multiple decision trees, features were ranked based on their ability to improve the purity of the nodes, measured by the decrease in Gini impurity. The features that exhibited the greatest decrease in impurity were considered the most important, while those with minimal impact were regarded as less significant. In total, 64 input variables were considered, encompassing a wide range of sociodemographic and workplace characteristics.

After performing preprocessing steps, including data normalization, addressing imbalanced classes, and missing data, the dataset was divided into training and validation subsets. To handle missing data, common imputation methods for supervised learning were employed. The k-nearest neighbors (KNN) algorithm was used for imputing missing values, median imputation was applied for continuous variables, and a separate category labeled as ‘unknown’ was created for categorical variables. (Malarvizhi und Thanamani, 2012).

The 10-fold cross-validation technique was employed to divide the dataset into training and validation subsets in order to measure the unbiased prediction accuracy of the ML models (Berrar, 2019). This approach involved splitting the data into ten folds, with each fold serving as the validation set once while the remaining nine folds were used for training. This process was repeated ten times, allowing every fold to serve as the validation set. The choice of ten folds was based on the optimal number suggested in the literature, balancing the time required to complete the testing process with minimizing the bias and variance associated with validation. The oversampling method was utilized as a data balancing technique. This method involves replicating samples from the minority class to achieve a more balanced class distribution in the dataset.

By following this methodology, the ML models were trained and evaluated multiple times, providing a robust assessment of their performance. The training involved iteratively using nine folds for training, while the remaining fold was used for validation. This approach enabled a comprehensive comparison of different ML classifiers, leveraging the training and validation datasets.

2.1.2 Primary outcome

The primary outcome of the study 1 was chronic stress measured using the German short version of the TICS-SSCS questionnaire. Chronic stress score was dichotomized using the median score (TICS = 23), with scores below 23 indicating low level of chronic stress and scores of 23 and above indicating high level of chronic stress.

2.1.3 Comparison of four machine learning and logistic regression models

The chosen four ML classifiers, including RF, SVM, KNN, and ANN, along with logistic regression (LR) as the classical statistical approach, were employed to analyze the factors influencing chronic stress in PrAs.

Each classifier possesses unique strengths and characteristics, enabling exploration of various aspects of the data and identification of significant variables associated with chronic stress.

Random forest: RF is a powerful ensemble method, a type of bagging technique that uses weak learners to create a collection of DT. It combines the predictions of multiple individual DT to enhance the overall predictive accuracy of the model. RF is particularly effective in handling complex interactions among variables by considering different subsets of features and training multiple trees on different subsets of the data. By aggregating the predictions of individual trees, RF can capture a broader range of patterns and reduce the impact of individual tree biases and variances. This ensemble approach improves the model's robustness and generalization ability, making it a powerful tool for predictive modelling tasks (Pavlov, 2000).

The hyperparameter tuning for the RF model in study 1, was constructed using a forest of 1,000 individual trees. The choice of a large number of trees contributes to increased predictive accuracy and enhances the robustness and reliability of the model. One notable advantage of RF is its ability to perform well without requiring extensive hyperparameter tuning compared to other models. For feature selection, the model employed a default approach where a random sample of $\sqrt{n}$ predictors, where n represents the total number of predictors considered, was selected at each node. This selection strategy takes the insensitivity of error rates to the number of features chosen to split each node into account. The predicted probability was obtained by averaging the predictions from all the trees in the forest, resulting in a comprehensive and aggregated estimation of the target variable.

Support vector machines: SVM is a powerful classification algorithm that constructs a hyperplane in the feature space to separate different outcome classes. The objective of SVM is to find the optimal hyperplane that maximizes the margin between classes, thereby improving the model's ability to generalize to new and unseen data. SVM is particularly effective in handling high-dimensional and non-linear datasets. By determining the optimal hyperplane, SVM can accurately classify new instances based on their position in the feature space (Hearst et al., 1998).

In study 1, a LR model is fitted to the output of the SVM to obtain probability estimates. This allows for the estimation of the probability of an instance belonging to a specific class, providing a measure of confidence in the predictions. Furthermore, the SVM classifier utilized the radial basis function (RBF) kernel, which is commonly used for non-linear classification tasks. The training error was set to a very low value of 1.0E-12, indicating a stringent tolerance for misclassification during the training process. The default boundary tolerance of 1.0E-03 was used to determine the hyperplane, balancing the trade-off between model complexity and performance. To ensure accurate probability estimates, the SVM outputs were calibrated using Platt scaling model (Böken, 2021). This additional step refines the probability estimates generated by the SVM and enhances their reliability as well as accuracy. By incorporating the RBF kernel, fine-tuning hyperparameter, and employing calibration models, the SVM classifier in study 1 aimed to provide precise and well-calibrated predictions for the target variable.

K-nearest neighbors: KNN is a non-parametric classification algorithm that utilizes the majority vote of its neighboring instances to classify an object. It makes predictions based on the proximity of instances in the feature space, allowing it to effectively capture local patterns and make accurate predictions. In KNN, the value of k determines the number of neighbors considered for classification (Peterson, 2009; Kramer, 2013). In study 1, KNN was applied with k set to 10 neighbors, which represents the ten closest observations in the multidimensional space. The proximity of these neighbors is determined using the Euclidean distance function. By selecting these nearest neighbors, the KNN model incorporates the local structure and patterns present in the training dataset, enabling it to make reliable classification decisions. This proximity-based approach makes KNN a suitable algorithm for analyzing and predicting data in various domains.

Artificial neural networks: ANN are computational models inspired by the structure and functioning of biological neural networks in the brain. They excel at learning complex patterns and relationships in data through interconnected layers of artificial neurons (Koprinkova-Hristova et al., 2014). ANNs are particularly effective at capturing non-linear and intricate relationships between variables, making them well-suited for solving complex problems and handling high-dimensional data. The architecture of an ANN includes input, hidden, and output layers, allowing it to process and transform input data through weighted connections and activation functions (Suzuki, 2013). In study 1, a multilayer perceptron (MLP) classifier with one hidden layer was applied. The input layer accommodated a number of nodes equal to the sum of the features, enabling the network to process the input data effectively. The output layer provided the final predictions based on the learned relationships. The MLP model was trained using the backpropagation algorithm, which iteratively adjusts the weights and biases of the network to minimize the error between predicted and actual outputs. To optimize the performance of the ANN model, specific hyperparameters were defined. The learning rate value of 0.3 with decay determined the step size at which the weights were updated during training. A higher learning rate can lead to faster convergence, but it may also result in overshooting. The momentum rate of 0.2 was employed for the backpropagation algorithm, which support in accelerating convergence by incorporating a fraction of the previous weight update into the current update.

The chosen values for the learning rate and momentum rate in study 1 fell within the suitable ranges of 0.15–0.8 and 0.1–0.4, respectively. These ranges have been found to be effective for training and convergence in neural network models. By selecting appropriate values for these hyperparameters, the ANN model aimed to learn complex patterns from the data and generalize well, leading to accurate predictions for the target variable. ANN's flexibility, capability to capture non-linear relationships, and ability to learn intricate representations make it a powerful tool for various applications, including image recognition, natural language processing, and predictive modelling.

In addition to the aforementioned ML classifiers, LR was employed as a standard approach for comparative analysis (Wright, 1995). LR, a widely-used statistical method, is well-suited for modelling the relationship between variables and a binary outcome, such as chronic stress in my study (DeMaris, 1995). It examines the influence of various factors

on the likelihood of experiencing chronic stress. All variables identified as significant association to chronic stress in the bivariate analyses were used for the LR model. As a classical statistical method, LR is a suitable tool for understanding the relationship between dependent dichotomous outcomes and independent variables, providing a benchmark for comparison with the ML classifiers used in the study. The programming in study 1 was performed in Python including the Scikit-Learn library.

2.1.4 Model interpretation: Variable rankings in machine learning models

Three of the four ML algorithms used in study 1 have the capability to quantify the importance of the various features in the prediction model.

KNN models do not provide explicit measures of variable importance or coefficients like other models such as LR. Unlike models that estimate explicit coefficients for each variable, KNN makes predictions based on the similarity or distance between data points without explicitly quantifying the impact of individual variables (Taunk et al., 2019).

In SVM models, the identification of support vectors is a key factor to determine the decision boundaries. Support vectors are the data points that are closest to the decision boundary and have the most influence on the model's classification or prediction. As a result, the variables associated with these support vectors are considered more significant in separating different classes or making accurate predictions. SVM models aim to find the optimal hyperplane that maximally separates the data points of different classes. This hyperplane is defined by a subset of support vectors, and their associated variables contribute to the model's decision-making process. By examining the weights or coefficients assigned to these variables in the SVM model, we can infer their importance in influencing the classification outcome (Cortes und Vapnik, 1995).

It is important to note that the interpretation of variable importance in SVM can vary depending on the type of SVM (e.g., linear, kernel-based) and the specific Kernel function used. In linear SVM, the coefficients directly indicate the contribution of each variable to the decision boundary. In kernel-based SVM, the interpretation of variable importance can be more complex as it involves transforming the data into a higher-dimensional space (Cristianini und Shawe-Taylor, 2012).

In ANN models, the importance of variables can be assessed by examining the weights assigned to the input variables (Olden et al., 2004). The weights represent the strength of

the connections between the input layer and the hidden layers or the output layer. Variables with higher weights are considered more influential as they have a greater impact on the network's computations and decision-making process. During the training phase of an ANN, the weights are adjusted iteratively to minimize the difference between the predicted output and the actual output. This optimization process assigns higher weights to variables that contribute more to the network's ability to accurately capture the underlying patterns in the data. By analyzing the weights of the input variables, we can infer their importance in influencing the predictions made by the ANN. Variables with larger weights play a more significant role in the network's computations and are more strongly associated with the target variable or the desired output. Conversely, variables with smaller weights have less influence on the predictions. It is worth noting that the interpretation of variable importance in ANNs can be complex, especially in DNN with multiple hidden layers. In such cases, the importance of variables can be distributed across different layers, making it challenging to attribute importance to individual variables. However, by examining the weights in the input layer or analyzing the feature maps and activations in the hidden layers, the relative importance of each variable in the network's decision-making process is quantified.

RF models provide a feature importance metric based on the mean decrease in Gini index (impurity) (Altmann et al., 2010). The impurity measures the disorder or randomness within each tree node, and the reduction in impurity represents the effectiveness of a variable in making accurate predictions. By calculating the average decrease in impurity across all trees in the RF, the model determines the relative importance of each variable. Variables that lead to a larger decrease in impurity when used for splitting nodes in the trees are considered more important, as they contribute more to the overall predictive power of the model. This feature importance metric identifies the key variables that have the most influence on the RF's performance and decision-making process.

It is crucial to recognize that variable rankings can differ among various ML models and datasets. As such, it is recommended to evaluate the importance of variables within the specific context of each model and dataset. This approach allows for a comprehensive understanding of the relevance of predictors in the given context.

In LR, variable rankings can be obtained by examining the coefficients or weights assigned to each predictor variable (Menard, 2011). These coefficients represent the

impact of each variable on the log-odds of the binary outcome being predicted. Variables with larger coefficients indicate a stronger influence on the predicted outcome. Positive coefficients suggest that an increase in the corresponding variable leads to a higher probability of the positive outcome, while negative coefficients indicate the opposite.

2.1.5 Evaluation of the models' performance

The performance of the models in study 1 was evaluated using several key metrics, including predictive accuracy, sensitivity, and positive predictive value. Predictive accuracy was assessed using the operating area under the curve (AUC) metric, which measures the models' ability to accurately distinguish between individuals with high chronic stress and those without. Higher AUC values indicate better overall predictive accuracy. Sensitivity, another important metric, quantifies the models' ability to correctly identify individuals who truly have high chronic stress. It represents the proportion of true positives that the model correctly identifies. Positive predictive value (PPV) was also utilized as a performance metric. PPV measures the probability that individuals identified as having high chronic stress by the model are genuinely experiencing it. It represents the proportion of true positives among all positive predictions made by the model (Hossin und Sulaiman, 2015).

By comparing the performance of the ML classifiers with LR, the study was able to determine which approach yielded the most favorable results in predicting chronic stress among PrAs, based on these evaluation metrics.

2.2 Study 2: Chronic stress in the German population

2.2.1 Dataset

This study 2 used nationally representative data from the DEGS1 study, which is part of the health monitoring program conducted by the Robert Koch Institute in Berlin, Germany. The DEGS1 study was carried out from 2008 to 2011 and involved interviews, examinations, and tests among a sample of the German population aged 18 to 79 years, with a total of 8,151 participants. The dataset used in study 2 included chronic stress measures in 5,801 respondents aged 18 to 64 years (Hapke et al., 2013).

The DEGS1 dataset includes a wide range of variables on sociodemographic characteristics, chronic diseases (e.g., coronary heart disease or stroke), general health, living conditions, preventive measures, and health-related behavior. For this analysis, 34 features were selected using the Powershap feature selection method.

Tab. 8 provides an overview of the demographic, clinical, and laboratory characteristics of the study participants. The study 2 aimed to examine the predictors associated with chronic stress using the DEGS1 dataset, which provided comprehensive information on various sociodemographic, health-related, and lifestyle factors. By analyzing these predictors, the study sought to identify the factors that contribute to chronic stress among the German population aged 18 to 64 years.

To address the challenges posed by small datasets in study 2, various data preprocessing techniques were applied. The dataset of DEGS1 consists of both continuous and discrete values. To ensure fair comparison and accurate modelling, the min-max normalization method was applied on the training dataset. This approach maintained the relationships within the data while avoiding bias.

The dataset had a relatively low missing value rate, with an overall rate of 13.91 %. To handle missing variables, the KNN imputation method was employed (Malarvizhi und Thanamani, 2012). KNN identifies the nearest neighbors based on Euclidean distance and replaces missing values using a majority vote on discrete variables and weighted averages on continuous features. Simultaneously, all features were imputed to ensure consistency.

The distribution of classes for chronic stress was imbalanced, with class 0 accounting for 52 %, class 1 for 38 %, and class 2 for 11 %. To address this imbalance, the SMOTE was applied (Chawla et al., 2002). This method generated new instances of the minority class to balance the class distribution without introducing additional information to the model.

For feature selection, the Powershap method, a wrapper-based Shapley feature selection approach, was employed (Verhaeghe et al., 2023). Powershap assesses the importance of features by evaluating their impact on predictions compared to random features. The method selects the most relevant features for modelling based on their importance.

These preprocessing steps were undertaken to ensure the data was appropriately prepared for subsequent analysis and modelling in the study. After preprocessing, the dataset, consisting of 34 features, was used to train ML classifiers for the classification

task. The dataset was split into training and validation sets using repeated K-fold cross-validation. This approach reduced bias in the model's estimated performance by averaging results across all folds. A value of K = 10 was chosen as the optimal number of folds, striking a balance between the time required to complete the tests and minimizing bias and variance associated with the validation process.

2.2.2 Primary outcome

In study 2, the TICS-SSCS scores obtained from the DEGS1 dataset were categorized into three classes based on the recommended DEGS1 approach. The three categories were defined as TICS-SSCS: 1–11 (≤median) = low stress, 12–22 = middle stress, and >22 = high stress (≥90th percentile).

2.2.3 eXtreme Gradient Boosting (XGBoost)

The XGBoost algorithm was employed as ML technique to predict chronic stress levels and identify factors that protect against chronic stress. XGBoost is an ensemble method based on DT that creates multiple weak learner classifiers, resulting in a scalable and high accurate model. The XGBoost model aims to establish a relationship between the input features X = {x_1, x_2, …, x_n} and the output variable Y. It uses K additive functions to make predictions, where each function corresponds to an independent tree structure with T leaves. The model minimizes a regularized objective function to learn the set of functions (Chen und Guestrin, 2016).

In the XGBoost model, K additive functions are applied to predict the output based on a given dataset with n samples and m features. The estimation (1) is used to calculate the predicted value $\hat{y}_j$ as the sum of the K functions $f_k(x_i)$:

$$\hat{y}_j = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Here, each f_k belongs to the regression tree's space $f(x) = \omega_q$, where q represents the independent structure of each tree with T leaves. Each f_k corresponds to a distinct tree

structure q and is associated with leaf weights $\omega \in \mathbb{R}^T$. To learn the set of functions, the model minimizes the following regularized objective (2):

$$L = \sum_i l(\hat{y}_j, y_j) + \sum_k \Omega(f_k) \quad (2)$$

In this objective, l represents the model loss function, and Ω denotes the regularized term, defined as $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2$. The regularization term controls the complexity of the model, with γ and λ being hyperparameters that determine the trade-off between model complexity and fitting the training data (Chen und Guestrin, 2016).

Hyperparameter tuning was performed using a grid-search approach with the scikit-learn class 'GridSearchCV'. The hyperparameters of XGBoost, such as the learning rate, number of estimators (trees), maximum depth, subsample, minimum child weight, L2 regularization term (lambda), and colsample-bytree, were optimized to achieve optimal model performance. The objective of the XGBoost model was set to 'multi:softmax' for multiclass classification (Tab. 1).

Tab. 1: Main hyperparameters for the XGBoost model

Hyperparameter	Value
Learning rate	0.3
Number of estimators	1,000
Max_depth	5
Subsample	0.8
Min_child_weight	3
L2 regularization term (Lambda)	2
Colsample-bytree	0.7
Objective	multi:softmax

The K-Fold Cross-Validation method was used in this study to train and evaluate the performance of the ML classifiers. After preprocessing the dataset and selecting the 34

relevant features, the data was divided into a ‘training’ and ‘validation’ set. To reduce bias and obtain a robust estimation of the model's performance, the repeated K-fold cross-validation approach was employed. With K = 10, the dataset was divided into 10 subsets or ‘folds’ for training and validation, and this process was repeated multiple times. By averaging the performance across all folds and repeats, we obtained a more reliable assessment of the model's performance while optimizing the time required for testing. This choice of K = 10 strikes a balance between minimizing bias and variance associated with the validation process and ensuring computational efficiency.

2.2.4 Model interpretation: SHapley Additive exPlanations (SHAP)

The study 2 emphasized the interpretability of the models to extract information that significantly impacts outcomes and identify factors protecting against chronic stress. The SHAP approach, proposed by Lundberg, is used to interpret predictions of complex models like XGBoost (Lundberg und Lee, 2017). SHAP calculates the impact of each feature on predictions and provides feature attribution, partial dependence plots, and other visualizations for improved understanding of the model. SHAP is a unified framework for interpreting the predictions of a wide range of models, including black box models. It provides an explanation for individual predictions by assigning importance values to each feature in the model.

Here is an overview of how SHAP works:

- Shapley values: SHAP is based on the concept of Shapley values from cooperative game theory. Shapley values measure the contribution of each feature in a prediction by evaluating its impact in different coalitions with other features. This provides a fair and consistent way to assign importance to each feature (Winter, 2002).
- Local explanations: SHAP provides local explanations, meaning it focuses on explaining individual predictions rather than providing a global view of the model. For a specific prediction, SHAP assigns an importance value to each feature, indicating how much each feature contributes to that particular prediction (Lundberg et al., 2020).
- Feature importance plot: SHAP generates a feature importance plot, commonly known as a SHAP summary plot. It displays the impact of each feature on predictions across

the entire dataset. The plot shows the SHAP values for each feature, indicating whether a feature contributes positively or negatively to the prediction.

- Interaction effects: SHAP can also capture interaction effects between features. It reveals how the presence or absence of certain features affects the importance of other features. This helps in understanding the complex relationships and dependencies between features.
- Model-agnostic: One of the strengths of SHAP is its model-agnostic nature. It can be applied to a wide range of ML models, including tree-based, deep learning, ensemble models, and more. This allows for consistent interpretation across different types of models (Molnar et al., 2022).
- Implementation: There are several libraries and framework to implement the SHAP methodology, such as the SHAP library in Python. These libraries provide functions and tools to calculate SHAP values, generate visualizations, and facilitate model interpretation.

The use of SHAP allows for a deeper understanding of the decision-making process of the model by discovering the contribution of different features toward individual predictions. It not only facilitates the identification of potential biases or interactions between features but also enhances the transparency and interpretability of black box models. This is particularly crucial in sensitive domains, as it fosters trust and facilitates the deployment of models.

The prediction of a specific input (X) is explained by calculating the impact of each feature on the prediction using Shapley values. The Shapley value $\hat{\phi}_j$ is estimated as the average difference between the predictions with and without a specific feature, where $\hat{g}\left(x_{-j}^{m}\right)$ is the prediction for x, but with a random number of feature values and divided by the number of iterations K (3) (Lundberg und Lee, 2017).

$$\hat{\phi}_j = \frac{1}{K}\sum_{k=1}^{K}((\hat{g}\left(x_{+j}^{m}\right) - \hat{g}\left(x_{-j}^{m}\right)) \tag{3}$$

TreeSHAP, a method suitable for gradient boosting models like XGBoost, is employed, offering visualizations of feature attributions and supporting partial dependence plots. Interaction values for TreeSHAP are estimated using equation (4), where φ_i represents the interaction value between features i and j. The equation involves feature subsets (S) and the delta function $(\delta_{ij}(S))$ to determine the interaction values. SHAP values contribute to the comprehension of tree models by providing insights into feature importance, local explanations, feature dependence plots, and summary plots (Lundberg und Lee, 2017).

$$\varphi_i = \sum_{S \subseteq N\{i,j\}} \frac{|S|!\,(M - |S| - 2i)}{2(M-1)!} \delta_{ij}(S) \tag{4}$$

For the implementation of study 2, Python 3.7 was employed along with various libraries from the Python data science ecosystem, including scikit-learn (version 1.0.2) and SHAP (version 0.40.0). The Powershap feature selection method was implemented using the Powershap Python library. The XGBoost classifier was trained and evaluated using scikit-learn, while the SHAP tool was used for model explainability.

2.2.5 Evaluation of the model's performance

The XGBoost model in study 2 was evaluated using several multiclass evaluation metrics. The primary metrics used were the AUC, precision, recall, and F1-score. Multiclass classification involves mutually exclusive classes, and the evaluation measures for individual classes were averaged using the macroaverage approach (Grandini et al.).

The receiver operating characteristic (ROC) curve was used to assess the classifier's performance. The macro true-positive rate and macro false-positive rate were plotted at different classification thresholds. The AUC value, ranging from 0 to 1, indicates the classifier's ability to distinguish between classes. A value of 1 represents a perfect classifier. In study 2, the ROC curve was plotted for each class using the One-vs-Rest approach, creating a series of binary problems. The macroaverage was computed by summing the values for true positive, true negative, false positive, and false negative across all classes. From these values, metrics such as precision (true positive instances), recall (true positive rate), specificity (true negative rate), and the F1-score (harmonic mean

of precision and recall) were calculated for each class. These metrics provide an overall assessment of the classifier's performance in terms of accuracy, sensitivity, and the balance between precision and recall. By using these evaluation metrics, the study aimed to determine the performance of the proposed method in accurately classifying instances into the respective stress categories. The AUC, precision, recall, and F1-score were used to assess the classifier's ability to distinguish between classes and capture the true positive instances.

3 Results

3.1 Study 1: Chronic stress in general practice assistants

3.1.1 Descriptive results

The dataset analyzed in study 1 included 550 PrAs from 136 general practices. The participants had an average age of 38 years, with a standard deviation of 12.6. Among the PrAs, 50.4 % (n = 277) were married (see Tab. 2).

Tab. 2: Sociodemographic characteristics and TICS score of practice assistants (n = 550)

	Participants (n = 550)	
Continuous variables, Mean [SD]		**Range**
Age	38 [12.61]	16–71
Persons in household > 18	2 [1.12]	0–6
Persons in household ≤ 18	1 [0.84]	0–6
Number of physicians in practice	3 [2.16]	1–10
Number of practice assistants in practice	8 [7.66]	0–35
Categorical variables, n (%)		
Female gender	544 (99.3)	
Marital status		
Married	277 (50.4)	
Single	221 (40.2)	
Divorced	45 (8.2)	
Widowed	7 (1.3)	
Number of persons in household	72 (13.1)	
Cares for next of kin	75 (13.6)	
Working hours/week		
1–9 hours	12 (2.2)	
10–19 hours	52 (9.5)	

	Participants (n = 550)
20–29 hours	116 (21.1)
30–39 hours	221 (40.2)
40–49 hours	116 (21.1)
50–59 hours	12 (2.2)
>60 hours	10 (1.8)
Working full time	364 (66.2)
Open-ended employment contract	466 (84.7)
Participated in stress seminar	31 (5.6)
Counseling for stress reduction	50 (9.1)
High level of chronic stress (TICS ≥ 23)	125 (22.7)

The TICS score showed that 22.7 % (n = 125) of the PrAs experienced high level of chronic stress, while 77.3 % (n = 425) experienced low level. Significant differences were identified in sociodemographic characteristics between the groups and level of chronic stress. The high chronic stress group consisted of PrAs, with an average age of 38.76, while the low stress group comprised younger PrAs, with an average age of 24.36. Furthermore, a higher percentage of unmarried PrAs (29.4 %) were found in the high chronic stress group compared to the low stress group, where only 17 % were not married.

Tab. 2 provides a comprehensive overview of sociodemographic characteristics of participants in study 1. Tab. 3 displays the practice and workplace characteristics of PrAs as well as the duties performed in the practice, including reception, telephone tasks, prescription handling, and blood pressure measurement. The three most frequent tasks were scheduling appointments (94.2 %), documenting in electronic health records (93.3 %), and preparing prescriptions (91.6 %).

Tab. 3: Practice and workplace characteristics of PrAs during the past three months (n = 550)

Practice characteristics	**n = 550**
Practice structure, n (%)	
Works in group practice	296 (53.8)
Works in single practice	147 (26.7)
Works in practice with several locations	50 (9.1)
Works in privately owned health center	6 (1.1)
Type of medical records in practice, n (%)	
Electronic medical records (EHR)	348 (63.3)
Paper and electronic records	187 (34.0)
Practice services for home care, n (%)	
Emergent home visits	515 (93.6)
Practice offers regular home visits	511 (92.9)
Nursing home visits*	508 (92.4)
Tasks of practice assistants during past 3 months, n (%)	
Scheduled appointments	518 (94.2)
Documented in patients´ EHR	513 (93.3)
Prepared prescriptions	504 (91.6)
Pulled up paper-based health records or opened electronic patient files	500 (90.9)
Performed phone service	499 (90.7)
Worked at reception	486 (88.4)
Obtained blood pressure readings	461 (83.8)
Performed ECGs	430 (78.2)
Prepared practice equipment for the day and switch them off in the evening	414 (75.3)
Performed laboratory work	393 (71.5)
Supported physician during patient-consultations	363 (66.0)
Supported billing of statutory health insurance patients	358 (65.1)
Performed disease-management examinations	332 (60.4)
Applied long-term blood pressure devices*	327 (59.5)

Practice characteristics	**n = 550**
Ordered medical practice supply	284 (51.6)
Applied long-term ECGs*	247 (44.9)
Ordered office supply	239 (43.5)
Performed treadmill testing	237 (43.1)
Supported billing of private patients*	236 (42.9)
Performed doppler examination of foot vessels/measured ankle-arm index*	103 (18.7)

**Missing values above 5 %; Electrocardiography = ECG*

The results of the workplace analysis using KFZA showed a versatility level of 3.6, a social support of 4.0, and a score on qualitative work demands of 2.9. For details, see Tab. 4.

Tab. 4: Result of short questionnaire for workplace analysis factor-level by PrAs (n = 550)

Work aspects	**Workplace characteristics**	**Mean (PrAs)**	**95 % CI**
Job content[1]	Versatility	3.6	3.6–3.7
	Completeness of task	3.5	3.4–3.6
Resources[1]	Scope of action	3.4	3.4–3.5
	Social support	4.0	4.0–4.1
	Cooperation	3.6	3.5–3.7
Stressors[2]	Qualitative work demands	2.2	2.1–2.3
	Quantitative work demands	2.9	2.8–3.0
	Work disruptions	2.7	2.7–2.8
	Workplace environment	2.2	2.1–2.3
Organizational culture[1]	Information and participation	3.6	3.6–3.7
	Benefits	2.9	2.8–2.9

[1]*High scores (>3) are considered positive; 2High scores (>3) are considered negative; Confidence interval = CI*

3.1.2 Performance of the machine learning algorithms

In terms of prediction accuracy, the ML classifiers were evaluated using the validation dataset. The AUC was calculated to assess the classifiers' performance. The results showed an AUC of 0.844 (95 % CI, 0.684–0.843) for the RF classifier, 0.760 (95 % CI, 0.605–0.777) for the ANN, 0.787 (95 % CI, 0.634–0.802) for the SVM, and 0.707 (95 % CI, 0.556–0.735) for the KNN classifier. Sensitivity and positive prediction value (PPV) for the ML classifiers were also calculated, with RF achieving 99 % sensitivity and 79 % PPV, ANN achieving 87 % sensitivity and 85 % PPV, SVM achieving 87 % sensitivity and 86 % PPV, and KNN achieving 99 % sensitivity and 78 % PPV.

LR analysis was also performed, and factors associated with chronic stress were identified through bivariate analysis. These factors included such as persons in the household below age 18 years, marital status, age, working hours per week, work status, and obtains blood pressure readings. The LR model achieved an AUC of 0.636 (95 % CI, 0.490–0.674) and predicted 316 cases correctly from a total of 425 cases, with a sensitivity of 75 % and a PPV of 44 %.

ML classifiers outperformed the LR model. The RF classifier showed the highest improvement (+20.8 %) compared to the LR model, resulting in a net increase of 104 cases correctly identified as high level of chronic stress. This classifier achieved a sensitivity of 99 % and a PPV of 79 % (see more in Tab. 5 and Tab. 6)

Tab. 5: Performance metrics of machine learning and logistic regression models

Algorithms	**AUC**	**95 % Confidence Interval**		**Absolute change in AUC (%)**
		LCL*	UCL*	
LR	0.636	0.490	0.674	[Reference]
ML: KNN	0.707	0.556	0.735	+7.1
ML: SVM	0.787	0.634	0.802	+15.1
ML: ANN	0.760	0.605	0.777	+12.4
ML: RF	0.844	0.684	0.843	+20.8

**LCL = lower control limit; UCL = Upper control limit*

Tab. 6: Full details of performance metrics for chronic stress prediction: Machine learning and logistic regression models

Algorithms	CSC (TP)*	CSC (FP)*	Total CSC	Non-CSC (TN*)	Non-CSC (FP)	Total non-CSC	Sensitivity (TP)	PPV*
LR	316	109	425	68	57	125	0.751	0.440
ML: RF	420	5	425	15	110	125	0.988	0.792
ML: KNN	421	4	425	6	119	125	0.991	0.780
ML: SVM	369	56	425	66	59	125	0.868	0.862
ML: ANN	369	56	425	59	66	125	0.868	0.848

**Chronic stress cases = CSC; True positive = TP; False positive = FP; True negative = TN; Positive predictive value = PPV*

3.1.3 Variable rankings in machine learning models

Comparing the results with the significant variables identified in the LR model, several factors showed consistency across all three models. These factors, including too much work, high demand to concentrate, time pressure, and complicated tasks, emerged as important variables in both the RF and ANN models, as well as being significant in the logistic regression model.

These findings highlight the crucial role of defined work characteristics in predicting chronic stress among PrAs. The consistent variable importance across different models underscores the significance of these factors and emphasizes their relevance in comprehending and addressing chronic stress in this population. Tab. 7 presents the top 10 influential factors identified by the three algorithms.

Tab. 7: Top 10 predictor variables associated with chronic stress listed by coefficient effect size (LR) weighting (ANN) and selection frequency (RF)

Standard model		**Machine learning models**			
LR	**Coefficient**	**ANN**	**Weight (%)**	**RF**	**Frequency**
Obtains blood pressure readings	0.951	Too much work	39.7	Too much work	0.73
Persons in household below age 18	0.349	High demand to concentrate	39.3	High demands to concentrate	0.71
Working hours/week more than 40	0.121	Time pressure	36.7	Time pressure	0.70
Work status	-0.109	Complicated tasks	31.5	Complicated tasks	0.67
Performs laboratory work	0.091	Insufficient practice room conditions	18.1	Age ≤ 35	0.63
Employment contract	0.063	Interruptions during work	14.9	Insufficient support by practice leaders	0.52
Age ≤ 35	0.045	Persons in household below age 18	13.8	Insufficient workplace environment	0.51
Insufficient workplace environment	0.028	Working hours/week more than 40 hours	12.7	Insufficient practice room conditions	0.50

Standard model		Machine learning models			
LR	**Coefficient**	**ANN**	**Weight (%)**	**RF**	**Frequency**
Measures ankle-arm index	0.018	Workplace environment	12.3	Holding together well	0.48
Marital status/single	0.006	Number of practitioners in the practice	10.6	Influence on work assigned	0.43

3.2 Study 2: Chronic stress in the German population

3.2.1 Descriptive results

The DEGS1 study involved 5,801 participants, with a mean age of 44 years. More than half of the population identified as female, comprising 53.1 % of the participants (n = 3,080). The average stress level across the entire population was 12.00 (95 % CI 11.79–12.20). Of the participants, 11 % (n = 625) reported having ‘high’ chronic stress (category 2), while 38 % (n = 2,188) had ‘middle’ chronic stress (category 1), and the majority, 52 % (n = 2,988), experienced ‘low’ chronic stress (category 0). Most of the participants rated their general state of health as either very good or good, accounting for 79.3 % (n = 4,599) of the population (Tab. 8).

Tab. 8: Demographic, clinical, and workplace characteristics of the German health interview and examination survey for adults study participants (n = 5,801)

Demographic characteristics	**Participants n = 5,801**	
Continuous variables, Mean [SD]		**Range**
Age	42 [13.11]	18–64
Number of persons in the household	3 [1.34]	1–11
Sleep hours per night in the past 4 weeks	7 [1.19]	2–12
Number of hospital nights in the past 12 months	1 [5.30]	0–150
Number of sick days in the past 12 months	13 [38.01]	0–365

Demographic characteristics	**Participants n = 5,801**
Categorical variables, n (%)	
Female gender	3,081 (49.6)
Marital status	
Married living with partner/separately from partner	3,697 (59.5)
Single	1,957 (31.5)
Divorced	376 (6.1)
Widowed	136 (2.2)
Provides care to someone in need or seriously ill	379 (6.1)
Renting or living in own apartment/house	
Rented apartment/house	2,689 (43.3)
Own apartment/house	3,268 (52.6)
Satisfaction with living space	
Very satisfied/satisfied	5,269 (84.8)
Neither satisfied nor dissatisfied	608 (9.8)
Dissatisfied/very dissatisfied	295 (4.8)
Residential area satisfaction	
Very satisfied/satisfied	5,091 (81.9)
Neither satisfied nor dissatisfied	727 (11.7)
Dissatisfied/very dissatisfied	320 (5.2)
General state of health	
Very good/good	4,942 (79.5)
Average	1,134 (18.3)
Poor/very poor	116 (1.8)
Intake of sleeping pills in the past 4 weeks	
Never	5,919 (95.3)
Less than 1 time	100 (1.6)
1 time or 2 times	73 (1.2)
3 times or more	86 (1.4)
Social support	
Low support	653 (10.5)

Demographic characteristics	**Participants n = 5,801**
Average support	3,082 (49.6)
Strong support	2,451 (39.5)
Health behavior consultation in the past 12 months	1,873 (30.4)
Has general practitioner	5,497 (88.5)
Visited to general practitioner in the past 12 months	4,870 (78.4)
Visited to neurologist in the past 12 months	463 (7.5)
Frequency of alcohol consumption	
Never	744 (12.0)
1 time per month or less	1,186 (19.1)
2–4 times per month	1,998 (32.2)
2–3 times per week	1,453 (23.4)
4 times per week or more	811 (13.1)
Tobacco use	
Yes, daily	1,701 (27.4)
Yes, occasionally	433 (7.0)
Not anymore	1,664 (26.8)
Never smoked	2,400 (38.7)
Comorbidities	
Has hypertension	1,625 (26.2)
Has diabetes	271 (4.4)
Has migraine	712 (11.5)
Has depression	682 (11.0)
Has anxiety disorder	327 (5.3)
Has burnout syndrome	292 (4.7)
Has one or more long-term chronic diseases	1,418 (22.8)
Prevention programs/sport activities	
Participated in prevention program in the past 12 months	988 (15.9)
Participated in relaxation or stress management program	188 (3.0)
Participated in gymnastics/fitness/balance sports program	832 (13.4)
Participated in alcohol cessation program	7 (0.1)

Demographic characteristics	**Participants n = 5,801**
Participated in smoking cessation program	17 (0.3)
Participated in weight reduction-healthy diet program	167 (2.7)
Sports activities per week (in the past 3 months)	
No sports activity	1,954 (31.5)
Up to 2 hours per week	2,584 (41.6)
Regularly, 2–4 hours per week	990 (15.9)
Regularly, more than 4 hours per week	645 (10.4)

3.2.2 Performance of the XGBoost algorithm

The evaluation metrics of the XGBoost model's performance, including AUC, precision, recall, specificity, and F1-score, are presented in Tab. 9. The model achieved the highest AUC score of 0.89 for class 2 (high chronic stress), indicating its good discriminatory ability for this class.

Tab. 9: Classification metrics: AUC, precision, recall, specificity, and F1-score for XGBoost model

Measure	**XGBoost**			
	Class 0	**Class 1**	**Class 2**	**Macroaverage**
AUC	0.83	0.71	0.89	0.81
Precision	0.73	0.56	0.58	0.63
Recall	0.80	0.55	0.37	0.52
Specificity	0.90	0.38	0.26	0.78
F1 score	0.76	0.60	0.45	0.54

Fig. 1 displays the ROC curves for the multiclass chronic stress prediction of the XGBoost model.

The macroaverage AUC score of 0.81 reflects the overall performance across all three stress classes. In terms of precision, which measures the model's ability to identify positive

instances correctly, the XGBoost model achieved values of 0.73, 0.56, and 0.58 for classes 0, 1, and 2, respectively. The macroaverage precision of 0.63 represents the average precision across all classes, suggesting a moderate level of accuracy in predicting chronic stress. The recall metric, which measures the model's ability to correctly identify all positive instances, yielded values of 0.80, 0.55, and 0.37 for classes 0, 1, and 2, respectively. The macroaverage recall of 0.52 indicates the model`s overall ability of the model to capture positive instances across all stress classes. Specificity, representing the model's ability to identify negative instances correctly, had values of 0.90, 0.38, and 0.26 for classes 0, 1, and 2, respectively. The macroaverage specificity of 0.78 suggests that the model performs relatively well in identifying negative instances on average.

The F1-score, which combines precision and recall into a single metric, provides an overall measure of the model's performance. The macroaverage F1-score of 0.54 represents the average F1-score across all stress classes. It indicates the harmonic mean of precision and recall and provides a balanced assessment of the model's effectiveness in capturing both positive and negative instances. The evaluation of the XGBoost model revealed its strong discriminatory ability in accurately classifying chronic stress, particularly for high stress instances (class 2). The macroaverage scores, which provide a comprehensive assessment across all stress classes, indicated that the model performed moderately well overall, with an F1-score of 0.54. This suggests that the model achieved a reasonable balance between precision and recall, effectively capturing both positive and negative instances of chronic stress.

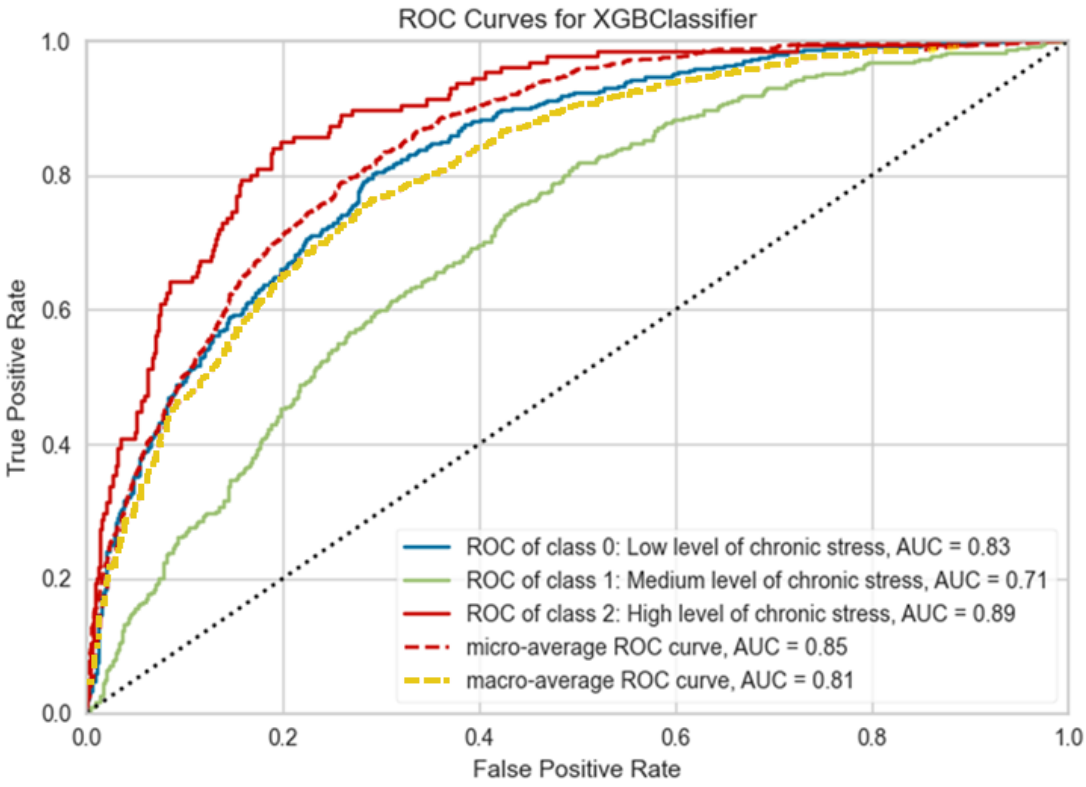


Fig. 1: ROC curves for 3 classes using the XGBoost multiclass classifier

3.2.3 Explanation of the behavior of individual features using SHAP

The SHAP analysis revealed that for class 0 (low level of chronic stress), features such as gender, general state of health, satisfaction with living space, and social support have a significant impact. These features play a crucial role in distinguishing low stress instances and contribute to the prediction of chronic stress at this level. Additionally, it is observed that class 2 (high level of chronic stress) uses features such as the number of sick days in the past 12 months, social support, sleeping hours per night in the past 4 weeks, gender, and general state of health. Both class 0 and class 2 show a reliance on various features, highlighting their importance in predicting chronic stress across different levels (Fig. 2).

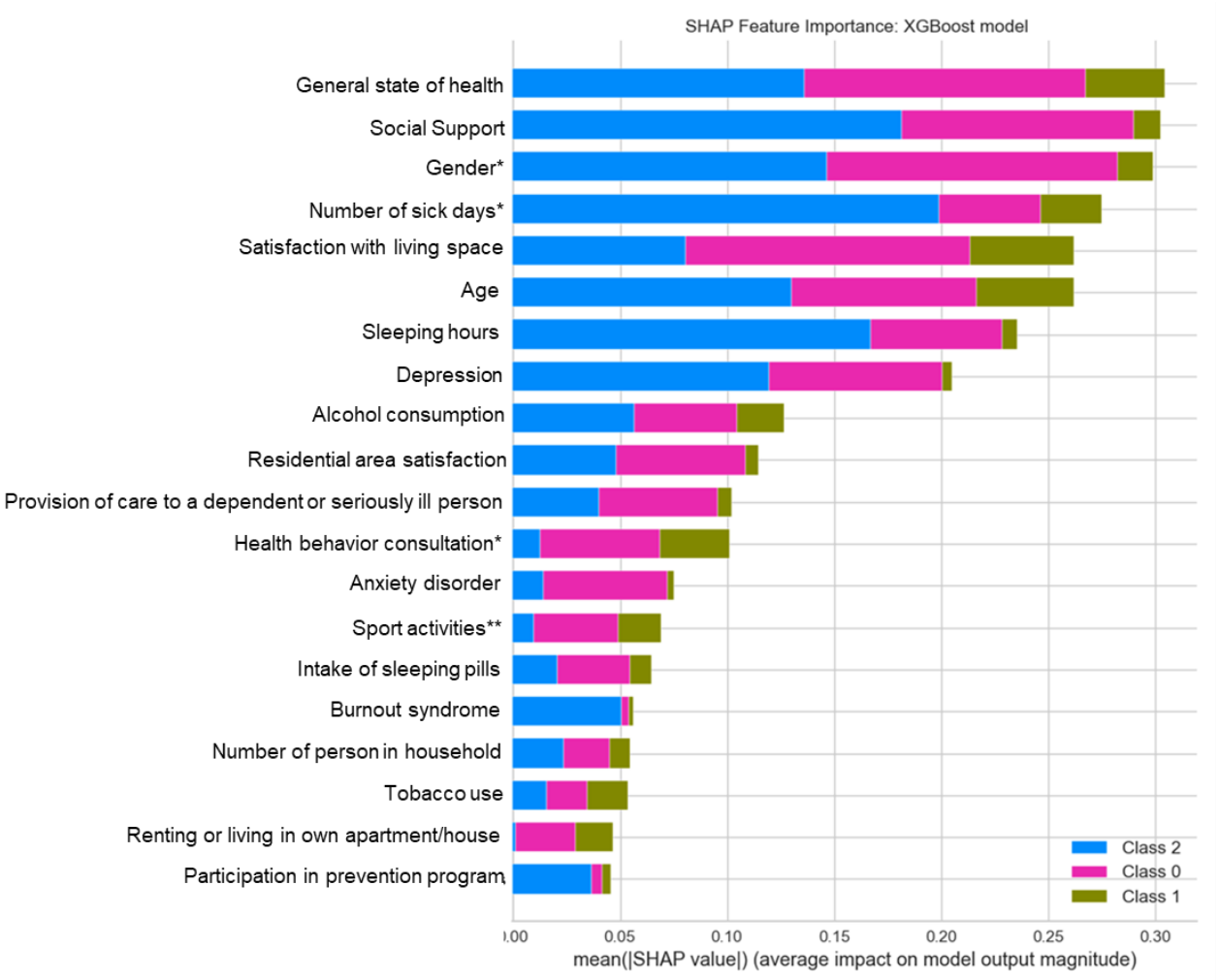
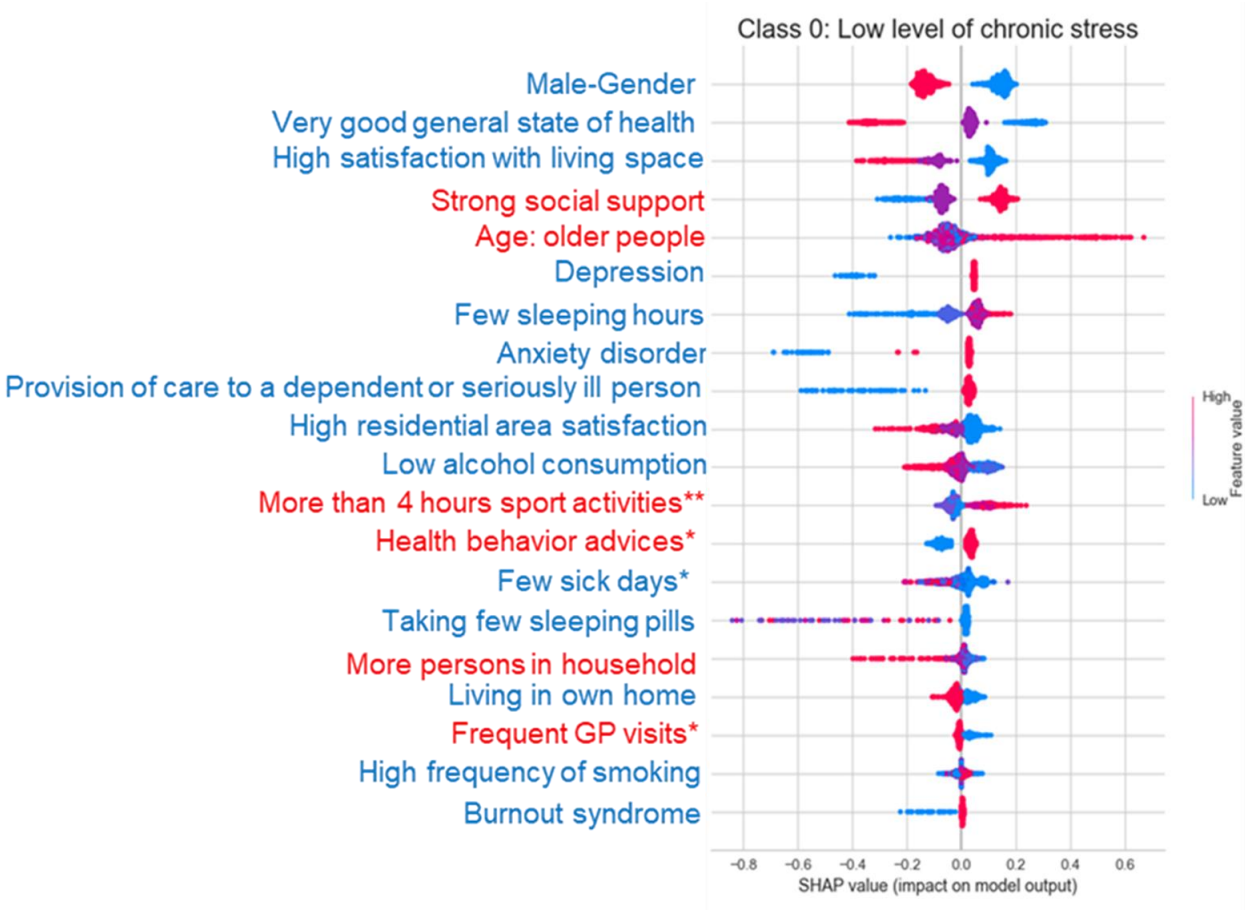


Fig. 2: SHAP feature plot of the 20 most important features: relative importance of each feature based on the average absolute value of the SHAP values; *in the past 12 months; **per week

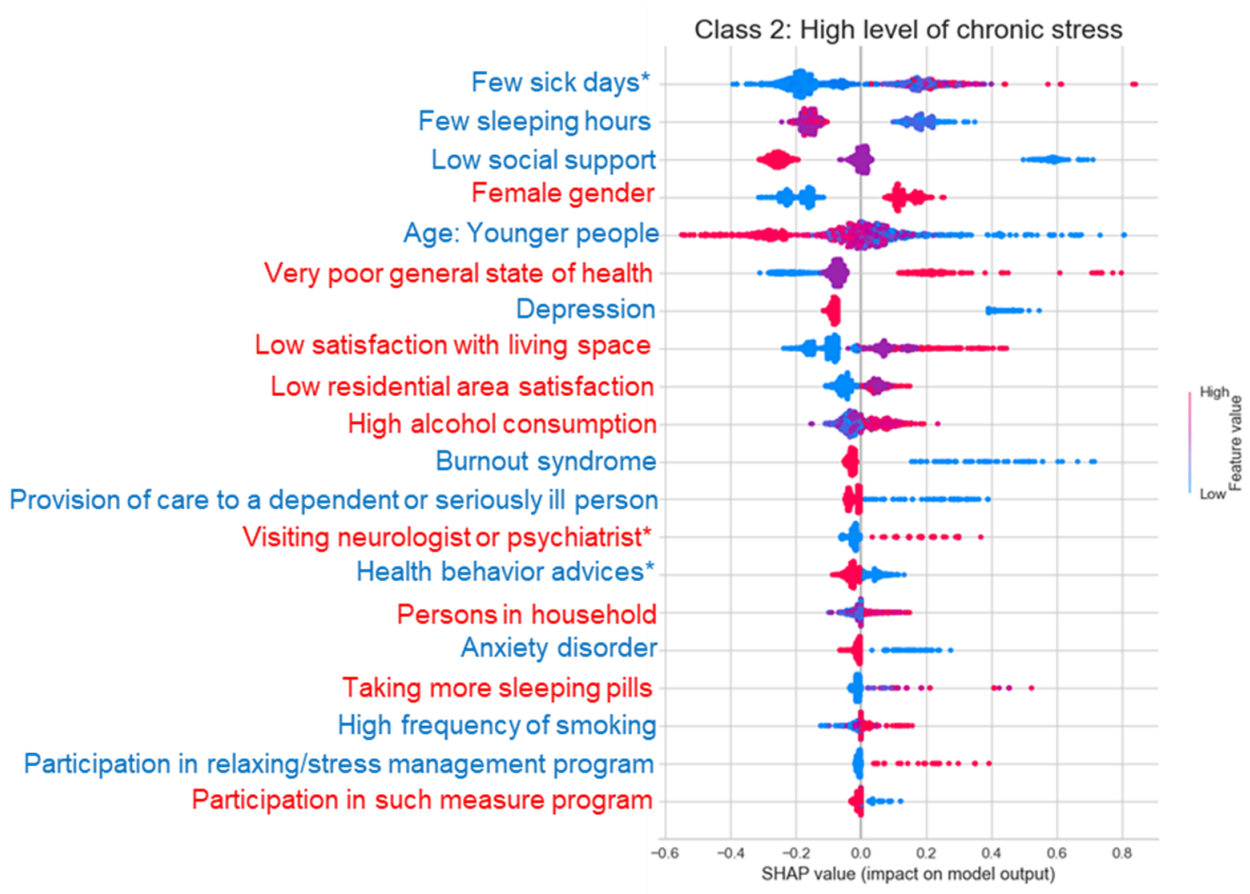
The five most important protective factors against chronic stress were identified as a very good general state of health, satisfaction with living space, strong social support, being male, and age ≥ 42 years (Fig. 3).



*Each dot is a Shapley value for a particular feature and reflects its impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low); *in the past 12 months; **per week*

Fig. 3: SHAP summary plot: Importance of the representative chronic stress features (top 20) in class 0

Based on the findings from the SHAP summary plot, the five most important risk factors for chronic stress were as follows: more sick days in the past 12 months, low social support, very poor general state of health, fewer sleeping hours in the past four weeks, and low satisfaction with living space (See Fig. 4).



*Each dot is a Shapley value for a particular feature and reflects the impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low); *in the past 12 months*

Fig. 4: SHAP summary plot: Importance of the representative chronic stress features (top 20) in class 2

4 Discussion

4.1 Main findings

Both conducted studies demonstrated the successful prediction of chronic stress levels in PrAs and the German population using ML algorithms, even with limited data. In study 1, the RF model demonstrated superior performance in predicting chronic stress compared to other ML algorithms (SVM, KNN, ANN) and the classical statistical method LR. With a small dataset containing 550 samples and more than 50 features, RF's ensemble learning approach effectively captured complex non-linear relationships and mitigated overfitting risks. Its decision tree-based structure coped well with high-dimensional data, providing a deeper understanding of feature importance while requiring simpler hyperparameter tuning. Additionally, RF's averaging mechanism across multiple trees contributed to robustness by minimizing the impact of outliers on performance. In study 2, XGBoost's boosted tree algorithm demonstrated high effectiveness in capturing complex non-linear relationships. Its regularization techniques were crucial in preventing overfitting in the high-dimensional dataset, ensuring more reliable predictions. Furthermore, XGBoost's robustness to outliers and scalability added to its suitability for the study. Unlike other studies on chronic stress, which used multivariate models with fewer parameters, my ML approaches allowed the incorporation of a wide range of characteristics, such as work characteristics, health status, lifestyle, living space, and social information.

The study 1 on factors influencing chronic stress in PrAs identified the five most important work characteristics using the RF model: excessive workload, high demand to concentrate, time pressure, complicated tasks, and insufficient support from practice leaders. The study 2 in the German population identified protective and risk factors for chronic stress in over 5,801 participants using SHAP. Here, the five most important protective factors against chronic stress were a very good general state of health, satisfaction with living space, strong social support, being male, and individuals of age ≥ 42 years. In contrast, the most significant risk factors were: more sick days in the past 12 months, low social support, a very poor general state of health, fewer sleeping hours in the past four weeks, and low satisfaction with living space. The differences between the risk-protective factors of the first and second study stem from differing input variables: while study 1 focused on work-related issues, study 2 focused on health and living

parameters in general. However, both studies implied that chronic stress can only be targeted effectively by complex interventions. The potential interventions and strategies encompass a wide range of individual and social factors, such as work characteristics, social information, health status, lifestyle, and living space.

4.2 Methodological considerations on analyzing chronic stress data with ML

In the abundance of studies that have addressed chronic stress in different populations using various stress measurement approaches, my conducted studies were compared with other ML approaches for chronic stress analysis. These studies measured chronic stress using four different approaches: biological signals (e.g., EEG signals, heart rate, electrodermal activity), facial expressions, social media text analyses and questionnaires (e.g., TICS and PSS scales).

The following studies are examples to predict chronic stress using ML methods based on biological signals:

- Gupta et al. investigated the detection of mental stress using EEG signals involving 14 human subjects with an average age of 26 years. with SVM utilized as the ML method.(Gupta et al., 2020).
- Omneya Attallah in her study aimed to develop an early detection system for mental stress using EEG signals from 36 participating worker. She employed various classification algorithms, including KNN, linear and cubic support vector machine (SVM), and RF classifiers to distinguish between stress and non-stress states (Attallah, 2020).
- The research conducted by Sriramprakash et al. analyzed the stress levels using various sensors. They used the SWELL-KW dataset, which contains data from 25 participants who worked under three conditions: neutral, interruptions, and time pressure, with an additional relaxation phase. The researchers employed SVM and KNN algorithms to classify the stress levels of the individuals as either normal or stressed (Sriramprakash et al., 2017).

The following studies are such examples that used social media interaction analyses and questionnaire surveys to detect mental stress using ML algorithms.

- Ahuja and Banga analyzed stress levels in a dataset of 206 students from Jaypee Institute of Information Technology. They used the PSS scale to measure stress levels, considering parameters like exam pressure and recruitment stress. Classification of stress levels was performed using ML classifiers: LR, Naïve Bayes, RF, and SVM (Ahuja und Banga, 2019).
- Chaware et al. proposed a model to estimate user stress levels using social media data including the extraction of Facebook posts and analysis of those posts. The transductive support vector machine (TSVM), SVM, Naïve Bayes (NB), RF, DT, and Adaboosted D-Tree algorithms were employed to categorize users' posts and estimate their stress levels based on positive and negative sentiments (Chaware et al., 2020).
- Yogesh Pingle aimed to predict chronic stress levels by using a dataset collected through real-time surveys conducted among 2,200 students from various streams in Mumbai colleges. The target attribute classified stress levels into ‘low’, ‘medium’, and ‘high’ categories. To achieve this, it was employed a combination of ML algorithms, including convolutional neural network (CNN), KNN, RF, and CNN-Adaboost (Pingle, 2020).
- Kene and Thakare used a dataset obtained through an online stress scale questionnaire, which involved 270 users. The questionnaire comprised 25 questions aimed at analyzing stress. The target attribute class label ranged from 0 to 15, where stress levels of 0-10 were classified as ‘no-stress’ and levels of 11-15 were classified as ‘stress’. To predict the stress level is used two ML algorithms, namely SVM and RF. (Kene und Thakare, 2022).
- The study conducted by Reddy et al. analyzed stress disorder patterns among working IT professionals. They used data from the open sourcing mental illness (OSMI) mental health survey, which included 750 participants. ML classifiers such as LR, KNN, DT, RF, boosting, and bagging were applied. The study identified stress factors, including gender, family illness history, and workplace mental health benefits (Reddy et al., 2018).

Tab. 10 provides an overview of the results from the aforementioned studies, which are difficult to compare regarding their outcomes due to the diverse methods of measuring stress and the various ML methods used. From the studies used ML to detect chronic

stress based on questionnaire surveys, only the study conducted by Reddy et al. identified the feature importance using a decision tree. These factors were gender, family history of illness, and the availability of mental health benefits in the workplace (Reddy et al., 2018). Additionally, this study employed six ML methods, with the best performer being boosting, an ensemble algorithm, achieving an AUC of 75 %.

This result is in line with my two studies, which both found the ensemble methods as performing best: in the study 1, RF was as the best model with an 84 % AUC, and in the study 2 XGBoost, another ensemble method, showed an AUC of 81 %. These findings indicate that ensemble methods, such as boosting, RF, and XGBoost, have the ability to handle small datasets and predict complex non-linear relationships in the context of chronic stress better than other ML methods.

Tab. 10: Overview of research studies on stress detection using machine learning

Authors	**Data**	**ML**	**Outcome**	**Model Accuracy (%)**
Gupta et al.	EEG-signals	SVM	Low, Medium, High, No stress level	SVM: 96.36
Omneya Attallah	EEG-signals	KNN, Linear-cubic SVM, RF	Stress, No stress	KNN: 99.98
Sriramprakash et al.	EEG- signals	KNN, SVM	Stressed, Normal	SVM: 93
Ahuja et al.	Questionnaire-based (PSS)	RF, SVM, NB, KNN	Highly stressed, Stressed, Normal	SVM: 86
Chaware et al.	Posts from social media	TSVM, SVM, DT	Positive, Negative	TSVM: 84

Authors	Data	ML	Outcome	Model Accuracy (%)
Yogesh Pingle	Questionnaire-based	CNN, KNN, RF, CNN-ADABOOST	Low, Medium, High	Stacking algorithms (RF, KNN, CNN-A): 88
Aksha and Thakare	Questionnaire-based	SVM, RF	No stress, Stress	SVM: 80.2
Authors	**Data**	**ML**	**Outcome**	**AUC (%)**
Reddy et al.	Questionnaire-based	LR, KNN, DT, RF, boosting, bagging	Stress Treatment, No-Treatment	Boosting: 75
Bozorgmehr et al.	Questionnaire-based	RF, SVM, KNN, ANN	No stress, Stress	RF: 84
Bozorgmehr and Weltermann	Questionnaire-based	XGBoost	Low, Medium, High	XGBoost: 81

4.3 Small datasets and feature importance

To address the challenges of higher variability and potential overfitting risks in small datasets, ensemble methods like RF and XGBoost were employed in both of my studies. These methods proved instrumental in enhanced the model´s performance and support the handling of the constraints posed by small datasets. By combining multiple weak learners, these techniques improve predictive accuracy, reduce the risk of overfitting, and are robust against outliers in order to make stress prediction models more reliable, especially on limited samples (Zhou, 2012; Sagi und Rokach, 2018). Additionally, using cross-validation techniques, specifically k-fold cross-validation, helped to overcome validation and evaluation issues related to small datasets. Furthermore, class imbalance, a common issue in small datasets, was addressed using oversampling techniques,

increasing the frequency of near-miss data points to improve reliability of stress prediction results.

ML approaches are widely used across various domains, but their interpretability is less considered. In medicine, integrating and understanding diverse data is a complex challenge. Explainable-AI is crucial in medicine as it helps medical professionals grasp machine decisions and builds trust in future AI systems (Holzinger et al.). My two conducted studies focused on identifying risk and protective factors for chronic stress using feature importance methods, especially SHAP, going beyond predicting stress levels. This enables a comparison with other research, particularly classical statistical methods that focus on measuring chronic stress and its contributing factors.

The classical methods in measuring chronic stress generally focused on identifying risk factors and less considering protective factors:

- The study by Hapke et al. examined chronic stress among adults in Germany using DEGS1 dataset and TICS. High levels of stress were more common among women and individuals with a lower socioeconomic status. It also highlighted the significant impact of chronic stress on mental wellbeing, including depressive symptoms, burnout syndrome, and sleep disturbances, but not differences between age groups (Hapke et al., 2013).
- The prospective cohort study by Herrera et al. assessed changes in chronic stress among young adults (n = 1,688) transitioning from high school to university or working life. The researchers measured two dimensions of stress at university or work: work overload and work discontent using TICS. The work overload increased during this transition, however, any significant difference in work discontent between the groups (Herrera et al., 2017).
- In another study by Herrera et al., the association between greenness around homes and occupational stress was examined. The findings of the study showed that residential green spaces, as measured by the vegetation index, were associated with two types of job-related chronic stress in young German adults who were transitioning from school to university or working life (Herrera et al., 2018).
- Kersting et al. focused on chronic stress among German GPs. Work-related factors, such as challenges related to personnel matters, practice software, complexity of

patients, and keeping medical records up-to-date, were found to be significantly associated with high chronic stress in GPs (Kersting et al., 2019).

- The study by Viehmann et al. investigated chronic stress prevalence among GPs and PrAs (n = 764) from 136 German general practices. The results showed that chronic stress was highest among female GPs and PrAs. On the practice level the PrAs reported high levels of stress. The study identified an association between high chronic stress and the number of working hours per week for both GPs and PAs. In addition, the GPs, who consistently employed more than five stress management measures had significantly lower levels of stress. (Viehmann et al., 2017).
- The German study by Petrowski et al. included 2,339 healthy participants aged 14 to 99 quantified chronic stress using TICS. Women and younger individuals, especially those aged 35 to 44, reported higher chronic stress levels. The TICS demonstrated good reliability and two higher-order factors: high demands and lack of satisfaction (Petrowski et al., 2012).
- The large cross-sectional study was conducted by Stubbs et al. and included 34,129 participants from six countries: China, Ghana, India, Mexico, Russia, and South Africa. The study found positive associations between perceived stress and various health conditions, namely multimorbidity, stroke, depression, and hearing problems (Stubbs et al., 2018).
- Ng and Jeffery conducted a study in the US, examining the association between perceived stress and health behaviors in 12,110 individuals from 26 worksites. They found that high stress was linked to unhealthy habits, including higher fat intake, less exercise, cigarette smoking, and reduced self-efficacy in quitting smoking. However, stress did not show a correlation with alcohol intake. The study underscores how stress may influence disease outcomes through its impact on unhealthy behaviors (Ng und Jeffery, 2003).
- Van der Horst et al. conducted a study involving 24,347 Canadian participants to investigate the impact of friendship network characteristics on subjective well-being (SWB). They found that higher frequency of contact, more friends, and less heterogeneity in the friendship network were linked to increased social trust, reduced stress, and better health (van der Horst und Coffé, 2012).

Tab. 11 provides an overview of these studies, the number of variables evaluated, and the factors, which were identified as risk or protective factors.

In comparison to my studies, almost all studies using classical statistical approaches applied bivariate-multivariate models with much fewer parameters: the number of variables evaluated ranged from 2 to 20, while my ML approaches evaluated 34 to 64 variables, encompassing a wide range of characteristics, such as work characteristics, health status, lifestyle, living space, and social information. This comparison shows the importance of using ML in complex scenarios such as workplaces. Several risk factors for chronic stress consistently emerged in the studies, including being female, having lower social support, experiencing work overload, and being younger. On the other hand, protective factors against chronic stress were relatively scarce, e.g., stress management measures and life satisfaction. My second study using SHAP revealed that a good general state of the health, satisfaction with living space and residential area, strong social support, being male, and age ≥ 42 years, more sleeping, and more than 4 hours sport activities per week were protective against chronic stress.

Tab. 11: Comparison of the number of variables evaluated in the various studies and risk-protective factors for chronic stress

Authors	**Sample size**	**Analytic Method**	**Variables**	**Risk/Protective Factors for Chronic Stress**
Hapke et al.	8,152 GP	MVA*	6	RF*: Female, lower SES, low social support
Herrera et al.	1,688 GrS*	GEE*	11	RF*: Work overload
Herrera et al.	1,632 GrS*	GEE*	13	PF*: Residential green spaces
Kersting et al.	109 GPs	BVA*	10	RF*: Work-related factors: challenges related to personnel matters, practice software, complexity of patients, and keeping medical records up-to-date

Authors	Sample size	Analytic Method	Variables	Risk/Protective Factors for Chronic Stress
Viehmann et al.	764 GPs-PrAs	MVA*	12	RF*: Female, number of working hours per week for both GPs and PrAs PF*: Applying more than five measures regularly to compensate for stress
Petrowski et al.	2,339 GrS*	LR*	2	RF*: younger women; aged 35 to 44
Stubbs et al.	34,129 Adults from 5 countries	LR*	7	RF*: Various health conditions, namely multimorbidity, stroke, depression, and hearing problems
Ng and Jeffery	12,110 US-workers	LR*	12	RF*: Unhealthy habits: higher fat intake, less exercise, cigarette smoking, and reduced self-efficacy in quitting smoking
Der Horst and Coffé	24,347 CaP*	DA*	20	PF*: Higher frequency of contact, more friends, and less heterogeneity in the friendship network
Bozorgmehr et al.	550 PrAs	ML: Random Forest	64	FI*: Too much work, high demands to concentrate, time pressure, Complicated tasks, Age <= 35, insufficient support by practice leaders
Bozorgmeh and Weltermann	5,801 GrP*	ML: XGboost	34	RF*: More sick days, few sleeping hours, low social support PF*: Male gender, very good general state of the health,

Authors	Sample size	Analytic Method	Variables	Risk/Protective Factors for Chronic Stress
				satisfaction with living space and residential area, strong social support, being male, age ≥ 42 years, more sleeping, more than 4 hours sport activities per week

Feature importance = FI; Risk factor = RF; Protective factor = PF; Multivariate analysis = MVA; Generalized estimating equations = GEE; Bivariate analysis = BVA; Linear regression = LR; Descriptive analysis = DA; German population = GrP; Canadian population = CaP; German samples = GrS

4.4 Strengths and limitations

The study has notable strengths, particularly in employing advanced ML algorithms which facilitated the identification of crucial factors influencing the model. The use of SHAP in developing an interpretable XGBoost model further enhanced the understanding of risk and protective factors for chronic stress. Additionally, the population-based DEGS1 dataset minimized selection bias risks. However, certain limitations should be acknowledged. The focus on only female PrAs limits the generalizability of findings to other populations or professions. The reliance on self-reported measures introduces potential response bias and may not fully capture the complexity of chronic stress. The cross-sectional design hinders causal information, underscoring the need for longitudinal studies to explore temporal dynamics of risk and protective factors over time. Furthermore, the DEGS1 data collected from 2008 to 2011 might not fully reflect the current living conditions in Germany, including potential effects of the pandemic, which were not assessed in this study. As a result, the transferability of the results to other settings should be approached with caution.

4.5 Conclusions and perspectives

In conclusion, my dissertation project comprised two studies to predict chronic stress and identify important variables influencing the models. Study 1 compared four ML classifiers to LR for predicting chronic stress in PrAs. The results showed that ML classifiers, specifically the RF, outperformed the LR model. The RF model identified important

predictors that influence chronic stress to provide Information for potential interventions. Additionally, an interpretable XGBoost model was developed to predict chronic stress in adults in Germany, which identified the risk-protective factors for chronic stress using the SHAP methodology. These findings emphasize the importance of addressing chronic stress among PrAs and the broader German population and highlight the potential of ML for data analyses on this topic. Future research should explore the experiences of healthcare professionals from other specialties to gain a comprehensive understanding of chronic stress in the healthcare workforce. ML methods have the capability to identify a diverse range of risk and protective factors, including work characteristics, health status, lifestyle, living space, and social circumstances. Based on my findings of risk and protective factors, targeted interventions and support systems like stress management sessions, career guidance, health awareness programs, and counselling assistance can be developed.

5 Abstract

Background: Chronic stress is widespread and adversely affects mental and physical health. The two studies in this dissertation used machine learning to predict chronic stress and identify its risk as well as protective factors. The first study examined workplace factors in German general practice assistants, while the second developed an interpretable model using a national dataset from the Robert Koch Institute to identify protective factors for improved well-being.

Methods: The first study analyzed 550 general practice assistants comparing 4 machine learning classifiers (random forest, support vector machine, K-nearest neighbors, and artificial neural network) and logistic regression to predict chronic stress. The model performance was evaluated using metrics such as the area under the curve (AUC), sensitivity, and positive predictive value. The second study investigated chronic stress in the German population using data from 5,801 in the representative DEGS1 study. Multiclass classification with extreme gradient boosting (XGBoost) and Shapley additive explanations (SHAP) was employed to predict stress levels and especially identify protective factors. The evaluation metrics included the area under the curve (AUC), precision, recall, and F1-score, which were averaged using the macro average.

Results: In the first study, machine learning models outperformed logistic regression in predicting chronic stress among practice assistants. The random forest model achieved the highest AUC of 84 %. Work characteristics were identified as relevant factors. In the second study, the XGBoost model obtained an AUC of 81 %, a precision of 63 %, a recall of 52 %, a specificity of 78 %, and a F1-score of 54 %. The study identified several protective factors against chronic stress like satisfaction with living space, strong social support, being male, and age ≥ 42 years.

Conclusion: Machine learning classifiers, specifically ensemble methods, demonstrated superior performance compared to logistic regression and other ML approaches in predicting chronic stress even with small data sets. Aiming at reliable feature importance detection, the SHAP technique identified significant protective factors that should be taken into account when designing interventions to alleviate chronic stress.

6 List of figures

7 List of tables

8 References

Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare. 2022. 10: 1–30

Ahuja R, Banga A. Mental Stress Detection in University Students using Machine Learning Algorithms. Procedia Computer Science. 2019. 152: 349–353

Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010. 26: 1340–1347

Apostolopoulos ID, Aznaouridis SI, Tzani MA. Extracting Possibly Representative COVID-19 Biomarkers from X-ray Images with Deep Learning Approach and Image Data Related to Pulmonary Diseases. J. Med. Biol. Eng. 2020. 40: 462–469

Attallah O. An Effective Mental Stress State Detection and Evaluation System Using Minimum Number of Frontal Brain Electrodes. Diagnostics (Basel, Switzerland). 2020. 10: 1–26

Barnard AS, Motevalli B, Parker AJ, Fischer JM, Feigl CA, Opletal G. Nanoinformatics, and the big challenges for the science of small things. Nanoscale. 2019. 11: 19190–19201

Berrar D. Cross-Validation. Elsevier. 2019. 1: 542–545

Böken B. On the appropriateness of Platt scaling in classifier calibration. Information Systems Elsevier. 2021. 95: 1–16

Bonaccorso G. Machine Learning Algorithms. Birmingham: Packt Publishing Limited. 2017

Brown GW, Tirril H. Social origins of depression. London: Tavistock Publication Limited. 1978

Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018. 300: 70–79

Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. CODATA. 2015. 14: 1–10

Chaware SM, Makashir C, Athavale C, Athavale M, Baraskar T. Stress Detection Methodology based on Social Media Network: A Proposed Design. IJITEE. 2020. 9: 3489–3492

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair. 2002. 16: 321–357

Chen T, Guestrin C. Xgboost: A scalable tree boosting system. ACM. 2016: 785–794

Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. JAMA. 2007. 298: 1685–1687

Cohen S, Kamarck T, Mermelstein R. A Global Measure of Perceived Stress. Journal of Health and Social Behavior. 1983. 24: 385

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995. 20: 273–297

Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge Univ. Press. 2012

Dai S, Mo Y, Wang Y, Xiang B, Liao Q, Zhou M, Li X, Li Y, Xiong W, Li G, Guo C, Zeng Z. Chronic Stress Promotes Cancer Development. Frontiers in oncology. 2020. 10: 1–10

Datta D, Arnsten AFT. Loss of Prefrontal Cortical Higher Cognition with Uncontrollable Stress: Molecular Mechanisms, Changes with Age, and Relevance to Treatment. Brain sciences. 2019. 9: 1–16

DeMaris A. A Tutorial in Logistic Regression. Journal of Marriage and the Family. 1995: 1–6

Dreher A, Theune M, Kersting C, Geiser F, Weltermann B. Prevalence of burnout among German general practitioners: Comparison of physicians working in solo and group practices. PLOS ONE. 2019. 14: 1-13

Dulhare U, Ahmad K, Bin Ahmad K. Machine Learning and Big Data. Boston, MA: Wiley-Scrivener; Safari. 2020

Eizirik DL, Cardozo AK, Cnop M. The role for endoplasmic reticulum stress in diabetes mellitus. Endocr Rev. 2008. 29: 42–61

Elrahman SMA, Abraham A. A review of class imbalance problem. Journal of Network and Innovative Computing. 2013. 1: 332–340

Elreedy D, Atiya AF. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences. 2019. 505: 32–64

Everson-Rose SA, Roetker NS, Lutsey PL, Kershaw KN, Longstreth WT, Sacco RL, Diez Roux AV, Alonso A. Chronic stress, depressive symptoms, anger, hostility, and risk of stroke and transient ischemic attack in the multi-ethnic study of atherosclerosis. Stroke. 2014. 45: 2318–2323

Ghahramani Z. Unsupervised Learning: Springer, Berlin, Heidelberg, 2004: 72–112

Goodfellow I, Courville A, Bengio Y. Deep learning. Cambridge, Massachusetts: The MIT Press. 2016

Gößwald A, Lange M, Dölle R, Hölling H. Die erste Welle der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1): Gewinnung von Studienteilnehmenden, Durchführung der Feldarbeit und Qualitätsmanagement. Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz. 2013. 56: 611–619

Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview: 1–17

Gupta R, Alam MA, Agarwal P. Modified Support Vector Machine for Detecting Stress Level Using EEG Signals. Computational intelligence and neuroscience. 2020. 2020: 1–14

Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003. 3: 1157–1182

Hapke U, Maske UE, Scheidt-Nave C, Bode L, Schlack R, Busch MA. Chronischer Stress bei Erwachsenen in Deutschland: Ergebnisse der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1). Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz. 2013. 56: 749–754

Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell. Syst. Their Appl. 1998. 13: 18–28

Henein MY, Vancheri S, Longo G, Vancheri F. The Impact of Mental Stress on Cardiovascular Health-Part II. Journal of Clinical Medicine. 2022. 11: 1–17

Herrera R, Berger U, Genuneit J, Gerlich J, Nowak D, Schlotz W, Vogelberg C, Mutius E von, Weinmayr G, Windstetter D, Weigl M, Radon K. Chronic Stress in Young German Adults: Who Is Affected? A Prospective Cohort Study. International journal of environmental research and public health. 2017. 14: 1–13

Herrera R, Markevych I, Berger U, Genuneit J, Gerlich J, Nowak D, Schlotz W, Vogelberg C, Mutius E von, Weinmayr G, Windstetter D, Weigl M, Heinrich J, Radon K. Greenness and job-related chronic stress in young adults: a prospective cohort study in Germany. BMJ open. 2018. 8: 1-12

Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?: 1–28

Hossin M, Sulaiman MN. A Review on Evaluation Metrics for Data Classification Evaluations. IJDKP. 2015. 5: 1–11

Hu Y, Visser M, Kaiser S. Perceived Stress and Sleep Quality in Midlife and Later: Controlling for Genetic and Environmental Influences. Behavioral sleep medicine. 2020. 18: 537–549

Iqbal MJ, Javed Z, Sadia H, Qureshi IA, Irshad A, Ahmed R, Malik K, Raza S, Abbas A, Pezzani R, Sharifi-Rad J. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int. 2021. 21: 1–11

Jain A, Patel H, Nagalapatti L, Gupta N, Mehta S, Guttula S, Mujumdar S, Afzal S, Sharma Mittal R, Munigala V. Overview and Importance of Data Quality for Machine Learning Tasks. 2020: 3561–3562

Karasek R, Brisson C, Kawakami N, Houtman I, Bongers P, Amick B. The Job Content Questionnaire (JCQ): an instrument for internationally comparative assessments of psychosocial job characteristics. Journal of Occupational Health Psychology. 1998. 3: 322–355

Katsis Y, Balac N, Chapman D, Kapoor M, Block J, Griswold WG, Huang J, Koulouris N, Menarini M, Nandigam V, Ngo M, Ong KW, Papakonstantinou Y, Smith B, Zarifis K, Woolf S, Patrick K. Big Data Techniques for Public Health: A Case Study. IEEE/ACM International Conference on Connected Health. 2017. 2017: 222–231

Kene A, Thakare S. Prediction of Mental Stress Level Based on Machine Learning Machine Intelligence and Smart Systems: Springer, Singapore, 2022: 525–536

Kersting C, Zimmer L, Thielmann A, Weltermann B. Chronic stress, work-related daily challenges and medicolegal investigations: a cross-sectional study among German general practitioners. BMC family practice. 2019. 20: 1–8

Kivimäki M, Steptoe A. Effects of stress on the development and progression of cardiovascular disease. Nature reviews. Cardiology. 2018. 15: 215–229

Klein EM, Brähler E, Dreier M, Reinecke L, Müller KW, Schmutzer G, Wölfling K, Beutel ME. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC psychiatry. 2016. 16: 1–10

Koprinkova-Hristova P, Mladenov V, Kasabov NK. Artificial Neural Networks. s.l.: Springer-Verlag. 2014

Kramer O. K-Nearest Neighbors Dimensionality Reduction with Unsupervised Nearest Neighbors: Springer, Berlin, Heidelberg, 2013: 13–23

Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016. 5: 221–232

Kumari S, Bhatia M. Machine Learning Techniques For Public Health System: A Scientometric Review. Second International Conference on Computer Science, Engineering and Applications (ICCSEA). IEEE. 2022. 2022: 1–6

Liu J, Li J, Li W, Wu J. Rethinking big data: A review on the data quality and usage issues. ISPRS Journal of Photogrammetry and Remote Sensing. 2016. 115: 134–142

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From Local Explanations to Global Understanding with Explainable AI for Trees. Nature machine intelligence. 2020. 2: 56–67

Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017. 2017: 1–10

Malarvizhi R, Thanamani AS. K-nearest neighbor in missing data imputation. International Journal of Engineering Research and Development. 2012. 5: 5–7

Marin M-F, Lord C, Andrews J, Juster R-P, Sindi S, Arsenault-Lapierre G, Fiocco AJ, Lupien SJ. Chronic stress, cognitive functioning and mental health. Neurobiology of Learning and Memory. 2011. 96: 583–595

Marr B. Big Data. Chichester, United Kingdom: Wiley; Ciando. 2016

McEwen BS. Protective and damaging effects of stress mediators: central role of the brain. Dialogues in clinical neuroscience. 2022. 8: 367–381

Menard S. Standards for Standardized Logistic Regression Coefficients. Social Forces. 2011. 89: 1409–1428

Molnar C. Interpretable machine learning. Morisville, North Carolina: Lean Publishing - Creative Comons. 2020

Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, Casalicchio G, Grosse-Wentrup M, Bischl B. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models: Springer, Cham, 2022: 39–68

Nasteski V. An overview of the supervised machine learning methods. 2017. 4: 51–62

Ng D, Jeffery RW. Relationships between perceived stress and health behaviors in a sample of working adults. Health psychology: official journal of the Division of Health Psychology, American Psychological Association. 2003. 22: 638–642

Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling. 2004. 178: 389–397

Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery. Molecules. 2020. 25: 1–17

Pavlov YL. Random Forests. Netherlands: Ridderprint. 2000

Peterson LE. K-nearest neighbor. Scholarpedia - Brain Corporation. 2009. 4: 1–13

Petrowski K, Paul S, Albani C, Brähler E. Factor structure and psychometric properties of the trier inventory for chronic stress (TICS) in a representative German sample. BMC medical research methodology. 2012. 12: 1–10

Pingle Y. Evaluation of mental stress using predictive analysis. International Journal of Engineering Research & Technology (IJERT). 2020. 9: 329–333

Power A, Burda Y, Edwards H, Babuschkin I, Misra V. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. ACM. 2022. 2022: 1–10

Prümper J, Hartmannsgruber K, Frese M. KFZA. Kurz-Fragebogen zur Arbeitsanalyse. Verlag für Angewandte Psycholoie. 1995: 125–132

Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. 2018. 1811: 1–49

Ratner B. Statistical and Machine-Learning Data Mining. Milton: Chapman and Hall/CRC. 2017

Reddy US, Thota AV, Dharun A. Machine Learning Techniques for Stress Prediction in Working Employees 2018 IEEE International Conference on Computational Intelligence and Computing Research. Piscataway, NJ: IEEE, 2018: 1–4

Ribeiro MT, Singh S, Guestrin C. Model-Agnostic Interpretability of Machine Learning. 2016: 91–95

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statist. Surv. 2022. 16: 1–85

Saeys Y, Abeel T, van de Peer Y. Robust Feature Selection Using Ensemble Feature Selection Techniques: Springer, Berlin, Heidelberg, 2008: 313–325

Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018. 8: 1-18

Saha B, Srivastava D. Data quality: The other face of Big Data 2014 IEEE 30th International Conference on Data Engineering (ICDE 2014). Piscataway, NJ: IEEE, 2014: 1294–1297

Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, Williams BA, Haggerty CM, Fornwalt BK. Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning. JACC. Cardiovascular imaging. 2019. 12: 681–689

Sanford LD, Suchecki D, Meerlo P. Stress, arousal, and sleep. Current topics in behavioral neurosciences. 2015. 25: 379–410

Schulz P, Schlotz W. Trierer Inventar zur Erfassung von chronischem Sre (TICS): Skalenkonstruktion, teststatistische Überprüfung und Validierung der Skala Arbeitsüberlastung. Hogrefe. 1999. 45: 1–19

Schulz P, Schlotz W, Becker P. Trierer Inventar zum Chronischen Stress (TICS) [Trier Inventory for Chronic Stress (TICS)]. Göttingen: Hogrefe. 2004

Sen J, Mehtab S, Engelbrecht A. Machine Learning. London, United Kingdom: IntechOpen. 2021

Siddiq M. ML-based Medical Image Analysis for Anomaly Detection in CT Scans, X-rays, and MRIs. Devotion Journal of Community Service. 2020. 2: 53–64

Siegrist J, Li J, Montano D. Psychometric properties of the effort-reward imbalance questionnaire. 2014: 1–28

Simon E, Darren V. Work-Related Quality of Life Scale. 2007: 1–8

Sriramprakash S, Prasanna VD, Murthy OR. Stress Detection in Working People. Procedia Computer Science. 2017. 115: 359–366

Stephens MAC, Wand G. Stress and the HPA axis: role of glucocorticoids in alcohol dependence. Alcohol Research: Current Reviews. 2012. 34: 468–483

Stubbs B, Vancampfort D, Veronese N, Schofield P, Lin P-Y, Tseng P-T, Solmi M, Thompson T, Carvalho AF, Koyanagi A. Multimorbidity and perceived stress: a population-based cross-sectional study among older adults across six low- and middle-income countries. Maturitas. 2018. 107: 84–91

Sutton RS, Barto A. Reinforcement learning. Cambridge, Massachusetts, London, England: The MIT Press. 2018

Suzuki K. Artificial Neural Networks. Rijeka, Croatia: IntechOpen. 2013

Taunk K, De S, Verma S, Swetapadma A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 2019 international conference on intelligent computing and control systems (ICCS). IEEE. 2019: 1255–1260

Thompson SC, Schlehofer MM. Perceived control. Health Behaviour Constructs: Theory, Measurement, and Research. National Cancer Institute. 2020: 1–17

van der Horst M, Coffé H. How Friendship Network Characteristics Influence Subjective Well-Being. Social indicators research. 2012. 107: 509–529

Vancheri F, Longo G, Vancheri E, Henein MY. Mental Stress and Cardiovascular Health-Part I. Journal of Clinical Medicine. 2022. 11: 1–17

Vemuri VK. The Hundred-Page Machine Learning Book. Journal of Information Technology Case and Application Research. 2020. 22: 136–138

Verhaeghe J, van der Donckt J, Ongenae F, van Hoecke S. Powershap: A Power-Full Shapley Feature Selection Method: Springer, Cham, 2023: 71–87

Viehmann A, Kersting C, Thielmann A, Weltermann B. Prevalence of chronic stress in general practitioners and practice assistants: Personal, practice and regional characteristics. PLOS ONE. 2017. 12: 1-13

Winter E. Chapter 53 The shapley value. Handbook of Game Theory with Economic Applications. 2002. 3: 2025–2054

Wright RE. Logistic regression. American Psychological Association. 1995: 217–244

Yuan Q, Cai T, Hong C, Du M, Johnson BE, Lanuti M, Cai T, Christiani DC. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal Cohort of Patients With Lung Cancer. JAMA Netw Open. 2021. 4: 1-14

Zafar MR, Khan N. Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. Machine Learning and Knowledge Extraction. 2021. 3: 525–541

Zhang C, Ma Y. Ensemble machine learning: methods and applications: Springer New York, NY. 2012

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2018. 4: 1–8

Zhenzhen, Yujie Y, Jing Z, Qi L, Denan L, Ye L, Jianping F, Wen C, Xie-Hui C, Yunpeng C. Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods. JMIR medical informatics. 2020. 8: 1–18

Zhou Z-H. Ensemble methods. Boca Raton, Fla.: CRC Press Taylor & Francis. 2012

9 Appendix

9.1 Publication 1

Chronic stress in practice assistants: An analytic approach comparing four machine learning classifiers with a standard logistic regression model

Arezoo Bozorgmehr, Anika Thielmann, Birgitta Weltermann

2021

DOI: https://doi.org/10.1371/journal.pone.0250842

RESEARCH ARTICLE

Chronic stress in practice assistants: An analytic approach comparing four machine learning classifiers with a standard logistic regression model

Arezoo Bozorgmehr[1]*, Anika Thielmann[1,2], Birgitta Weltermann[1,2]

1 Institute of General Practice and Family Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany, **2** Institute for General Medicine, University Hospital Essen, University of Duisburg Essen, Essen, Germany

* Arezoo.bozorgmehr@ukbonn.de

OPEN ACCESS

Citation: Bozorgmehr A, Thielmann A, Weltermann B (2021) Chronic stress in practice assistants: An analytic approach comparing four machine learning classifiers with a standard logistic regression model. PLoS ONE 16(5): e0250842. https://doi.org/10.1371/journal.pone.0250842

Editor: Alfredo Vellido, Universitat Politecnica de Catalunya, SPAIN

Received: July 30, 2020

Accepted: April 15, 2021

Published: May 4, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0250842

Data Availability Statement: The manuscript's data cannot be shared publicly because of ethical restrictions as our dataset includes potentially identifying information of personnel in general

Abstract

Background

Occupational stress is associated with adverse outcomes for medical professionals and patients. In our cross-sectional study with 136 general practices, 26.4% of 550 practice assistants showed high chronic stress. As machine learning strategies offer the opportunity to improve understanding of chronic stress by exploiting complex interactions between variables, we used data from our previous study to derive the best analytic model for chronic stress: four common machine learning (ML) approaches are compared to a classical statistical procedure.

Methods

We applied four machine learning classifiers (random forest, support vector machine, K-nearest neighbors', and artificial neural network) and logistic regression as standard approach to analyze factors contributing to chronic stress in practice assistants. Chronic stress had been measured by the standardized, self-administered TICS-SSCS questionnaire. The performance of these models was compared in terms of predictive accuracy based on the 'operating area under the curve' (AUC), sensitivity, and positive predictive value.

Findings

Compared to the standard logistic regression model (AUC 0.636, 95% CI 0.490–0.674), all machine learning models improved prediction: random forest +20.8% (AUC 0.844, 95% CI 0.684–0.843), artificial neural network +12.4% (AUC 0.760, 95% CI 0.605–0.777), support vector machine +15.1% (AUC 0.787, 95% CI 0.634–0.802), and K-nearest neighbours +7.1% (AUC 0.707, 95% CI 0.556–0.735). As best prediction model, random forest showed a sensitivity of 99% and a positive predictive value of 79%. Using the variable frequencies at the decision nodes of the random forest model, the following five work characteristics

practices. Data requests may be sent to the institutional ethics committee of Universitatsklinikum Bonn (ethik@ukbonn.de).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

influence chronic stress: too much work, high demand to concentrate, time pressure, complicated tasks, and insufficient support by practice leaders.

Conclusions

Regarding chronic stress prediction, machine learning classifiers, especially random forest, provided more accurate prediction compared to classical logistic regression. Interventions to reduce chronic stress in practice personnel should primarily address the identified workplace characteristics.

1. Introduction

Occupational stress is an important issue in health care and other workers worldwide [1]. Following stress models introduced by Selye, Lazarus and others, it was shown that chronic stress can lead to adverse (mental) health effects such as burnout or depression [2, 3]. Also, stress can produce temporary or even permanent alterations in memory [4], cognition [5], arousal/sleep [6, 7], and coping behaviours [8]. In our prior study with 214 general practitioners (GPs) and 550 practice assistants from 136 German general practices, we showed that 19.9% of the male GPs (n = 141), 35.6% of the female GPs (n = 73) and 26.4% of the practice assistants (PrAs) had high chronic stress [9]. Overall, the mean prevalence of high chronic stress was 26.3% in this workforce, which is more than twice as prevalent compared to the general population (11%) studied in the representative German Health Interview and Examination Survey for Adults (DEGS1) with more than 7.900 participants [10, 11]. Analyzing for various work and (regional) practice characteristics, we showed that only the weekly working hours correlated with high chronic stress in GPs and PrAs.

However, aiming to develop effective prevention strategies, a more profound understanding of factors causing and/or contributing to high psychological strain on an individual and group level is needed. As workplaces typically are complex and multifactorial social organizations, appropriate statistical methods are needed to analyse for complex associations and cause-effect relationships. Prior studies addressing impaired psychological well-being in primary care workers used standard statistical procedures such as prevalence ratios and logistic regression models to evaluate for associations [9, 12, 13]. These statistical approaches usually simplify the complex relationships between independent variables (features) and response variable (dependent variable): they assume that each independent variable is linked to the outcome by a linear statistical function. This is especially problematic when datasets with large numbers of non-linear interactions and interaction effects between independent variables occur, which make the model more complex [14]. Nowadays, machine learning (ML) approaches offer new opportunities to evaluate complex relationships. Conceptually, ML has the benefit that it efficiently exploits complex and non-linear interactions between variables by minimizing the error between predicted and observed response variables and improve the accuracy of the models compared to standard approaches [15, 16]. By using a large dataset available on practice assistants from our prior study, we aim to develop better understanding workplace factors, associated with chronic stress in practice assistants using machine learning. Thus, we compare four machine learning classifiers (random forest, support vector machine, K-nearest neighbors', artificial neural network) with a standard logistic regression model using standard measurements to compare test accuracy, i.e. to derive the best prediction model for chronic stress in practice assistants in primary care.

Regarding terminology, we like to point out that we use the term "prediction" as used in the context of machine learning: it refers to the output of an algorithm after it has been trained on a dataset and applied to new data to forecast the likelihood of a particular outcome. In contrast, in epidemiological analyses, a (risk) prediction model refers to a mathematical equation that uses patient characteristics (risk factors) to estimate the probability of a defined outcome prospectively.

2. Methods

2.1 Data source

The dataset used for the analyses was derived from our cross-sectional study addressing stress among general practice personnel (GPs, PrAs), which was performed among general practices belonging to the teaching practice network of the Institute for General Medicine, University Hospital Essen, Essen, Germany. A total of 764 professionals from 136 practices had taken part in the survey, which was performed in 2014. The design of the study and key results addressing the 214 GPs (practice owners and employed physicians) and 550 practice assistants (PrAs) (including medical secretaries and practice assistants in trainees) are published [9]. This analysis addresses chronic stress in 550 practice assistants (PrAs), which are the largest professional group in general practices. We documented that 26.4% of the 550 practice assistants (PrAs) had high chronic stress, as well as 19.9% of the male (n = 141) and 35.6% of the female (n = 73) general practitioners (GPs) [9]. In this workforce, the average of workers with high chronic stress was 26.3% (n = 201).

2.2 Ethics statement

Ethical approval for the survey had been obtained from the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen (reference number: 13-5536-BO, date of approval: 24/11/2014). All participants had received written information and signed informed consent forms. The principal investigator of the study (B.W) and coauthor of this manuscript provided the data for this analysis.

2.3 Outcome

The primary outcome is strain due to chronic stress over the past three months. Chronic stress was measured using the German short version of the standardized, validated, self-administered TICS-SSCS questionnaire [17, 18]. This instrument measures strain due to chronic stress for the past three months. It consists of 12 items on 5-point Likert scales (0 = 'never' und 4 = 'very often'). The TICS-SSCS values are added to a sum-score. The score ranges from 0 to 48 with 0 denoting 'never stressed' and 48 'very often stressed', and reflects subjective strain due to chronic stress [17, 18]. Following the definition of chronic stress of our prior analysis, the TICS scores were dichotomized using the median (TICS = 23) as cut-off (0 = no chronic stress (TICS $<$ 23), 1 = strain due to chronic stress (TICS $\geq$ 23)).

2.4 Socio-demographic and workplace characteristics

A total of 64 sociodemographic and workplace characteristics were used for the analyses. The sociodemographic characteristics included e.g., age, marital status, number of persons in household. Work-related characteristics comprised details on the employment (e.g., number of hours per week, work status, employment contract), duties in practice (e.g., reception, telephone, prescription, blood pressure measurement) and subjective perceptions of workload (e.g., self-determination of sequence of work steps, influence on work assigned, plan the work

independently). The standardized 'short questionnaire for workplace analysis' (German: Kurzfragebogen zur Arbeitsanalyse (KFZA)) was used to assess workplace characteristic [19]. For details on the work characteristics see Tables 1–3. In line with the TICS instrument, which addresses strain due to chronic stress during the past three months, all workplace characteristics had been requested regarding the past three months (see Table 4).

2.5 Statistical analysis

2.5.1 Handling of missing data. Missing values were observed in 0.2% to 11%. If missing data were above 5%, this is indicated in the Tables 1–3. Common imputation methods for supervised learning were applied to handle missing data [20]. The K-nearest neighbors algorithm was used for imputing missing values in TICS scores with k = 10. For continuous variables we used median imputation and for categorical variables a separate category 'unknown' [20].

2.5.2 Preparation of datasets for machine learning. After pre-processing the data to compare machine learning classifiers, the dataset was split into a 'training' and a 'validation' dataset. Fig 1 illustrates the study process flow. We used the 10-fold cross validation approach

Table 1. Sociodemographic characteristics of practice assistants (n = 550) and strain due to chronic stress (measured by the standardized and validated TICS tool): Items and sum scores.

	Participants (N = 550)		
Continuous variables	**Mean**	**SD**	**Range**
Age	38	12.61	16–71
Persons in household more age 18	2	1.12	0–6
Persons in household below age 18	1	0.84	0–6
Number of physicians in practice	3	2.16	1–10
Number of practice assistants in practice	8	7.66	0–35
Categorical variables	**n**		**%**
Female gender	544		99.3
Marital status			
Married	277		50.4
Single	221		40.2
Divorced	45		8.2
Widowed	7		1.3
Number of persons in household	72		13.1
Cares for next of kin	75		13.6
Working hours/week			
1–9 hours	12		2.2
10–19 hours	52		9.5
20–29 hours	116		21.1
30–39 hours	221		40.2
40–49 hours	116		21.1
50–59 hours	12		2.2
>60 hours	10		1.8
Working full time	364		66.2
Has open-ended employment contract	466		84.7
Had participated in stress seminar in the past	31		5.6
Had used counseling for stress reduction	50		9.1
High strain due to chronic stress (TICS ≥ 23)	125		22.7

https://doi.org/10.1371/journal.pone.0250842.t001

Table 2. Practice and workplace characteristics during the past three months (n = 550 practice assistants).

Practice characteristics		
Practice structure		
Working in group practice	296	53.8
Working in single physician practice	147	26.7
Working in practice with several locations	50	9.1
Working in practice with an employed physician	39	7.1
Working in privately owned health center	6	1.1
Medical records		
Electronic medical records (EHR)	348	63.3
Paper and electronic records	187	34.0
Practice services		
Emergent home visits	515	93.6
Practice offers regular home visits	511	92.9
Nursing home visits*	508	92.4
Tasks of practice assistant during past 3 months		
Scheduled appointments	518	94.2
Documented in patients´ EHR	513	93.3
Prepared prescriptions	504	91.6
Pulled up paperhealth records or opened electronic patient files	500	90.9
Performed phone service	499	90.7
Worked at reception	486	88.4
Obtained blood pressure readings	461	83.8
Performed ECGs	430	78.2
Prepared practice equipment for the day and switch them off in the evening	414	75.3
Performed laboratory work	393	71.5
Supported physician during patient-consultations	363	66.0
Supported billing of statutory health insurance patients	358	65.1
Performed disease-management examinations	332	60.4
Applied long-term blood pressure devices*	327	59.5
Ordered medical supply	284	51.6
Applied long-term ECG*	247	44.9
Ordered office supply	239	43.5
Performed treadmill testing	237	43.1
Supported billing of private patients*	236	42.9
Performed doppler examination of foot vessels/measured ankle-arm index*	103	18.7

*__Missing values above 5%__

https://doi.org/10.1371/journal.pone.0250842.t002

in machine learning models to measure the unbiased prediction accuracy of the models (see Fig 2). Based on the literature, 10 was chosen as optimal number of folds, which optimizes the time to complete the test while minimizing the bias and variance associated with the validation process [21–23]. The K-Fold cross validation method also called rotation estimation is used to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more machine learning methods. In this method the complete dataset (D) is randomly split into k mutually exclusive subsets (the folds: D1, D2,. . ., Dk) of approximately equal size. The classification model is trained and tested k times. Each time (t 2 {1, 2,. . ., k}), it is trained on all but one folds (Dt) and tested on the

Table 3. Self-assessment of workplace situation (n = 550 practice assistants).

Work aspects	Workplace factor	Mean Score (PrAs)	95% CI
Job content	Versatility	3.6	3.58–3.7
	Completeness of task	3.5	3.41–3.57
Resources	Scope of action	3.4	3.37–3.49
	Social support	4.0	3.98–4.12
	Cooperation	3.6	3.53–3.66
Stressors	Qualitative work demands	2.2	2.14–2.29
	Quantitative work demands	2.9	2.83–3.01
	Work disruptions	2.7	2.67–2.81
	Workplace environment	2.2	2.13–2.3
Organizational culture	Information and participation	3.6	3.57–3.73
	Benefits	2.9*	2.77–2.94

https://doi.org/10.1371/journal.pone.0250842.t003

remaining single fold (Dt). The cross validation estimate of the overall accuracy is calculated as the average of the k individual accuracy measures by formula:

$$CVA = \sum_{i=1}^{k} A_i \quad (1)$$

Where CVA stands for cross-validation accuracy, k is the number of folds used, and A is the accuracy measure of each fold [21].

2.5.3 Logistic regression as standard statistical procedure. *Logistic Regression (LR)* is a classical statistical modelling procedure to analyze one dependent dichotomous or binary outcome and one or more nominal, ordinal, interval or ratio-level independent variables. LR models are frequently applied to exposure-event studies in medical research, because they can be used to estimate the model predictors' odds ratio [24]. All variables significant in bivariate analysis were included in the logistic regression model.

2.5.4 Machine learning approaches. *1) K-Nearest Neighbors (KNN)* classifies an object by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer). If k = 1, the object is simply assigned to the class of its nearest neighbor. KNN is a type of instance-based or lazy learning where the

Table 4. Chronic stress of practice assistants: Results of TICS (Trierer Inventory of Chronic Stress) (n = 550).

How often in the last 3 months did you experience . . .	Never	Rarely	Sometimes	Frequently	Very Frequently
	n(%)	n(%)	n(%)	n(%)	n(%)
Fear, something unpleasant might occur	72 (13.1)	213 (38.7)	190 (34.5)	54 (9.8)	21 (3.8)
Lack of recognition for good performance	158 (28.7)	157 (28.5)	121 (22.0)	71 (12.9)	42 (7.6)
Times with too many obligations	38 (6.9)	119 (21.6)	167 (30.4)	157 (28.5)	67 (12.2)
Times when being unable to suppress worrying thoughts	90 (16.4)	174 (31.6)	182 (33.1)	83 (15.1)	21 (3.8)
Work is not appreciated despite doing the best	157 (28.5)	200 (36.4)	116 (21.1)	56 (10.2)	20 (3.6)
Everything is too much	86 (15.7)	174 (31.7)	174 (31.7)	85 (15.5)	30 (5.5)
Times of worry and one cannot stop it	138 (25.1)	186 (33.9)	139 (25.3)	57 (10.4)	29 (5.3)
Times when being unable to perform as expected	120 (21.8)	299 (54.4)	107 (19.5)	19 (3.5)	5 (0.9)
Times in which the responsibility for others is a burden	162 (29.5)	215 (39.1)	123 (22.4)	42 (7.6)	8 (1.5)
Times when the work gets too much	85 (15.5)	205 (37.3)	183 (33.3)	60 (10.9)	17 (3.1)
Fear of not being able to perform the tasks	126 (22.9)	229 (41.6)	137 (24.9)	43 (7.8)	15 (2.7)
Times when being overwhelmed with worries	165 (30.0)	189 (34.4)	128 (23.3)	45 (8.2)	23 (4.2)

https://doi.org/10.1371/journal.pone.0250842.t004

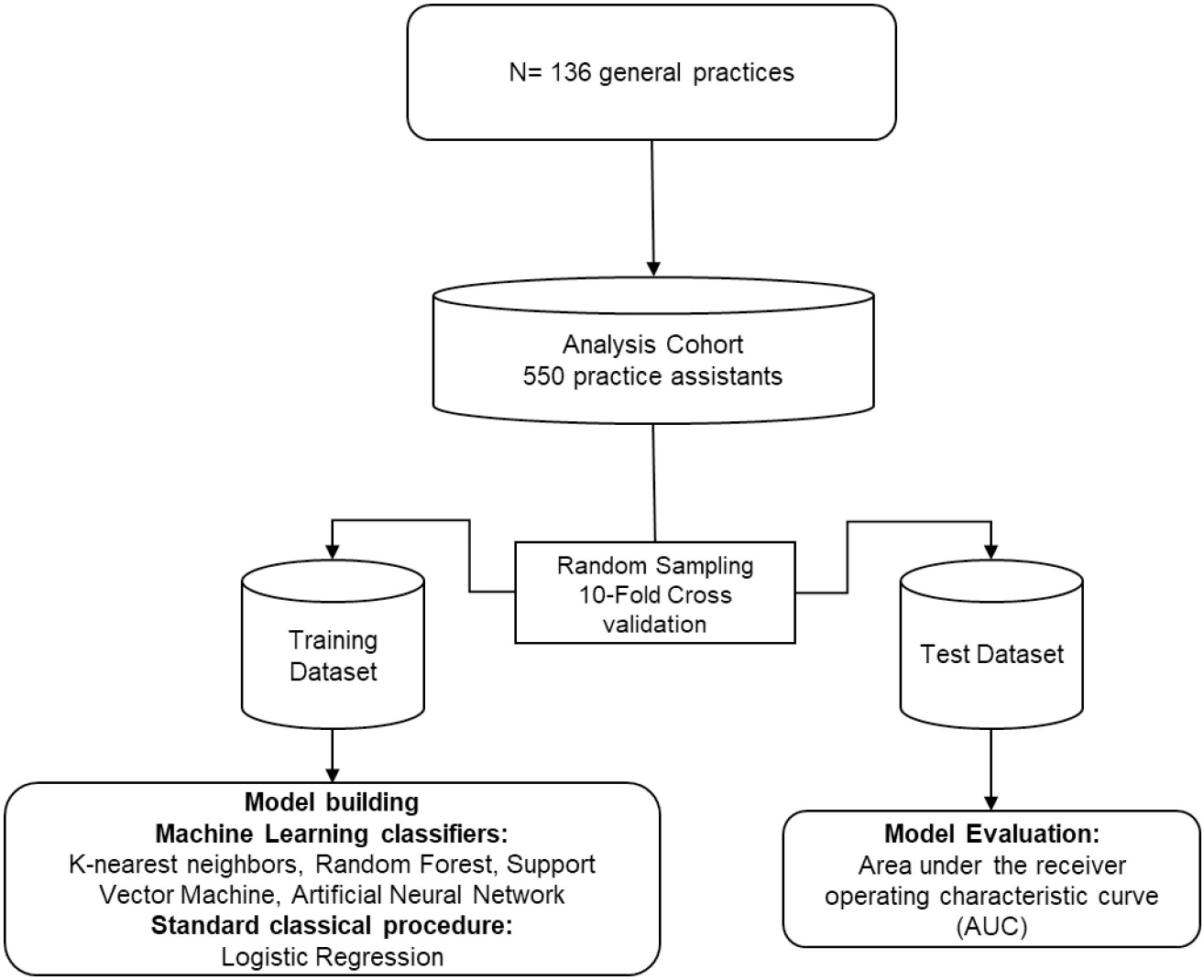


Fig 1. Machine learning data extraction process flow.

https://doi.org/10.1371/journal.pone.0250842.g001

function is only approximated locally and all computation is deferred until classification [25, 26]. In this study, we used KNN applying k = 10 neighbors, which are the ten closest observations in multidimensional space based on Euclidean distance function to model the training dataset.

2) Support Vector Machine (SVM) represents different outcome classes in a hyperplane in multidimensional space to find the maximum marginal hyperplane. SVM generates the hyperplane in an iterative manner to minimize the error. A basic SVM is a non-parametric linear classifier that creates a hyperplane using the Euclidean distance function from the nearest input values to determine the target states. In order to obtain probability estimates, a logistic regression model is fitted to the output of the support vector machine [25]. In this study, the SVM classifier used RBF (Radial basis function) kernel, a training error of 1.0E-12, and a default boundary tolerance of a 1.0E-03 hyperplane. To obtain proper probability estimates, we used the option that fits calibration models to the outputs of the SVM.

3) Random Forest (RF) is a collection of decision trees, each constructed in a bootstrapped sample and from a random subset of the possible predictors at each node. RF is used to reduce

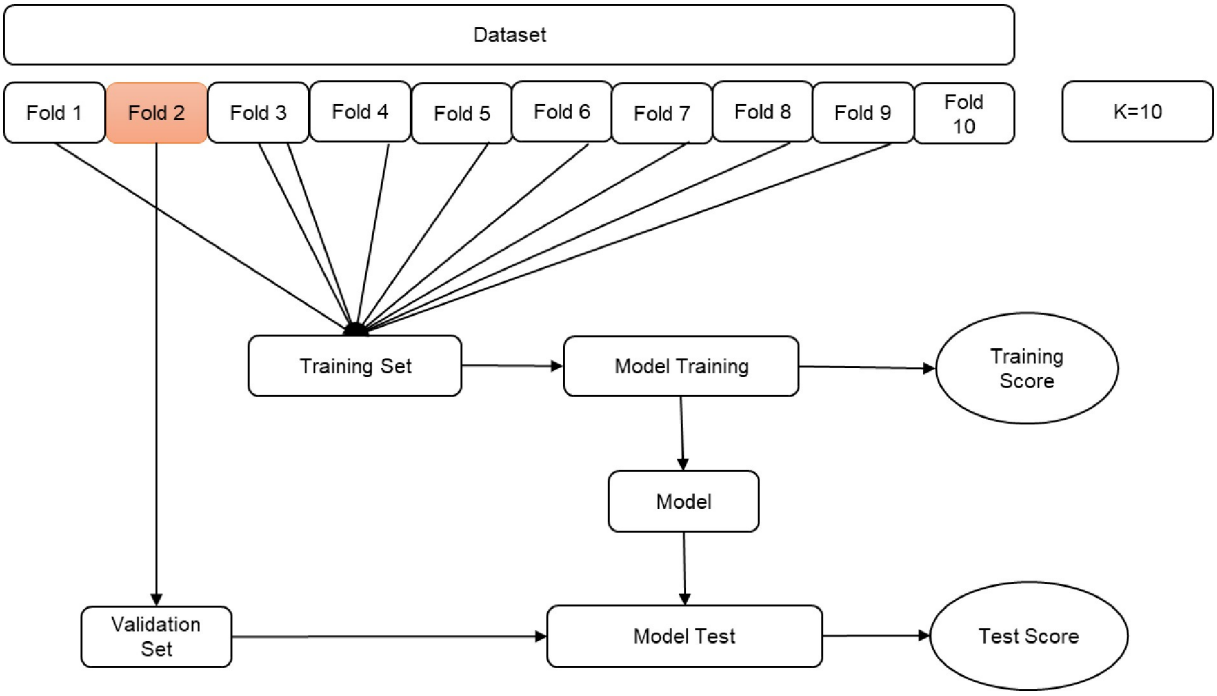


Fig 2. K-Fold cross validation.

https://doi.org/10.1371/journal.pone.0250842.g002

variance associated with decision trees [27, 28]. In this study, the forest is constructed consisting of randomly 1,000 individual trees. A large number of trees increases the predictive accuracy of RF models and the forest does not require extensive tuning [29]. Due to the insensitivity of error rates to the number of features selected to split each node, we used the default of a random sample of $\sqrt{n}$ of predictors at each node with n being the total number of predictors under consideration. The predicted probability was derived based on average prediction across all of the trees.

4) Artificial Neural Network (ANN) is a computational and flexible model that expresses complex non-linear relationships among features, which consist of an interconnected group of variables. A basic ANN model consists of three layers of neurons, i.e. input, output, and hidden layer. These layers can learn from data iteratively through a backpropagation classifier. It trains a multilayer perceptron with one hidden layer, an input layer with the number of nodes equal to the sum of features, and an output layer [30]. This study used a multilayer Perceptron classifier with one hidden layer, a learning rate value with decay of 0.3, and a momentum rate for the backpropagation classifier of 0.2. Suitable ranges for these parameters are within 0.15–0.8 for learning rate and 0.1–0.4 for momentum [30].

Development of the models was completed using Python (Version 3.7.3) and Python's Scikit-Learn library (https://scikit-learn.org/stable/).

3. Results

3.1 Sociodemographic and workplace characteristics of the study population

The dataset comprised results of 550 PrA from 136 general practices. The vast majority of the total of PrAs were females (98.9%) with a mean age of 38 years (SD 12.6). Regarding the

marital status, 50.6% (n = 277) of the PrAs were married. On average, they worked in the current practice for 18.8 years (SD 12.5), 32.5% in part-time.

3.2. Primary outcome: Strain due to chronic stress

The TICS score of the population ranged from 0 to 44 with a mean of 17.2 and median of 17.0. In the total dataset, 22.7% (n = 125) had high strain due to chronic stress versus 77.3% (n = 425) low strain due to chronic stress. Regarding socio-demographic characteristics personnel with high strain due to chronic stress showed the following significant differences compared to those with low strain: older PrAs (mean 38.76) vs. younger PrAs (mean 24.36), unmarried PrAs (29.4%) vs. married PrAs (17%). While caring for next of kin did not differ between groups. No gender-specific distribution was applied, because PrAs were predominantly female (98.9%). All regression and machine learning approaches were applied to the dataset with female subjects only (n = 546).

3.3. Results of four machine learning classifiers

3.3.1 Prediction accuracy. The performance of the machine learning classifiers was assessed using the validation dataset by calculating Harrell's c-statistic, a measure of the total area under the receiver operating characteristic curve (AUC) [31]. The results showed an AUC of 0.844 (95%CI, 0.684–0.843) for RF, 0.760 (95%CI, 0.605–0.777) for ANN, 0.787 (95%CI, 0.634–0.802) for SVM, and 0.707 (95%CI, 0.556–0.735) for KNN.

3.3.2 Classification analysis. Corresponding results of sensitivity and positive prediction value (PPV) for machine learning were 99% and 79% for RF, 87% and 85% for ANN, 87% and 86% for the SVM, and 99% and 78% for KNN.

3.4. Results of Logistic regression analysis

In bivariate analysis, the following factors were associated with strain significantly: persons in household below age 18, marital status, age, working hours/week, room equipment, work status, performed laboratory work, obtained blood pressure readings, and performed doppler examination of foot vessels/measured ankle-arm index as duties in practice. C statistics for logistic regression showed an AUC of 0.636 (95%CI, 0.490–0.674). This model predicted 316 cases correctly from 425 total cases, with a sensitivity of 75% and positive prediction value (PPV) of 44%.

3.5. Comparison of ML and regression analysis

The prediction accuracy according to the discrimination (AUC c-statistic) value is shown in Table 5 for all models. All machine learning models achieved statistically improvements in compared to the standard logistic regression model: +20.8% for RF, +15.1% for SVM, +12.4% for ANN, and +7.1% for KNN. Random forest is performing well out of all four machine learning classifiers. RF classifier resulted in a net increase of 104 strain due to chronic stress cases from the logistic regression baseline model, increasing the sensitivity to 99% and PPV to 79%. See Table 6 for more details of machine learning models.

3.6. Variable rankings in machine learning models

Of the 4 ML approaches used, variable importance can only be determined in artificial neural network and random forest. Artificial neural network model uses the overall weighting of the variables within the model. Random forest ranks variable importance based on decision-trees on the selection frequency of the variable as a decision node. For KNN does not provide a

Table 5. Performance of the machine learning algorithms predicting chronic stress derived from applying training algorithms on the validation dataset. Higher c-statistics results in better algorithm discrimination. The baseline (BL) standard logistic regression model is provided for comparative purposes.

Algorithms	AUC c-statistic	95% Confidence Intervall		Absolute change in AUC (%)
		LCL	UCL	
BL: Logistic Regression	0.636	0.490	0.674	[Reference]
ML: K-nearest Neighbours	0.707	0.556	0.735	+7.1%
ML: Support Vector Machine	0.787	0.634	0.802	+15.1%
ML: Artificial Neural Network	0.760	0.605	0.777	+12.4%
ML: Random Forest	0.844	0.684	0.843	+20.8%

https://doi.org/10.1371/journal.pone.0250842.t005

method for the importance or coefficients of variables. We used a nonlinear SVM classifier with RBF kernel, which has no variable importance methods. The variable importance was determined by the coefficient effect size for logistic regression model. The identified factors such as persons in household below age 18, age below 35 years old, and insufficient room equipment that have identified by logistic regression, has also identified by ANN and RF. The most determined factors by both of ANN and RF included work related characteristics such as too much work, high demand to concentrate, time pressure, complicated tasks, and insufficient practice room conditions (See Table 7).

4. Discussion

To the best of our knowledge, this study is the first to use machine learning for a better understanding of stress in primary care practice personnel. Comparing four common machine learning (ML) approaches to a classical statistical procedure, we showed that all four machine learning approaches provided more accurate models for the prediction of strain due to chronic stress than as standard regression analysis. Random forest showed the highest accuracy with workload, high demand to concentrate, and time pressure being the most important factors associated with chronic stress. These factors were also identified in other studies in the target populations GPs and GP practice personnel. Addressing job satisfaction, Harris et al. identified time pressure as the most frequent stressor in a study with 626 Australian practice staff in 96 general practices [12]. Studying 158 Canadian family physicians, Lee et al. determined the following occupational stressors as relevant: challenging patients, high workload, time limitations, competency issues, challenges of documentation and practice management and changing roles within the workplace [13, 32]. Similarly, Hoffmann et al. showed that the work disruption was a negative relevant workplace factor in study with 550 practice assistants [33].

Table 6. Full details on classification analysis.

Algorithms	Chronic stress cases correct (True Positive)	Chronic stress cases incorrect (False Negative)	Total chronic stress cases	Non-chronic stress cases correct (True Negative)	Non-chronic stress cases incorrect (False Positive)	Total non-chronic stress cases	Sensitivity (True Positive)	Positive Predictive Value (PPV)
Logistic Regression	316	109	425	68	57	125	0.751	0.440
ML: Random Forest	420	5	425	15	110	125	0.988	0.792
ML: K-nearest Neighbours	421	4	425	6	119	125	0.991	0.780
ML: Support Vector Machine	369	56	425	66	59	125	0.868	0.862
ML: Artificial Neural Network	369	56	425	59	66	125	0.868	0.848

https://doi.org/10.1371/journal.pone.0250842.t006

Table 7. The most influential predictor variables associated with chronic stress listed by coefficient effect size (Standard logistic regression) weighting (Artificial neural network) and selection frequency (Random forest).

Standard model		Machine learning models			
Logistic regression	Coefficient	Artificial Neutral Network	Weight (%)	Random Forest	Frequency
Obtained blood pressure readings	0.951	Too much work	39.7	Too much work	0.73
Persons in household below age 18	0.349	High demand to concentrate	39.3	High demands to concentrate	0.71
Working hours/week more than 40	0.121	Time pressure	36.7	Time pressure	0.70
Work status	-0.109	Complicated tasks	31.5	Complicated tasks	0.67
Performed laboratory work	0.091	Insufficient practice room conditions	18.1	Age $\leq$ 35	0.63
Employment contract	0.063	Interrupted during work	14.9	Insufficient support by practice leaders	0.52
Age $\leq$ 35	0.045	Persons in household below age 18	13.8	Insufficient workplace environment	0.51
Insufficient workplace environment	0.028	Working hours/week more than 40 hours	12.7	Insufficient practice room conditions	0.50
Performed doppler examination of foot vessels/measured ankle-arm index	0.018	Workplace environment	12.3	Holding together well	0.48
Marital status/single	0.006	Number of practitioners in the practice	10.6	Influence on work assigned	0.43

https://doi.org/10.1371/journal.pone.0250842.t007

These stressors are described to influence poor physician well-being and adverse patient outcomes such as low patient satisfaction [34]. The relevance of such chronic psychological burden is tremendous as it was shown that physiological responses due to stress negatively affect e.g. memory, immune system functions, the function of the cardiovascular system, and brain electric activity [35, 36].

4.1 Comparison to other ML analyses

There are a few other studies from other medical fields, which compared standard statistical and ML approaches, similar to our results. Machine learning is considered a branch of artificial intelligence, which extracts meaningful patterns from data and develops prediction models using several algorithms [37]. ML approaches integrate many different levels of data to develop a new approach to classification based on medical issues such as chronic stress and linked more precisely to interventions for a given individual. Better model accuracy by machine learning was also found in an UK study on cardiovascular risk prediction. Using routine clinical data of 378,256 patients four machine learning algorithms (random forest, logistic regression, gradient boosting, and neural network) were compared to an established algorithm (American College of Cardiology guidelines) to predict first cardiovascular event over 10-years [38]. Neural network performed best, with a predictive accuracy improving by 3.6% compared to baseline algorithm. Using a dataset with 9.502 heart failure patients and a one-year follow-up, a US study compared four machine learning methods (least absolute shrinkage and selection operation regression, classification and regression trees, random forests, and gradient boosted modeling (GBM)) with logistic regression as a classical statistical procedure to predict four heart failure outcomes. The C statistic results for all outcomes show that ML methods were better calibrated and that gradient-boosted (GMB) model was the most consistent ML modeling approach [39]. In the field of oncology, a large American study on breast cancer survival compared two ML algorithms (artificial neural network and decision trees) to classical statistical logistic regression using a large dataset with more than 200,000 cases. The decision tree approach was the best predictor with 93.6% prediction accuracy, followed by

artificial neural network with 91.2% and LR with 89.2% [40]. Overall, machine learning approaches yielded more accurate results than classical methods in our and the above-mentioned studies.

4.2 Strength and limitations

The key strength of this study is the comparison of a range of machine learning approaches in the field of healthcare workers´ well-being. Chronic stress measurement approaches based on self-reported questionnaires [17, 41] are subjective and cannot provide immediate information about the state of a person. A continuous stress monitoring using data mining technology helps to better understand stress patterns and also provide better insights about possible future interventions.

Limitations of this study include the rather small sample size and the large number of predictor variables (features), which poses a risk for overfitting [42, 43]. One of the key components of predictive accuracy is the amount and quality of the data to provide better results. Furthermore, our data source contained practice assistants from the German region only, which limits generalizability and requires validation in populations from other countries where job tasks and challenges might be different. Although the data collection was conducted in 2014, the results still apply to German practices, except that the COVID pandemic likely increased workload and psychological burden, which we are currently evaluating in an ongoing study [11]. Prospectively, research using continuous stress monitoring and data mining technologies will help to better understand stress patterns and provide even deeper insights for possible future interventions.

5. Conclusion

Compared to logistic regression as a classical statistical procedure, this study showed that all machine learning classifiers provided more accurate models for the prediction of chronic stress in practice assistants with random forest performing best. Identification of chronic stress is of importance for the well-being and productivity of practice assistants. RF identified prominent predictor variables (features) that influence chronic stress which should be considered when developing interventions to reduce chronic stress.

Acknowledgments

We would like to thank all participating practices for their support of the study.

Author Contributions

Conceptualization: Arezoo Bozorgmehr.

Data curation: Arezoo Bozorgmehr.

Formal analysis: Arezoo Bozorgmehr.

Methodology: Arezoo Bozorgmehr, Birgitta Weltermann.

Project administration: Birgitta Weltermann.

Software: Arezoo Bozorgmehr.

Supervision: Birgitta Weltermann.

Validation: Arezoo Bozorgmehr.

Visualization: Arezoo Bozorgmehr.

Writing – original draft: Arezoo Bozorgmehr.

Writing – review & editing: Anika Thielmann.

References

1. Schreibauer EC, Hippler M, Burgess S, Rieger MA, Rind E. Work-Related Psychosocial Stress in Small and Medium-Sized Enterprises: An Integrative Review. Int J Environ Res Public Health. 2020; 17. Epub 2020/10/13. https://doi.org/10.3390/ijerph17207446 PMID: 33066111.
2. Dreher A, Theune M, Kersting C, Geiser F, Weltermann B. Prevalence of burnout among German general practitioners: Comparison of physicians working in solo and group practices. PLoS One. 2019; 14: e0211223. https://doi.org/10.1371/journal.pone.0211223 PMID: 30726284.
3. Luken M, Sammons A. Systematic Review of Mindfulness Practice for Reducing Job Burnout. Am J Occup Ther. 2016; 70:7002250020p1–7002250020p10. https://doi.org/10.5014/ajot.2016.016956 PMID: 26943107.
4. Alzoubi KH, Abdel-Hafiz L, Khabour OF, El-Elimat T, Alzubi MA, Alali FQ. Evaluation of the Effect of Hypericum triquetrifolium Turra on Memory Impairment Induced by Chronic Psychosocial Stress in Rats: Role of BDNF. Drug Des Devel Ther. 2020; 14:5299–314. Epub 2020/12/01. https://doi.org/10.2147/DDDT.S278153 PMID: 33299301.
5. Datta D, Arnsten AFT. Loss of Prefrontal Cortical Higher Cognition with Uncontrollable Stress: Molecular Mechanisms, Changes with Age, and Relevance to Treatment. Brain Sci. 2019; 9. Epub 2019/05/17. https://doi.org/10.3390/brainsci9050113 PMID: 31108855.
6. Sanford LD, Suchecki D, Meerlo P. Stress, arousal, and sleep. Curr Top Behav Neurosci. 2015; 25:379–410. https://doi.org/10.1007/7854_2014_314 PMID: 24852799.
7. Hu Y, Visser M, Kaiser S. Perceived Stress and Sleep Quality in Midlife and Later: Controlling for Genetic and Environmental Influences. Behav Sleep Med. 2020; 18:537–49. Epub 2019/06/23. https://doi.org/10.1080/15402002.2019.1629443 PMID: 31232098.
8. Kaldewaij R, Koch SBJ, Volman I, Toni I, Roelofs K. On the Control of Social Approach-Avoidance Behavior: Neural and Endocrine Mechanisms. Curr Top Behav Neurosci. 2017; 30:275–93. https://doi.org/10.1007/7854_2016_446 PMID: 27356521.
9. Viehmann A, Kersting C, Thielmann A, Weltermann B. Prevalence of chronic stress in general practitioners and practice assistants: Personal, practice and regional characteristics. PLoS One. 2017; 12: e0176658. https://doi.org/10.1371/journal.pone.0176658 PMID: 28489939.
10. Hapke U, Maske UE, Scheidt-Nave C, Bode L, Schlack R, Busch MA. Chronischer Stress bei Erwachsenen in Deutschland: Ergebnisse der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1). Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2013; 56:749–54. https://doi.org/10.1007/s00103-013-1690-9 PMID: 23703494.
11. Weltermann BM, Kersting C, Pieper C, Seifried-Dübon T, Dreher A, Linden K, et al. IMPROVEjob—Participatory intervention to improve job satisfaction of general practice teams: a model for structural and behavioural prevention in small and medium-sized enterprises—a study protocol of a cluster-randomised controlled trial. Trials. 2020; 21:532. https://doi.org/10.1186/s13063-020-04427-7 PMID: 32546256.
12. Harris MF, Proudfoot JG, Jayasinghe UW, Holton CH, Powell Davies GP, Amoroso CL, et al. Job satisfaction of staff and the team environment in Australian general practice. Med J Aust. 2007; 186:570–3. https://doi.org/10.5694/j.1326-5377.2007.tb01055.x PMID: 17547545.
13. Lee FJ, Stewart M, Brown JB. Stress, burnout, and strategies for reducing them: what's the situation among Canadian family physicians. Can Fam Physician. 2008; 54:234–5. PMID: 18272641
14. Jaccard J. Interaction effects in factorial analysis of variance. Thousand Oaks, Calif.: Sage; 2005.
15. Obermeyer Z, Emanuel EJ. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016; 375:1216–9. https://doi.org/10.1056/NEJMp1606181 PMID: 27682033.
16. Weng W-H. Machine Learning for Clinical Predictive Analytics. In: Celi LA, Majumder MS, Ordóñez P, Osorio JS, Paik KE, et al., editors. LEVERAGING BIG DATA IN GLOBAL HEALTH. [S.l.]: SPRINGER NATURE; 2020. pp. 199–217.
17. Petrowski K, Paul S, Albani C, Brähler E. Factor structure and psychometric properties of the trier inventory for chronic stress (TICS) in a representative German sample. BMC Med Res Methodol. 2012; 12:42. https://doi.org/10.1186/1471-2288-12-42 PMID: 22463771.
18. Schulz P, Schlotz W. Trierer Inventar zur Erfassung von chronischem Streß (TICS): Skalenkonstruktion, teststatistische Überprüfung und Validierung der Skala Arbeitsüberlastung. Diagnostica. 1999; 45:8–19. https://doi.org/10.1026//0012-1924.45.1.8

19. Prümper J., Hartmannsgruber K. & Frese, 1995 (M.). KFZA–Kurzfragebogen zur Arbeitsanalyse. Available from: https://fragebogen-arbeitsanalyse.at/help.
20. Poulos J, Valle R. Missing Data Imputation for Supervised Learning. Applied Artificial Intelligence. 2018; 32:186–96. https://doi.org/10.1080/08839514.2018.1448143
21. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection.; 1995.
22. Jiang G, Wang W. Error estimation based on variance analysis of k -fold cross-validation. Pattern Recognition. 2017; 69:94–106. https://doi.org/10.1016/j.patcog.2017.03.025
23. Steyerberg EW. Validation in prediction research: the waste by data splitting. J Clin Epidemiol. 2018; 103:131–3. Epub 2018/07/29. https://doi.org/10.1016/j.jclinepi.2018.07.010 PMID: 30063954.
24. Hilbe JM. Logistic regression models. Boca Raton, London, New York: CRC Press; 2017.
25. Kuhn M, Johnson K. Applied predictive modeling. 5th ed. New York: Springer; 2016.
26. Boehmke BC, Greenwell B. Hands-on machine learning with R. Boca Raton, London, New York: CRC Press; 2020.
27. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32. https://doi.org/10.1023/A:1010933404324
28. Denisko D, Hoffman MM. Classification and interaction in random forests. Proc Natl Acad Sci U S A. 2018; 115:1690–2. Epub 2018/02/12. https://doi.org/10.1073/pnas.1800256115 PMID: 29440440.
29. Probst P, Wright MN, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. WIREs Data Mining Knowl Discov. 2019; 9. https://doi.org/10.1002/widm.1301
30. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay. US Naval Research Laboratory Technical Report. 2018. Available from: https://arxiv.org/pdf/1803.09820.
31. Newson R. Confidence Intervals for Rank Statistics: Somers' D and Extensions. The Stata Journal. 2006; 6:309–34. https://doi.org/10.1177/1536867X0600600302
32. Lee FJ, Brown JB, Stewart M. Exploring family physician stress: helpful strategies. Can Fam Physician. 2009; 55:288–289.e6. Available from: https://www.cfp.ca/content/55/3/288.short. PMID: 19282541
33. Hoffmann J, Kersting C, Weltermann B. Practice assistants´ perceived mental workload: A cross-sectional study with 550 German participants addressing work content, stressors, resources, and organizational structure. PLoS One. 2020; 15:e0240052. https://doi.org/10.1371/journal.pone.0240052 PMID: 33002064.
34. Shanafelt TD, West C, Zhao X, Novotny P, Kolars J, Habermann T, et al. Relationship between increased personal well-being and enhanced empathy among internal medicine residents. J Gen Intern Med. 2005; 20:559–64. https://doi.org/10.1111/j.1525-1497.2005.0108.x PMID: 16050855.
35. Yaribeygi H, Panahi Y, Sahraei H, Johnston TP, Sahebkar A. The impact of stress on body function: A review. EXCLI J. 2017; 16:1057–72. https://doi.org/10.17179/excli2017-480 PMID: 28900385.
36. Lotfan S, Shahyad S, Khosrowabadi R, Mohammadi A, Hatef B. Support vector machine classification of brain states exposed to social stress test using EEG-based brain network measures. Biocybernetics and Biomedical Engineering. 2019; 39:199–213. https://doi.org/10.1016/j.bbe.2018.10.008
37. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics. 2002; 35:352–9. https://doi.org/10.1016/s1532-0464(03)00034-0 PMID: 12968784
38. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data. PLoS One. 2017; 12:e0174944. https://doi.org/10.1371/journal.pone.0174944 PMID: 28376093.
39. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Netw Open. 2020; 3:e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962 PMID: 31922560.
40. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005; 34:113–27. https://doi.org/10.1016/j.artmed.2004.07.002 PMID: 15894176.
41. Slavich GM, Shields GS. Assessing Lifetime Stress Exposure Using the Stress and Adversity Inventory for Adults (Adult STRAIN): An Overview and Initial Validation. Psychosom Med. 2018; 80:17–27. https://doi.org/10.1097/PSY.0000000000000534 PMID: 29016550.
42. Jovic A, Brkic K, Bogunovic N. A review of feature selection methods with applications. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE; 5/25/2015–5/29/2015. pp. 1200–5.

43. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019; 14:e0224365. Epub 2019/11/07. https://doi.org/10.1371/journal.pone.0224365 PMID: 31697686.

9.2 Publication 2

Prediction of Chronic Stress and Protective Factors in Adults:

Development of an Interpretable Prediction Model Based on XGBoost and SHAP Using National Cross-sectional DEGS1 Data

Arezoo Bozorgmehr, Birgitta Weltermann

2023

DOI: https://ai.jmir.org/2023/1/e41868

Original Paper

Prediction of Chronic Stress and Protective Factors in Adults: Development of an Interpretable Prediction Model Based on XGBoost and SHAP Using National Cross-sectional DEGS1 Data

Arezoo Bozorgmehr, MSc; Birgitta Weltermann, MD, MPH

Institute of General Practice and Family Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany

Corresponding Author:
Arezoo Bozorgmehr, MSc
Institute of General Practice and Family Medicine
University Hospital Bonn
University of Bonn
Venusberg-Campus1
Bonn, 53127
Germany
Phone: 49 228 287 11160
Email: arezoo.bozorgmehr@ukbonn.de

Abstract

Background: Chronic stress is highly prevalent in the German population. It has known adverse effects on mental health, such as burnout and depression. Known long-term effects of chronic stress are cardiovascular disease, diabetes, and cancer.

Objective: This study aims to derive an interpretable multiclass machine learning model for predicting chronic stress levels and factors protecting against chronic stress based on representative nationwide data from the German Health Interview and Examination Survey for Adults, which is part of the national health monitoring program.

Methods: A data set from the German Health Interview and Examination Survey for Adults study including demographic, clinical, and laboratory data from 5801 participants was analyzed. A multiclass eXtreme Gradient Boosting (XGBoost) model was constructed to classify participants into 3 categories including low, middle, and high chronic stress levels. The model's performance was evaluated using the area under the receiver operating characteristic curve, precision, recall, specificity, and the F_1-score. Additionally, SHapley Additive exPlanations was used to interpret the prediction XGBoost model and to identify factors protecting against chronic stress.

Results: The multiclass XGBoost model exhibited the macroaverage scores, with an area under the receiver operating characteristic curve of 81%, precision of 63%, recall of 52%, specificity of 78%, and F_1-score of 54%. The most important features for low-level chronic stress were male gender, very good general health, high satisfaction with living space, and strong social support.

Conclusions: This study presents a multiclass interpretable prediction model for chronic stress in adults in Germany. The explainable artificial intelligence technique SHapley Additive exPlanations identified relevant protective factors for chronic stress, which need to be considered when developing interventions to reduce chronic stress.

(JMIR AI 2023;2:e41868) doi: 10.2196/41868

KEYWORDS

artificial intelligence; machine learning; prognostic; model; chronic stress; resilience factors; interpretable model; explainability; stress; disease; diabetes; cancer; dataset; clinical; data; gender; social support; support; intervention; SHAP

Introduction

Chronic stress has many negative effects, primarily on mental health, for example burnout and depression [1]. Long-term chronic stress is associated with various illnesses including cardiovascular disease, diabetes, cancer, and asthma [2-5]. High chronic stress is prevalent with multiple mental health problems in the German population, and this value has increased to 61.1% [6]. However, the vast majority of the population does not develop high chronic stress. While most research has focused on the development of pathology and risk factors, it is paramount to better understand protective factors that prevent chronic stress. In our prior study [7] with 764 participants including general

practitioners (GPs) and practice assistants (PrAs) from 136 German general practices, we analyzed the level of strain due to stress stratified for personal, practice, and regional characteristics. We showed that GPs and PrAs, who individually applied more than 5 measures regularly to compensate for stress, had markedly lower stress levels as measured by the Screening Scale of the Trier Inventory for the Assessment of Chronic Stress (TICS-SSCS) instrument [8].

The psychological construct of resilience, developed over the last decades, addresses this perspective. The American Psychological Association (in 2014) defines resilience as "the process of adapting well in the face of adversity, trauma, tragedy, threats or even significant sources of stress" [9]. Resilience in the context of chronic stress has been characterized by the ability to "bounce back from negative emotional experiences and by flexible adaptation to the changing demands of stressful experiences" [10]. It involves the ability to maintain healthy functioning in different domains of life, such as work and family. Holz et al [11] provided an overview of the current literature investigating the neural mechanisms of resilience focusing on social background. They discussed possible prevention and early intervention approaches targeting the individual and the social environment to lower the risk of psychiatric disorders and to foster resilience [11]. Schetter et al [12] reviewed the traditions of research and definitions of resilience to chronic stress in adults and gained an understanding of resilience in general. They developed a taxonomy of resilience resources to guide future research [12]. Other studies focused on neurobiological cascades involving, for example, enkephalins and associated opioid receptors, μ-opioid peptide receptor, and δ-opioid peptide receptor, to better understand the biological mechanisms of natural adaptation. Prospectively, this bares the potential for effective preventive or therapeutic strategies [13].

To better understand the chronic stress in epidemiological studies, machine learning (ML) offers new approaches to evaluate and model complex relationships in data [14,15]. ML strategies are based on algorithms, which describe the relationships between variables. Two areas in medicine that benefit from ML techniques are diagnosis and outcome prediction [16,17]. Focusing on chronic stress prediction, our prior study [18] compared 4 supervised ML classifiers and 1 standard approach based on data of 550 PrAs from 136 German general practices. We showed that all 4 ML approaches, especially random forest, provided more accurate models for predicting chronic stress than standard regression analysis [18].

Aiming at an interpretable multiclass ML model for predicting chronic stress, we developed an eXtreme Gradient Boosting (XGBoost) model based on nationally representative German Health Interview and Examination Survey for Adults (DEGS1) data. The unified framework SHAP (SHapley Additive exPlanations) is used to interpret the prediction model and to identify factors protecting against chronic stress.

Methods

Overview

This study used nationally representative data from the DEGS1 study, which is a part of the health monitoring program of the Robert Koch Institute, Berlin, Germany. It was conducted from 2008 to 2011 by means of interviews, examinations, and tests among the German population aged 18-79 years (n=8151). The DEGS1 data set, which is available for public use on request, included measurements for chronic stress among 5801 respondents aged 18 to 64 years [6,19].

Primary Outcome

Chronic stress was assessed using the 12-item German short version of TICS-SSCS (n=5850) [6]. It was developed by Schultz et al [8] based on the systemic-requirement-resource model of health [8,20]. The 12-item scale addresses 5 stress areas: chronic worrying, work overload, social overload, excessive demands of work, and lack of social recognition. Its internal consistency showed a Cronbach α of .91 and a good to very good reliability with values ranging from .84 to .91 (mean α=.87) [8]. All 12 questionnaire items use a 5-point Likert scale answer format (0="never" to 4="very often") to measure chronic stress in the past 3 months [21,22]. A sum score (scale 0-48) was calculated for each participant, which is categorized in 3 classes based on a reference population with the TICS-SSCS: 1-11 (≤median)=low stress, 12-22=middle stress, and >22=high stress (≥90th percentile). This multiclass outcome is the recommended DEGS1 approach [6].

Predictors

In addition, the DEGS1 data set included variables on sociodemographic characteristics, chronic diseases (eg, coronary heart disease, stroke, diabetes mellitus, depression, and anxiety disorder), living conditions, health-related behavior, preventive measures, and general health. Based on a literature review and using the Powershap feature selection method, 34 features were included in this analysis. Table 1 depicts descriptive information about the variables used.

Table 1. Demographic, clinical, and workplace characteristics of the German Health Interview and Examination Survey for Adults study participants (N=5801).

Demographic characteristics	Values
Continuous variables, mean (SD; range)	
Age (years)	42 (13.11; 18-64)
Number of persons in the household	3 (1.34; 1-11)
Sleep hours per night in the past 4 weeks	7 (1.19; 2-12)
Number of hospital nights in the past 12 months	1 (5.30; 0-150)
Number of sick days in the past 12 months	13 (38.01; 0-365)
Categorical variables	
Gender (female), n (%)	3081 (49.6)
Marital status, n (%)	
Married living with partner or separately from partner	3697 (59.5)
Single	1957 (31.5)
Divorced	376 (6.1)
Widowed	136 (2.2)
Provides care to someone in need or seriously ill, n (%)	379 (6.1)
Renting or living in own apartment/house, n (%)	
Rented apartment or house	2689 (43.3)
Own apartment or house	3268 (52.6)
Satisfaction with living space, n (%)	
Very satisfied or satisfied	5269 (84.8)
Neither satisfied nor dissatisfied	608 (9.8)
Dissatisfied or very dissatisfied	295 (4.8)
Residential area satisfaction, n (%)	
Very satisfied or satisfied	5091 (81.9)
Neither satisfied nor dissatisfied	727 (11.7)
Dissatisfied or very dissatisfied	320 (5.2)
General state of health, n (%)	
Very good or good	4942 (79.5)
Average	1134 (18.3)
Poor or very poor	116 (1.8)
Intake of sleeping pills in the past 4 weeks, n (%)	
Never	5919 (95.3)
Less than 1 time	100 (1.6)
1 time or 2 times	73 (1.2)
3 times or more	86 (1.4)
Social support, n (%)	
Low support	653 (10.5)
Average support	3082 (49.6)
Strong support	2451 (39.5)
Health behavior consultation in the past 12 months, n (%)	1873 (30.4)
Has general practitioner	5497 (88.5)
Visited to general practitioner in the past 12 months	4870 (78.4)

Demographic characteristics	Values
Visited to neurologist in the past 12 months	463 (7.5)
Frequency of alcohol consumption, n (%)	
Never	744 (12.0)
1 time per month or less	1186 (19.1)
2-4 times per month	1998 (32.2)
2-3 times per week	1453 (23.4)
4 times per week or more	811 (13.1)
Tobacco use, n (%)	
Yes, daily	1701 (27.4)
Yes, occasionally	433 (7)
Not anymore	1664 (26.8)
Never smoked	2400 (38.7)
Comorbidities, n (%)	
Has hypertension	1625 (26.2)
Has diabetes	271 (4.4)
Has migraine	712 (11.5)
Has depression	682 (11)
Has anxiety disorder	327 (5.3)
Has burnout syndrome	292 (4.7)
Has one or more long-term chronic diseases	1418 (22.8)
Prevention programs or sport activities, n (%)	
Participated in prevention program in the past 12 months	988 (15.9)
Participated in relaxation or stress management program	188 (3)
Participated in gymnastics, fitness, or balance sports program	832 (13.4)
Participated in alcohol cessation program	7 (0.1)
Participated in smoking cessation program	17 (0.3)
Participated in weight reduction or a healthy diet program	167 (2.7)
Sports activities per week (in the past 3 months), n (%)	
No sports activity	1954 (31.5)
Up to 2 hours per week	2584 (41.6)
Regularly, 2-4 hours per week	990 (15.9)
Regularly, more than 4 hours per week	645 (10.4)

Data Preprocessing

Data Normalization

The DEGS1 study features include both discrete and continuous values. When these features are combined, the range of the values differs. Therefore, the training data set was normalized using the min-max normalization method. This normalization technique accurately preserves all relationships in the data, thereby avoiding the introduction of bias [23].

Handling of Missing Data

For single features, missing values were low (<2%), yielding an overall missing rate of 13.91% in our data set. We used the K-Nearest Neighbors (KNN) approach to impute the missing variables. This method identifies the KNNs on the Euclidean distance. Missing values were replaced using a majority vote for discrete variables and weighted means for continuous features. All features are imputed simultaneously without the need to treat features individually [24].

Addressing the Imbalanced Data Set

For chronic stress, the distribution of classes was unequal (class 0: 52%, class 1: 38%, and class 2: 11%). This imbalanced multiclass classification was addressed using the Synthetic Minority Oversampling TEchnique to increase the frequency of near-miss data points within the training data set. This oversampling method randomly generated new instances of

minority class to balance the number of classes without any additional information to the model [25].

Feature Selection

We used Powershap as a wrapper-based Shaply feature selection method. This technique is based on the core assumption that an informative feature will have a larger impact on the prediction compared to a known random feature [26].

Machine Learning Approach: XGBoost

Overview

To predict chronic stress levels and detect factors protecting against chronic stress, we applied the decision tree–based ensemble ML technique, XGBoost [27,28]. XGBoost is a scalable and accurate implementation gradient boosting machine developed by the Distributed Machine Learning Community in the form of open-source libraries. It combines a recursive gradient boosting method called Newton boosting. Based on a decision tree model, it efficiently provides accurate predictions because each tree is boosted recursively and in parallel.

The ML technique generally aims to identify a relationship between the input $\boldsymbol{X}=\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots \boldsymbol{x_n}\}$ and the output **Y**. For a given data set with **n** samples and **m** features, **K** additive functions are used in the XGBoost model to predict the output through the following estimation (equation 1) [27]:

$$\hat{y}_j = \sum_{k=1}^{K} f_k(x_i) \quad (1)$$

where $\mathbf{f_k}$ $\{\mathbf{f(x)} = \omega_{\mathbf{q}}\}$ $(\mathbf{q}: R^{\mathbf{m}} \rightarrow \mathbf{T}, \omega\ R^{\mathbf{T}})$ is the regression tree's space, and **q** denotes the independent structure of each tree with **T** leaves. Each $\mathbf{f_k}$ corresponds to an independent tree structure **q** and leaf weights ω**.** The following regularized objective is minimized to learn the set of functions (equation 2).

$$\mathbf{L} = \sum_{i} \boldsymbol{l}(\hat{y}_j, \boldsymbol{y}_j) + \sum_{k} \boldsymbol{\Omega}\,(\boldsymbol{f}_k) \quad (2)$$

where $\Omega\ (\mathbf{f}) = \gamma\mathbf{T} + ½\ \lambda\ \|\omega\|^2$, **I** represents the model loss function, and Ω denotes the regularized term.

Hyperparameter Tuning

In this study, a grid-search approach from scikit-learn class "GridSearchCV" was applied toward the optimal tuning of XGBoost hyperparameters. The number of estimators was set to 1000 to represent the maximum number of trees created during the training phase. The Softmax function is used to convert logits of the XGBoost classifier into a probability distribution. Each element of the output lies in the interval (0,1) and the output elements sum up to 1. Table 2 summarizes the hyperparameters´ values used to the XGBoost model (see Multimedia Appendix 1).

Table 2. Main hyperparameters for the Extreme Gradient Boosting model.

Hyperparameter	Value
learning rate	0.3
Estimators, n	1000
max_depth	5
Subsample	0.8
min_child_weight	3
L2 regularization term (Lambda)	2
colsample-bytree	0.7
Objective	multi:softmax

K-Fold Cross-Validation

After preprocessing, the 34 features were fed into ML classifiers to train the model for classification. The data set was split into a "training" and a "validation" data set. We used the repeated K-fold cross-validation approach, repeating the mean performance across all folds and all repeats to reduce the bias in the model's estimated performance with K=10. K=10 was chosen as the optimal number of folds, which optimizes the time to complete the test while minimizing the bias and variance associated with the validation process.

Model Performance Evaluation

To evaluate the method proposed in this study, we used the following most promising multiclass evaluation metrics: the area under the receiver operating characteristic curve (AUC), precision, recall, and F_1-score. Multiclass classification works on data sets in which all classes are mutually exclusive. In a multiclass classifier, the evaluation measures of individual classes are averaged out to determine the performance on overall system across the data. We applied the macroaverage approach [29].

The receiver operating characteristic (ROC) curve was used to evaluate the performance of the classifier. For different classification thresholds, the macro true-positive rate (equation 3) is plotted against the macro false-positive rate (equation 4). The AUC indicates the classifier's ability to distinguish between classes. The value of the AUC is in the range (0,1), in which 1 is for a perfect classifier. In this study, the ROC curve is plotted for each class broken down into a series of binary problems using the One-vs-Rest approach. The macroaverage is computed by summing the individual values for true positive, true negative, false positive, and false negative. Then, macroaverage scores

of true positive instances (precision; equation 5), true positive rate (recall; equation 6), true negative rate (specificity; equation 7), and the harmonic mean of the precision and recall computed on each class (F_1-score; equation 8) were computed. Mathematically, they are defined as follows:

$$TPR_{macro} = \frac{\sum_{i=1}^{k} TPR_i}{k} \quad (3)$$

$$FPR_{macro} = \frac{\sum_{i=1}^{k} FPR_i}{k} \quad (4)$$

$$Precision = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \quad (5)$$

$$Recall = \frac{1}{n}\sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \quad (6)$$

$$Specificity = \frac{1}{n}\sum_{i=1}^{n} \frac{TN_i}{TN_i + FP_i} \quad (7)$$

$$F_1 - score = \frac{1}{n}\sum_{i=1}^{n} \frac{2\ Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (8)$$

We used Python 3.7 (Python Software Foundation) to implement our ML framework. In addition, several libraries from the python data science ecosystem were used to execute the experiments and the integrated development environment PyCharm. To implement the Powershap feature selection method, we used the Powershap Python library. The scikit-learn package (version 1.0.2) was used to train and evaluate the ML classifier. SHAP tool (version 0.40.0) was used to assess the explainability the model; that is, to identify factors protecting against chronic stress.

In addition to the performance evaluation, this study maximizes the interpretability of the underlying models. It focuses particularly on the explainability of the model, which can serve as an indispensable tool in the era of precision medicine.

Model Interpretation: SHAP

Per our understanding, the interpretation of the prediction models is as crucial as the prediction accuracy because it extracts information that significantly affects outcomes and identifies the factors protecting against chronic stress from subjects with lower chronic stress. However, the ensemble learning method XGBoost represents a black-box model. To overcome this problem, Lundberg [30,31] proposes the SHAP approach for interpreting predictions of complex models created by different techniques; for example, NGBoost, CatBoost, XGBoost, LightGBM, and scikit-learn tree models. SHAP was initially developed by Shapley in 1953 and is based on the game theory [32]. It explains the prediction of a specific input (**X**) by calculating the impact of each feature on the prediction. The estimated Shapley values are calculated as follows (equation 9):

$$\hat{\phi}_j = \frac{1}{K}\sum_{k=1}^{K}((\hat{g}\left(x_{+j}^{m}\right) - \hat{g}\left(x_{-j}^{m}\right)) \quad (9)$$

where $\hat{g}(x_{-j}^{m})$ is the prediction for *x*, but with a random number of feature values. TreeSHAP is used for gradient boosting models including XGBoost. It offers a rich visualization of each feature attribution and allows for partial dependence plots.

The TreeSHAP interaction values estimates as follows (equation 10):

$$\varphi_i = \sum_{S \subseteq N\{i,j\}} \frac{|S|!\,(M - |S| - 2i)}{2(M-1)!}\delta_{ij}(S) \quad (10)$$

where $i \neq j$, $\delta_{ij}(S) = f_x(S \cup \{I,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S)$, *M* is the number of features, and *S* denotes all feature subsets. SHAP values advance the understanding of tree models by including feature importance, feature dependence plots, local explanations, and summary plots [30].

Ethical Considerations

Ethics approval for the DEGS1 survey was obtained from the Charité – Universitätsmedizin Berlin Ethics Committee (EA2/047/08). All participants received written information and provided informed consent before the interview and examination. The analysis described here builds on a data set from the DEGS1 study, which was kindly provided by the Robert Koch Institute. This secondary analysis of anonymized data does not require a separate ethics vote.

Results

Characteristics of the DEGS1 Study Population

The mean age of the 5801 DEGS1 study participants was 44 years, with more than half of the population being female (n=3080, 53.1%). The mean stress level of the total population was 12.00 (95% CI 11.79-12.20): 11% (n=625) of the participants had “high chronic stress” (category 2), while 38% (n=2188) had “middle” (category 1), and 52% (n=2988) of them had “low chronic stress” (category 0). Most participants reported their general state of health as very good or good (79.3%, n=4599). Table 1 shows the weighted demographic, clinical, and laboratory characteristics of the participants.

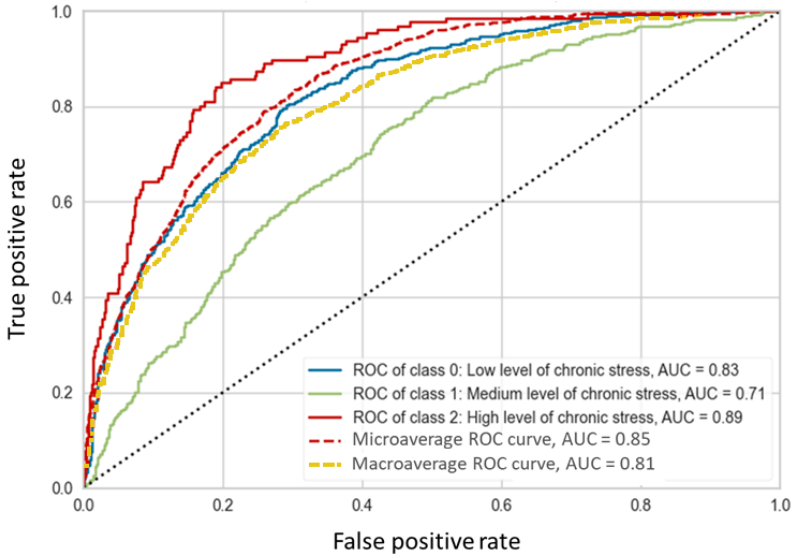
Results of the Machine Learning Analysis

The evaluation metrics of the XGBoost model’s performance are presented in Table 3 differentiated by chronic stress classes. We see that the XGBoost model achieved the highest AUC score for class 2 with 0.89% and a good macroaverage AUC score of 81% for the overall model. The metrics for the 3 stress classes and the average results are reported in Table 3. The ROC curves for the multiclass chronic stress prediction of the XGBoost model are shown in Figure 1.

Table 3. Classification metrics: area under the receiver operating characteristic curve (AUC), precision, recall, specificity, and F_1-score for XGBoost.

Measure	XGBoost			
	Class 0	Class 1	Class 2	Macroaverage
AUC	0.83	0.71	0.89	0.81
Precision	0.73	0.56	0.58	0.63
Recall	0.80	0.55	0.37	0.52
Specificity	0.90	0.38	0.26	0.78
F_1-score	0.76	0.60	0.45	0.54

Figure 1. ROC curves for 3 classes using the XGBoost multiclass classifier. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic curve.



Explanation of the Behavior of Individual Features

The result of the SHAP analysis is displayed in Figure 2. In this plot, the impact of a feature on the respective classes (stress classes 0-2) is stacked to illustrate the feature importance. This means that the features with large absolute Shapley values are more important than those with lower values. The plot shows that class 0 (low level of chronic stress) hardly uses the features gender, general state of health, satisfaction with living space, and social support. Class 2 as the high level of chronic stress uses the features number of sick days in the past 12 months, social support, sleeping hours per night in the past 4 weeks, gender, and general state of health. Interestingly, classes 0 and 2 use many identical features.

While the SHAP feature plot provides an overview of the role of each variable irrespective of the direction of these effects, the SHAP summary plot provides such additional information for classes. The impact distribution of each feature on the model output for classes with low and high levels of chronic stress is shown in Figures 3 and 4. Each row in this plot represents a single feature in order of their mean absolute SHAP values. It can be a negative or positive value and represents the importance of each feature. Each dot is a Shapley value for a particular feature and reflects its impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low). The color is the actual feature value in the data set. For example, the red values for age as a continuous feature represent older people, while blue values represent younger people, and blue values for gender as a categorical feature (low value=1) represent males and red values (high value= 2) represent females. Overlapping points are jittered toward the y-axis, giving a sense of the distribution of the Shapley values per feature.

According to the SHAP summary plot result, gender is the most significant feature for class 0, and the number of sick days in the past 12 months has the highest impact on class 2. We note that the general state of health (shown in red) with high values has negative SHAP values and a relatively negative effect on the model for the low level of chronic stress and a positive impact (positive SHAP values) for class 2. Higher values on the social support scale have a positive impact on class 0 and negative effects on class 2, which means that chronic stress is less likely with strong social support.

Figure 2. SHAP feature plot of the 20 most important features: relative importance of each feature based on the average absolute value of the SHAP values. SHAP: SHapley Additive exPlanations; XGBoost: Extreme Gradient Boosting. *In the past 12 months; **per week.

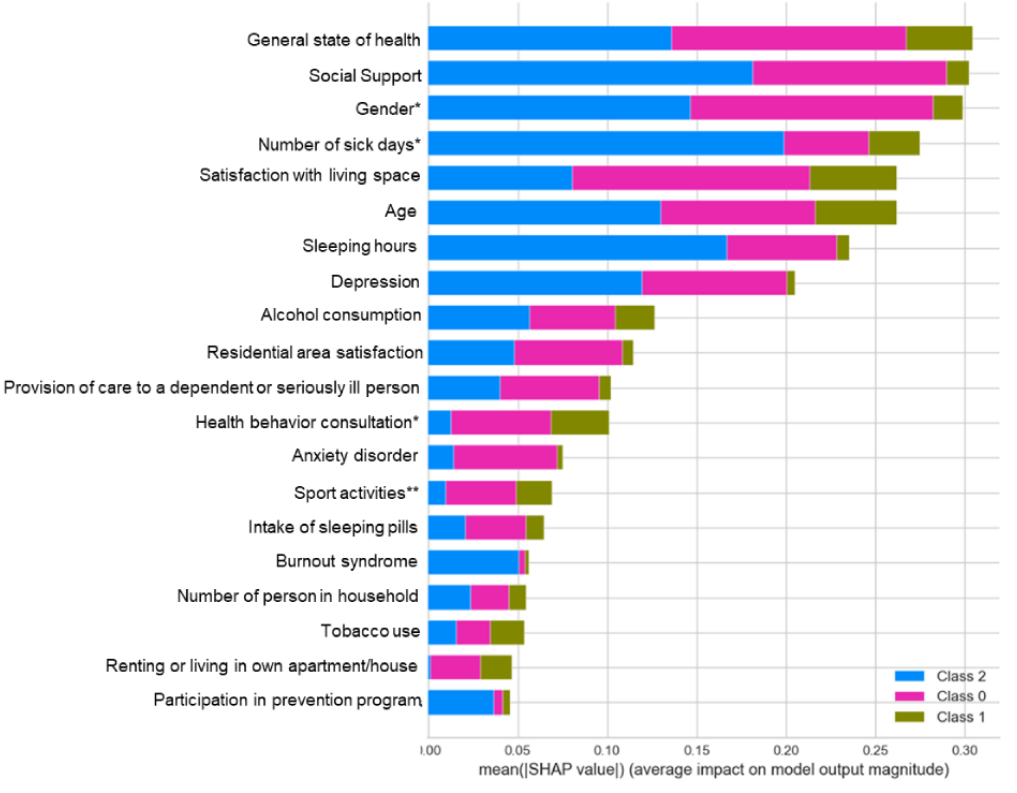


Figure 3. SHAP summary plot. Importance of the representative chronic stress features (top 20) in class 0: each dot is a Shapley value for a particular feature and reflects its impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low). GP: general practitioner; SHAP: SHapley Additive exPlanations. *In the past 12 months; **per week.

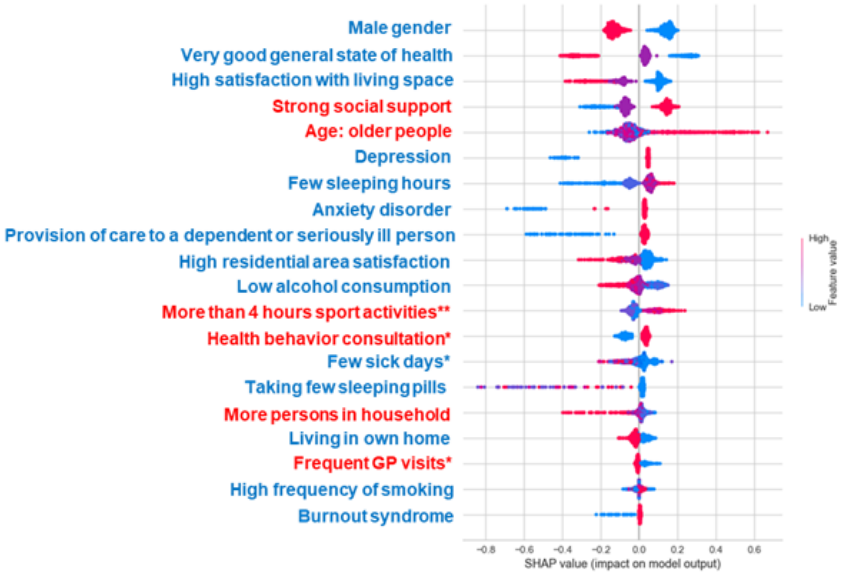
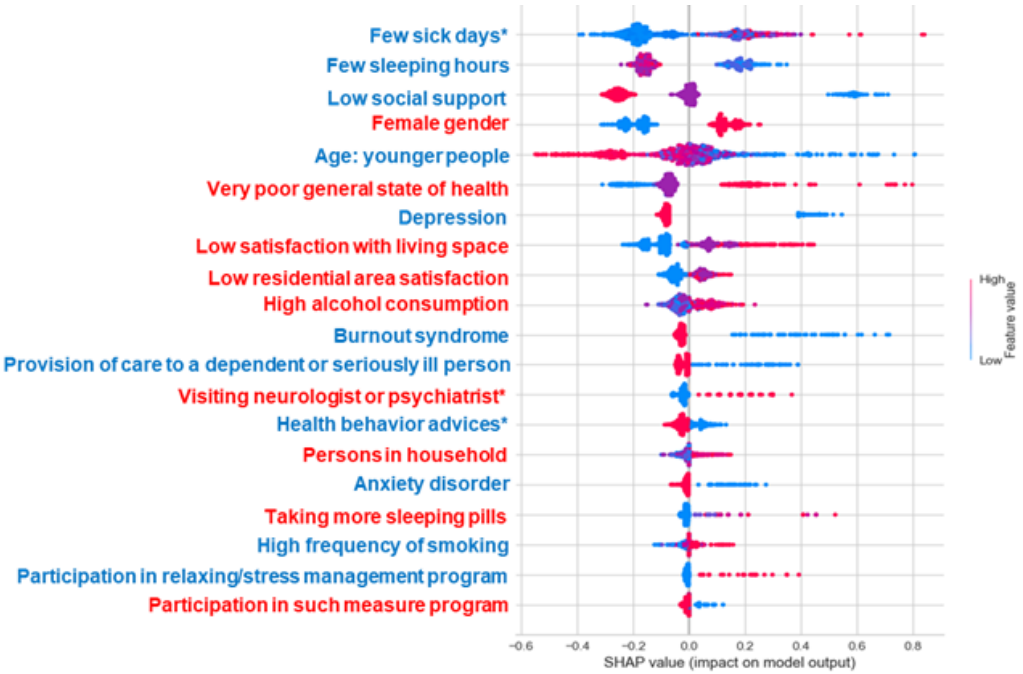


Figure 4. SHAP summary plot. Importance of the representative chronic stress features (top 20) in class 2: each dot is a Shapley value for a particular feature and reflects its impact on a specific class for a given instance, and dots stack up to show density. It is color-coded in accordance with the magnitude to which the value contributes to the model impact (red=high and blue=low). SHAP: SHapley Additive exPlanations. *In the past 12 months.



Discussion

Principal Findings

To our knowledge, this is the first study to select the XGBoost algorithm as an ML multiclass classifier in the prediction of chronic stress as well as the SHAP method to interpret the model's prediction. Based on nationally representative German data, chronic stress was predicted using 34 characteristics of adult participants. We identified male gender, a very good general state of health, high satisfaction with living space, strong social support, enough sleep, and more than 4 hours of sports activities per week as protective factors against chronic stress. These results are in line with those of other studies, which showed that resilience against chronic stress is promoted by social support, family connectedness, and friendship networks in the community [33-36]. For example, with a sample of 24,347 participants from the Canadian General Social Survey, Van der Horst et al [36] determined that good friendship networks are positively associated with less stress, better health, and more social support. A cross-sectional study of 538 nursing students from an Australian university showed that social support positively affect the psychological well-being [37].

Our ML approach allowed for the inclusion of a broad spectrum of individual characteristics, which comprised medical, lifestyle, living space, and social information, while other studies on chronic stress used multivariate models with fewer parameters only. For example, a large cross-sectional study with 34,129 participants from China, Ghana, India, Mexico, Russia, and South Africa showed positive associations of multimorbidity, stroke, depression, and hearing problems with perceived stress without assessing potential protective factors such as living space and social support [38]. A US cross-sectional telephone survey with 340,847 participants aged between 18 and 85 years documented that psychological well-being, especially stress, improved, but integrated only 5 parameters such as gender, employment status, partnership, and underage children in the household in their model analyzed [39]. In a study with 12,110 working adults from Minnesota, United States, a high level of perceived stress was associated with a higher-fat diet, less exercising, and being a smoker using a multivariate model with 6 variable topics but did not include medical and living circumstances [40].

Strengths and Limitations

This study used the population-based, representative DEGS1 data set, which implies a low risk of selection bias; yet, the results may not be transferrable to other settings. The DEGS1 data, which were collected from 2008 to 2011, may not fully describe current living conditions in Germany, especially the potential effects of the pandemic, which were shown in other studies, were not measured [41]. In our study, the SHAP methodology allowed for a detailed visualization of single feature attributions, which improved the understanding of the ML model.

Conclusions

In this study, we developed an XGBoost ML model to predict chronic stress in adults. The SHAP methodology identified various relevant factors protecting against chronic stress, which need to be considered when developing interventions for stress reduction and improving resilience.

Acknowledgments

We owe special thanks to the Robert Koch Institute, Berlin, Germany, for kindly providing the data set and additional information on the DEGS1 survey.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Hyperparameter Tuning for XGBoost.
[PDF File (Adobe PDF File), 530 KB-Multimedia Appendix 1]

References

1. Marin M, Lord C, Andrews J, Juster R, Sindi S, Arsenault-Lapierre G, et al. Chronic stress, cognitive functioning and mental health. Neurobiol Learn Mem 2011 Nov;96(4):583-595. [doi: 10.1016/j.nlm.2011.02.016] [Medline: 21376129]
2. Cohen S, Janicki-Deverts D, Miller GE. Psychological stress and disease. JAMA 2007 Oct 10;298(14):1685-1687. [doi: 10.1001/jama.298.14.1685] [Medline: 17925521]
3. Kivimäki M, Steptoe A. Effects of stress on the development and progression of cardiovascular disease. Nat Rev Cardiol 2018 Apr 7;15(4):215-229. [doi: 10.1038/nrcardio.2017.189] [Medline: 29213140]
4. Marcovecchio ML, Chiarelli F. The effects of acute and chronic stress on diabetes control. Sci Signal 2012 Oct 23;5(247):pt10. [doi: 10.1126/scisignal.2003508] [Medline: 23092890]
5. Landeo-Gutierrez J, Celedón JC. Chronic stress and asthma in adolescents. Ann Allergy Asthma Immunol 2020 Oct;125(4):393-398 [FREE Full text] [doi: 10.1016/j.anai.2020.07.001] [Medline: 32653405]
6. Hapke U, Maske UE, Scheidt-Nave C, Bode L, Schlack R, Busch MA. [Chronic stress among adults in Germany: results of the German Health Interview and Examination Survey for Adults (DEGS1)]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2013 May;56(5-6):749-754. [doi: 10.1007/s00103-013-1690-9] [Medline: 23703494]
7. Viehmann A, Kersting C, Thielmann A, Weltermann B. Prevalence of chronic stress in general practitioners and practice assistants: personal, practice and regional characteristics. PLoS One 2017;12(5):e0176658 [FREE Full text] [doi: 10.1371/journal.pone.0176658] [Medline: 28489939]
8. Schulz P, Schlotz W, Becker P. Trierer Inventar zum Chronischen Stress (TICS) Trier Inventory for Chronic Stress (TICS). Newburyport, MA: Hogrefe Publishing Corp; 2004.
9. Southwick SM, Bonanno GA, Masten AS, Panter-Brick C, Yehuda R. Resilience definitions, theory, and challenges: interdisciplinary perspectives. Eur J Psychotraumatol 2014;5 [FREE Full text] [doi: 10.3402/ejpt.v5.25338] [Medline: 25317257]
10. Tugade MM, Fredrickson BL. Resilient individuals use positive emotions to bounce back from negative emotional experiences. J Pers Soc Psychol 2004 Feb;86(2):320-333 [FREE Full text] [doi: 10.1037/0022-3514.86.2.320] [Medline: 14769087]
11. Holz NE, Tost H, Meyer-Lindenberg A. Resilience and the brain: a key role for regulatory circuits linked to social stress and support. Mol Psychiatry 2019 Oct 18;25(2):379-396. [doi: 10.1038/s41380-019-0551-9]
12. Schetter CD, Dolbier C. Resilience in the context of chronic stress and health in adults. Soc Personal Psychol Compass 2011 Sep;5(9):634-652 [FREE Full text] [doi: 10.1111/j.1751-9004.2011.00379.x] [Medline: 26161137]
13. Henry MS, Gendron L, Tremblay M, Drolet G. Enkephalins: endogenous analgesics with an emerging role in stress resilience. Neural Plast 2017;2017:1546125 [FREE Full text] [doi: 10.1155/2017/1546125] [Medline: 28781901]
14. Alpaydin E. Machine learning. Cambridge, MA: The MIT Press; 2021.
15. Bonaccorso G. Machine learning algorithms (first edition). Birmingham: Packt Publishing Limited; 2017.
16. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019 Mar 19;19(1):64 [FREE Full text] [doi: 10.1186/s12874-019-0681-4] [Medline: 30890124]
17. Deo RC. Machine learning in medicine. Circulation 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
18. Bozorgmehr A, Thielmann A, Weltermann B. Chronic stress in practice assistants: An analytic approach comparing four machine learning classifiers with a standard logistic regression model. PLoS One 2021;16(5):e0250842 [FREE Full text] [doi: 10.1371/journal.pone.0250842] [Medline: 33945572]
19. Gößwald A, Lange M, Dölle R, Hölling H. [The first wave of the German Health Interview and Examination Survey for Adults (DEGS1): participant recruitment, fieldwork, and quality management]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2013 May 25;56(5-6):611-619. [doi: 10.1007/s00103-013-1671-z] [Medline: 23703477]
20. Beise U. Prävention und Gesundheitsförderung. In: Beise U, Heimes S, Schwarz W, editors. Gesundheits- und Krankheitslehre. Berlin: Springer; 2013:27-34.

21. Petrowski K, Paul S, Albani C, Brähler E. Factor structure and psychometric properties of the trier inventory for chronic stress (TICS) in a representative German sample. BMC Med Res Methodol 2012 Apr 01;12:42 [FREE Full text] [doi: 10.1186/1471-2288-12-42] [Medline: 22463771]
22. Schulz P, Schlotz W. Trierer Inventar zur Erfassung von chronischem Streß (TICS): Skalenkonstruktion, teststatistische Überprüfung und Validierung der Skala Arbeitsüberlastung. Diagnostica 1999 Jan;45(1):8-19. [doi: 10.1026//0012-1924.45.1.8]
23. Borkin D, Némethová A, Micha??onok G, Maiorov K. Impact of data normalization on classification model accuracy. Research Papers Faculty of Materials Science and Technology Slovak University of Technology 2019;27(45):79-84 [FREE Full text] [doi: 10.2478/rput-2019-0029]
24. Murti DMP, Pujianto U, Wibawa AP, Akbar MI. K-nearest neighbor (K-NN) based missing data imputation. 2019 Presented at: 5th International Conference on Science in Information Technology (ICSITech); October 23-24, 2019; Yogyakarta, Indonesia. [doi: 10.1109/icsitech46713.2019.8987530]
25. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences 2019 Dec;505:32-64. [doi: 10.1016/j.ins.2019.07.070]
26. Verhaeghe J, Van Der Donckt J, Ongenae F, Van Hoecke S. Powershap: a power-full Shapley feature selection method. arXiv. Preprint posted online June 16, 2022 2022 [FREE Full text] [doi: 10.1007/978-3-031-26387-3_5]
27. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
28. Che D, Liu Q, Rasheed K, Tao X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. Adv Exp Med Biol 2011;696:191-199. [doi: 10.1007/978-1-4419-7046-6_19] [Medline: 21431559]
29. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv. Preprint posted online August 13, 2020 2020 [FREE Full text]
30. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online May 22, 2017 2017 [FREE Full text]
31. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
32. Winter E. The Shapley value. In: Handbook of Game Theory with Economic Applications. Amsterdam: Elsevier; 2002:2025-2054.
33. Taylor SE. Social support: a review. In: Friedman HS, editor. The Oxford handbook of health psychology. Oxford: Oxford University Press; 2011:189-214.
34. Lepore SJ, Evans GW, Schneider ML. Dynamic role of social support in the link between chronic stress and psychological distress. J Pers Soc Psychol 1991 Dec;61(6):899-909. [doi: 10.1037//0022-3514.61.6.899] [Medline: 1774628]
35. Thomas PA, Liu H, Umberson D. Family relationships and well-being. Innov Aging 2017 Nov;1(3):igx025 [FREE Full text] [doi: 10.1093/geroni/igx025] [Medline: 29795792]
36. van der Horst M, Coffé H. How friendship network characteristics influence subjective well-being. Soc Indic Res 2012 Jul;107(3):509-529 [FREE Full text] [doi: 10.1007/s11205-011-9861-2] [Medline: 22707845]
37. He FX, Turnbull B, Kirshbaum MN, Phillips B, Klainin-Yobas P. Assessing stress, protective factors and psychological well-being among undergraduate nursing students. Nurse Educ Today 2018 Sep;68:4-12. [doi: 10.1016/j.nedt.2018.05.013] [Medline: 29870871]
38. Stubbs B, Vancampfort D, Veronese N, Schofield P, Lin P, Tseng P, et al. Multimorbidity and perceived stress: a population-based cross-sectional study among older adults across six low- and middle-income countries. Maturitas 2018 Jan;107:84-91 [FREE Full text] [doi: 10.1016/j.maturitas.2017.10.007] [Medline: 29169587]
39. Stone AA, Schwartz JE, Broderick JE, Deaton A. A snapshot of the age distribution of psychological well-being in the United States. Proc Natl Acad Sci U S A 2010 Jun 01;107(22):9985-9990 [FREE Full text] [doi: 10.1073/pnas.1003744107] [Medline: 20479218]
40. Ng DM, Jeffery RW. Relationships between perceived stress and health behaviors in a sample of working adults. Health Psychol 2003 Nov;22(6):638-642. [doi: 10.1037/0278-6133.22.6.638] [Medline: 14640862]
41. Schelhorn I, Ecker A, Lüdtke MN, Rehm S, Tran T, Bereznai JL, et al. Psychological burden during the COVID-19 pandemic in Germany. Front Psychol 2021;12:640518 [FREE Full text] [doi: 10.3389/fpsyg.2021.640518] [Medline: 34557124]

Abbreviations

AUC: area under the receiver operating characteristic curve
DEGS1: German Health Interview and Examination Survey for Adults
GP: general practitioner
KNN: K-nearest neighbors
ML: machine learning

PrA: practice assistant
ROC: receiver operating characteristic
SHAP: SHapley Additive exPlanations
TICS-SSCS: Screening Scale of the Trier Inventory for the Assessment of Chronic Stress
XGBoost: Extreme Gradient Boosting

Edited by K El Emam; submitted 12.08.22; peer-reviewed by W Klement, J Li; comments to author 14.11.22; revised version received 06.01.23; accepted 03.03.23; published 16.05.23

Please cite as:
Bozorgmehr A, Weltermann B
Prediction of Chronic Stress and Protective Factors in Adults: Development of an Interpretable Prediction Model Based on XGBoost and SHAP Using National Cross-sectional DEGS1 Data
JMIR AI 2023;2:e41868
URL: https://ai.jmir.org/2023/1/e41868
doi: 10.2196/41868
PMID:

10 Acknowledgements

I would like to express my sincere gratitude to my parents for their unwavering support throughout my academic journey. Their love, encouragement, and belief in me have been instrumental in my success.

I would also like to extend my heartfelt appreciation to my first supervisor, Professor Dr. Med. Birgitta Weltermann, for her guidance, and expertise. Her mentorship and dedication to my research have been invaluable in shaping the trajectory of my dissertation.

I am deeply grateful to all the participants who generously shared their time and perspectives for the studies conducted in this dissertation.

11 Publications

2023 **Bozorgmehr, A.,** & Weltermann, B. (2023). Prediction of Chronic Stress and Protective Factors in Adults: Development of an Interpretable Prediction Model Based on XGBoost and SHAP Using National Cross-sectional DEGS1 Data. JMIR AI, 2, e41868.

2021 **Bozorgmehr, A.,** Thielmann, A., & Weltermann, B. (2021). Chronic stress in practice assistants: An analytic approach comparing four machine learning classifiers with a standard logistic regression model. Plos one, 16(5), e0250842.

2020 Ghofrani, J., Kozegar, E., Divband Sourati, M., **Bozorgmehr, A.,** Chen, H., Naake, M. (2020). Repository for Reusing Artifacts of Artificial Neural Networks (ACM M, New York, NY, USA, Article 4, 7 pages.), DOI: 10.475/123_4.

2019 Ghofrani, J., Kozegar, E**., Bozorgmehr, A.,** Divband Sourati, M. (2019). Reusability in Artificial Neural Networks: An Empirical Study. International Conference on Compute and Data Analysis (SPLC '19: Proceedings of the 23rd International Systems and Software Product Line Conference - ACM), DOI: 10.18420/SE2020_07.

2019 Ghofrani, J., **Bozorgmehr, A.** (2019). Migration to Microservices: Barriers and Solutions. International Conference on Applied Informatics (Springer 2019), DOI: 10.1007/978-3-030-32475-9_20.

2018 Ghofrani, J., **Bozorgmehr, A.,** Panah, A. (2018). A Fast Algorithm Based on Apriori Algorithms to Explore the Set of Repetitive Items of Large Transaction Data. International Conference on Compute and Data Analysis (ICCDA 2018), DOI: 10.1145/3193077.3193089 2017 IEEE International Conference on Knowledge-Based Engineering and Innovation (KBEI).

2017 Ghofrani, J., Mohseni, M**., Bozorgmehr, A.** (2017). A Conceptual Framework for Clone Detection using Machine Learning. IEEE

International Conference on Knowledge-Based Engineering and Innovation (KBEI 2017), DOI: 10.1109/KBEI.2017.8324908.

2017 P. Ivanovic, H. Richter and **A. Bozorgmehr,** "Cloud-Efficient Modelling and Simulation of Magnetic Nano Materials," Simulationswissenschaftliches Zentrum Clausthal-Göttingen, Clausthal/Göttingen, 2017.

Published contributions to congresses

2021 Bozorgmehr A, Filbert A, Lowitsch V, Jezuita J, Buchner D, Stieber Ch, Lehmann L, Weltermann B. Aufbau einer IT-Struktur für das hausärztliche Forschungspraxennetz NRW (HAFO.NRW): Konzeption and Charakteristika. 20. Deutscher Kongress für Versorgungsforschung, Deutsches Netzwerk Versorgungsforschung e. V. 2021. doi: 10.3205/21dkvf019.

2021 Bozorgmehr A, Thiem SK, Stieber Ch, Weltermann B. Aufbau einer IT-Infrastructure for Supporting the North Rheine-Westphalian General Practice Research Network (NRW.GPRN): Results from the Pilot Study. 93rd EGPRN Meeting, Halle - Germany, 14-17 October 2021