Exploring the Application of Machine Learning Algorithms in Stroke Prediction

Qiyuan Dong

AIEN Institute, Shanghai Ocean University & the University of Tasmania, Shanghai, 201306, China
qiyuand@utas.edu.au

Abstract. Strokes significantly affect global public health and economic stability, with over 12.2 million new cases annually and being a leading cause of death worldwide. This essay takes a dataset containing 15000 entries and 22 features as an example to analyze its preprocessing, feature engineering, and algorithm optimization processes. The relevant results indicate that ML has the ability to identify complex risk patterns and provide personalized health interventions, which greatly advances stroke prediction and prevention strategies. In addition, the study evaluated the model using strict indicators such as accuracy, sensitivity, and Receiver Operating Characteristic and Area Under the Curve score to ensure reliable and applicable results. The significance of this research lies in its contribution to personalized medicine, highlighting how ML can be pivotal in developing targeted treatments and preventive measures. By improving early detection and enabling tailored healthcare solutions, the study enhances individual patient care and optimizes resource allocation across public health systems, setting a benchmark for future research in medical Artificial Intelligence applications.

Keywords: Stroke prediction, Machine learning, Personalized medicine

1 Introduction

Strokes significantly affect public health and economies worldwide. Their impact extends beyond immediate health issues, influencing survivors' quality of life, raising mortality rates, and placing an economic burden [1]. The likelihood of experiencing a stroke is high. The World Stroke Organization (WSO) reports over 12.2 million new stroke cases annually [1]. Globally, one in four individuals over the age of 25 will experience a stroke in their lifetime [1]. Strokes are a major cause of death globally. According to the World Health Organization (WHO), strokes account for about 11% of all deaths worldwide, making them the second leading cause of death after heart disease, resulting in approximately 6 million deaths each year [2]. The costs associated with treating and caring for stroke survivors are substantial. In the US, the total cost of strokes was $103.5 billion in 2016 [3]. Therefore, predicting strokes is crucial.

Traditional methods for predicting stroke, such as the Framingham Stroke Risk

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_12

Profile and CHADS2 score, face several limitations. These models often rely on a limited number of variables, employing linear assumptions that fail to capture complex interactions among risk factors or adapt to changes over time [4]. Consequently, their predictive accuracy may be compromised, particularly when applied to diverse populations or updated with new data. Furthermore, traditional methods struggle with scalability and are unable to efficiently process the vast amounts of data available today through modern healthcare systems [5].

Machine learning (ML) algorithms offer substantial advantages over traditional models in predicting stroke risk. They excel in managing complex and high-dimensional data, allowing them to identify intricate patterns and interactions that traditional methods cannot [6]. ML algorithms can handle large datasets efficiently, making them well-suited for use with the extensive data generated by electronic health records and various digital health technologies [5]. These algorithms also provide personalized risk assessments by learning from individual patient data, thus offering predictions that are more directly applicable to individual care scenarios [7]. Additionally, ML models are inherently adaptable, capable of being continuously updated as new data becomes available or as medical knowledge evolves [8]. This adaptability ensures that ML-based predictions remain relevant and accurate over time. Overall, machine learning significantly enhances the predictive power and utility of stroke risk models, representing a transformative advance in the approach to predicting and ultimately mitigating stroke risk.

This essay explores the advantages of ML algorithms over traditional stroke prediction models by demonstrating their superior data handling capabilities. Additionally, it addresses the challenges inherent in adopting these technologies and proposes potential solutions to facilitate their integration into clinical practice.

2 Explain Machine Learning Algorithms

2.1 Definition and Overview

ML is a branch of artificial intelligence dedicated to creating systems capable of learning from data and making decisions or predictions independently, without specific programming for each task. The main objective of ML is to allow computers to learn autonomously, minimizing the need for human intervention, and to adjust their actions accordingly. This ability is highly transformative across various industries, as it facilitates a dynamic method of data analysis, yielding insights that might be difficult for humans to uncover manually [9].

Machine learning is divided into three main categories based on how models learn from data. The first type is Supervised Learning. Supervised learning involves training a model on a labeled dataset, where the outcomes are known. This allows the model to learn to predict the output from the input data. A classic example is stroke prediction, where models are trained using datasets containing various patient parameters and whether they had a stroke. Techniques such as logistic regression can be used for binary predictions, while neural networks are capable of handling non-linear relationships within large datasets [10].

The second type is Unsupervised Learning. Unsupervised learning identifies patterns or intrinsic structures in data that have not been labeled or classified. Techniques like clustering help discover these hidden patterns by grouping similar data points together. For instance, principal component analysis (PCA) reduces the dimensionality of data by transforming the original variables into a set of new variables (principal components), which are easier to analyze and visualize [11].

The third type is Reinforcement Learning. Reinforcement learning is based on the method of learning optimal actions through trial and error, primarily using rewards and penalties as signals. It is particularly effective in environments where the machine must make decisions without prior knowledge and learn from the consequences of its actions, as seen in robotics and advanced game systems [12].

2.2 The mechanism of algorithms

Data segmentation and cross-validation techniques are crucial for assessing the performance of machine learning models. They work together to provide a thorough understanding of a model's effectiveness on various datasets. In machine learning, data is generally split into two sets: a training set for building and training the model, and a testing set for evaluating its performance. This division is essential for determining how well the model generalizes to new, unseen data [11].

Cross-validation is a reliable statistical method for evaluating machine learning models by dividing the dataset into multiple partitions, ensuring the model is tested on unseen data. The most common approach, K-fold cross-validation, splits the dataset into 'K' equal parts or folds. Each fold is used as a test set in turn, while the remaining folds serve as the training set. This process is repeated until each fold has been used as the test set. Cross-validation assesses how well the results of a statistical analysis will generalize to an independent dataset. It involves partitioning a data sample into complementary subsets, performing the analysis on one subset (the training set), and validating the analysis on another subset (the validation or testing set). This technique helps reduce model overfitting and enhances the model's robustness [11].

Feature selection and engineering are crucial for optimizing model performance, closely linked with data segmentation and cross-validation. By refining features through selection and engineering, models become more efficient and accurate. Data segmentation ensures these features are tested on independent data sets, enhancing the model's reliability. Cross-validation complements this by using multiple data splits to validate the model, ensuring it generalizes well across different data subsets, thus providing a robust evaluation of its performance. Feature selection and engineering are critical steps in improving the performance of ML models. Feature selection is the process of choosing the most relevant features for training, which reduces dimensionality and enhances model efficiency. Feature engineering involves applying domain expertise to create new features from raw data to boost the performance of machine learning algorithms. This is particularly significant in medical datasets where specific features can significantly impact outcomes, such as stroke prediction [13].

3 The factors affecting stroke and the application of machine learning algorithms

3.1 Modifiable and Non-modifiable Risk Factors

Modifiable risk factors are those aspects of an individual's health and lifestyle that can be changed to reduce the risk of developing certain diseases, including stroke. Addressing these factors can significantly lower the risk of stroke. Here are some key modifiable risk factors for stroke:

Hypertension is one of the most critical modifiable risk factors for stroke. It damages blood vessels and can lead to blockages or bursts, directly contributing to the occurrence of strokes. ML models leverage patient data to effectively predict and manage hypertension. By analyzing trends in blood pressure readings alongside lifestyle habits and medication adherence, these models can suggest personalized adjustments in treatment, potentially reducing the risk of stroke significantly [14].

Smoking significantly increases the risk of stroke by accelerating clot formation, thickening blood, and increasing the amount of plaque buildup in arteries. ML algorithms help design and implement personalized smoking cessation programs by analyzing individual behavior patterns, past cessation attempts, and physiological responses to different cessation methods. These models can predict which individuals are likely to respond best to specific interventions, thereby increasing the likelihood of successful quitting [15].

Diabetes is a major risk factor due to its role in contributing to high blood pressure and arterial damage. ML can analyze data from continuous glucose monitors and other health inputs to forecast the risk of stroke and recommend personalized management plans. These might include dietary recommendations, exercise, and medication adjustments tailored to individual needs and responses, helping to maintain optimal blood glucose levels and reduce stroke risk [14].

Excessive weight increases the burden on the circulatory system and is often associated with other stroke risk factors, including hypertension and diabetes. ML models analyze dietary habits, physical activity levels, and genetic predispositions to recommend personalized weight loss programs. These programs are optimized for effectiveness based on an individual's unique health profile and response to previous weight loss strategies [14].

High levels of LDL cholesterol contribute to the buildup of fats in the arteries, leading to atherosclerosis, which can cause ischemic strokes. ML models are used to analyze dietary patterns and genetic data to identify individuals at high risk and recommend personalized dietary or pharmacological interventions aimed at reducing cholesterol levels [15].

Excessive alcohol intake is linked to an increased risk of stroke. ML can help in identifying patterns of alcohol consumption that are likely to increase stroke risk and develop personalized intervention strategies, which may include counseling, medication, or other support mechanisms designed to reduce alcohol intake [14].

Non-modifiable Risk Factors Non-modifiable risk factors are those that cannot be changed or influenced through lifestyle or medical interventions. These factors inherently increase the risk of diseases such as stroke, and awareness of them is crucial for understanding personal risk and implementing appropriate monitoring strategies. Here are the key non-modifiable risk factors for stroke:

The risk of stroke increases with age, particularly after the age of 55. While age itself cannot be changed, understanding its impact on stroke risk is crucial for early intervention. ML models can analyze age-related health data to optimize screening and preventive measures for older adults, focusing on more aggressively managing other modifiable risk factors [16]. Genetic predisposition plays a significant role in an individual's risk of stroke. ML algorithms analyze genetic markers to identify individuals at increased risk. This information can be used to implement early preventive measures, including lifestyle modifications and regular monitoring for other stroke risk factors [11]. Certain ethnic groups, such as African Americans, have a higher risk of stroke. ML models tailored to these populations can improve the accuracy of risk predictions and the effectiveness of targeted prevention programs. This approach allows healthcare providers to address specific genetic, lifestyle, and environmental factors influencing stroke risk in these groups [16].

Men are generally at a higher risk of stroke at a younger age, while women see their risk increase later in life, especially after menopause. ML models that incorporate gender-specific data can help design gender-appropriate prevention and treatment strategies, considering hormonal and physiological differences [17]. ML excels at uncovering complex interactions between risk factors that may not be apparent through traditional analysis methods. For instance, the combined effect of smoking and hypertension on stroke risk can be more severe than the sum of their individual effects. ML models analyze these interactions in large datasets, providing insights that can lead to more effective combined interventions [17].

Socio-economic status and environmental conditions are significant contributors to stroke risk. Lower socio-economic status is often associated with higher stroke risk due to factors like limited access to healthcare, poor diet, and high stress. ML algorithms incorporate these socio-economic variables to understand their impact on stroke risk better and to develop targeted interventions [15]. Environmental factors, including air quality and exposure to pollutants, are increasingly recognized for their impact on health. ML models that analyze environmental data can identify regions with high stroke risks, helping to direct public health resources and interventions to areas where they are most needed [15].

3.2 The Application of Machine Learning Algorithms in Stroke Prediction

Data Sources and Integration. ML in stroke prediction heavily depends on the diversity and quality of data sources. Key data sources include.

Firstly, they are from Electronic Health Records (EHRs). EHRs are a foundational data source, providing detailed patient histories, treatment records, and outcomes. ML algorithms use this data to identify patterns and risk factors associated with stroke incidents. Advanced analytics can highlight critical correlations and trends that might be overlooked in routine clinical assessments.

Secondly, they are from Wearable Technology Data. Devices such as smartwatches and fitness trackers offer real-time monitoring of physiological data, including heart rate, blood pressure, and activity levels. ML models can analyze this continuous data stream to detect early signs of conditions like atrial fibrillation, which significantly increases stroke risk.

Thirdly, they are from Genetic Information. Genetic testing reveals predispositions to various conditions, including stroke. By integrating genetic data, ML models can provide insights into the hereditary aspects of stroke risk, allowing for early interventions tailored to an individual's genetic profile.

Fourthly, they are from Imaging Data. Advanced imaging techniques, such as MRI and CT scans, provide detailed visualizations of brain health. ML algorithms can analyze these images to detect early signs of cerebral small vessel disease and other pre-stroke conditions, often before they manifest clinically.

Finally, they are from Social Determinants of Health. Information on socioeconomic status, education level, and environmental factors is increasingly recognized as crucial to understanding health outcomes. ML models that include these parameters can provide more accurate risk assessments, accounting for external factors that may influence an individual's health.

Model Development and Training This essay takes the stroke prediction dataset as an example, comprising 15,000 entries and 22 distinct features to predict stroke likelihood. This dataset encapsulates a comprehensive range of attributes, providing a holistic view of each patient's health and lifestyle, which are critical for effective stroke prediction. The development and training of ML models in stroke prediction involve several key steps. The first step is preprocessing. This study cleans the data by handling missing values and standardize it by normalizing or standardizing numerical features. Cleaning the data involves handling missing values which might occur within the dataset. Missing data can lead to inaccurate predictions and compromise the statistical power of the machine learning model. Cleaning can be achieved through various methods such as imputation, where missing values are filled based on other data points or removed entirely if they represent a significant portion of the data. Standardization (rescaling the data to have a mean of zero and a standard deviation of one) ensures that each feature contributes equally to the analysis, improving the stability and performance of the machine learning algorithms.

The second step is the feature selection. The study uses Truncated Singular Value Decomposition (Truncated SVD) for reducing dimensionality and selecting meaningful features. The feature selection step in machine learning, particularly employing Truncated Singular Value Decomposition (Truncated SVD), plays a critical role in refining the predictive model by reducing dimensionality and isolating the most impactful features. Truncated SVD is a matrix factorization technique that reduces the number of features in a dataset by transforming the data into a lower-dimensional space while preserving essential information. This method effectively identifies and retains the most significant features that capture the majority of the data's variance, thereby enhancing the model's performance. By focusing on these key features, Truncated SVD reduces computational complexity and potential overfitting and improves the interpretability of the model, making it easier to understand and explain.

The third step is the algorithm selection, with the Random Forest algorithm chosen for the model. The Random Forest algorithm is a sophisticated ensemble learning technique used for both classification and regression tasks, which builds upon the simplicity and foundational principles of decision trees. It involves creating a multitude of decision trees during the training phase and making predictions by obtaining the mode of the classes (in classification tasks) or the mean prediction (in regression tasks) from the individual trees. This method leverages the strengths of multiple trees to produce a more accurate and robust model compared to using a single decision tree. The ensemble approach helps achieve better generalization of unseen data, a core goal in machine learning applications.

Key features of Random Forest include its ability to handle large datasets with higher dimensionality efficiently and its robustness against overfitting, thanks to the random selection of features and bootstrap sampling used in constructing the trees. Each tree in the forest is built from a random sample of data points and a subset of features, which ensures diversity among the trees and contributes to a more stable combined prediction. This feature randomness helps in de-correlating the trees, significantly reducing the model's variance without substantially increasing the bias. Moreover, Random Forest can automatically rank the importance of different features for the prediction task, providing valuable insights into the data. This combination of high accuracy, ease of use, and interpretability makes Random Forest popular among data scientists across various sectors.

The fourth step involves training and validation, with the data split into training and validation sets to build and validate the model.

The fifth step is evaluation, which assesses the model using metrics such as accuracy, sensitivity, and the area under the ROC curve (AUC).

Accuracy measures the ratio of correct predictions (both true positives and true negatives) to the total number of cases examined. In simpler terms, it reflects how many predictions made by the model were correct. Accuracy ranges from 0 to 1, where 0 means no predictions were correct (0% accurate), and 1 means all predictions were correct (100% accurate).

Sensitivity, also known as recall, measures the proportion of actual positives (in this case, actual stroke cases) that the model correctly identified. It is particularly important in medical settings like stroke prediction, as it reflects how effective the model is at detecting positive cases, which can be critical for timely intervention. Sensitivity also ranges from 0 to 1, where 0 means no positive cases were correctly identified, and 1 means all positive cases were correctly identified.

The ROC AUC score is a metric used to evaluate the performance of classification models across different threshold settings. AUC stands for "Area under the ROC Curve." The ROC curve displays the true positive rate (sensitivity) versus the false positive rate (1 - specificity) at various thresholds. AUC measures how well the model distinguishes between classes. A higher AUC indicates a better model, capable of correctly predicting both positive and negative classes. The AUC value ranges from 0 to 1, where 0.5 suggests no discriminatory ability (similar to random guessing), and 1.0 indicates perfect discrimination between positive and negative classes.

Real-world Application. ML models are increasingly being integrated into clinical practice for stroke prediction and management:

The first one is Risk Assessment Tools. Hospitals and clinics use ML-based tools to assess stroke risk in patients. These tools analyze patient data in real time to identify high-risk individuals who may benefit from preventative interventions.

The second one is Clinical Decision Support Systems. ML models provide recommendations directly to healthcare providers, helping them make informed decisions about patient care. This can include suggestions for further testing, preventive measures, or immediate interventions.

The third one is Patient Monitoring. In critical care settings, ML algorithms continuously analyze patient data to detect any changes that might indicate an increased risk of stroke. Alerts can be generated to prompt immediate medical action, potentially saving lives.

The fourth one is Public Health Initiatives. ML can also guide public health strategies by identifying populations at higher risk based on aggregated health data. The Public Health Initiatives can enable targeted health campaigns and resource allocation to areas most in need.

Personalized Medicine. Personalized medicine leverages individual patient data to tailor medical treatment, optimize health outcomes, and enhance preventive care strategies, particularly in areas as critical as stroke prediction and management. Machine Learning (ML) algorithms play a crucial role in realizing the full potential of personalized medicine by enabling more accurate, individualized predictions and treatments based on comprehensive data analysis.

Firstly, Individualized Risk Assessments can be made. To create individualized stroke risk assessments, ML algorithms can analyze complex, multi-dimensional data such as genetic profiles, medical histories, and lifestyle factors. This allows healthcare providers to identify patients at high risk and implement early intervention strategies tailored to each individual's specific risk factors.

Secondly, Tailored Treatment Plans can be made. Once risk levels are determined, ML aids in crafting personalized treatment plans that may include specific medications, recommended lifestyle adjustments, and regular monitoring schedules. These plans are dynamically adjusted based on continuous learning from patient progress and new health data, ensuring that the treatment remains optimal as the patient's condition evolves.

Thirdly, predictive analytics for preventive measures can be used. ML models utilize predictive analytics to recommend preventive measures that are significantly more precise than those derived from general population data. For example, by predicting the likelihood of stroke from minor but frequent symptoms, ML can prompt preventive measures well before traditional methods would.

By integrating these ML-driven capabilities, personalized medicine can significantly enhance stroke prevention strategies' effectiveness, offering treatments specifically designed for individual patient profiles and thereby improving overall healthcare outcomes.

4 Challenges and Solutions

4.1 Challenges

Firstly, data privacy and security are among the foremost challenges in implementing machine learning (ML) in healthcare, particularly in stroke prediction, ensuring patient data privacy and security. Health data is incredibly sensitive, and the potential for breaches poses significant risks. Advanced encryption methods, secure data environments, and stringent access controls are essential to safeguard this data. Moreover, adhering to regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in the EU is essential for maintaining trust and legality [14]. Machine learning models should incorporate privacy-preserving technologies, such as federated learning, which allows the model to learn from decentralized data without sharing it, thereby enhancing privacy [11].

Integrating ML models into existing healthcare frameworks presents the second challenge. These include data compatibility issues, the need for healthcare professionals to adapt to new technologies, and the systems' ability to handle real-time data inputs effectively. Achieving interoperability between different electronic health record (EHR) systems and ML platforms is crucial. This requires standardizing data formats, developing APIs that facilitate smooth data exchange, and training personnel to use these new tools efficiently. Ensuring that ML predictions are timely, relevant, and accessible within the clinical workflow is critical for their adoption and effectiveness [10].

Thirdly, Algorithmic bias is a significant concern in ML applications. If the data used to train algorithms is not representative of the entire population, there is a risk that the models developed will perpetuate existing biases or create new ones. This can lead to unequal healthcare outcomes among different demographic groups. To mitigate this, it's essential to use diverse training datasets and employ algorithmic fairness approaches to ensure that ML models treat all patient groups equitably. Regular audits and updates to the algorithms are necessary to maintain fairness over time [17].

Another challenge is navigating the complex landscape of regulations that govern medical data and patient privacy. Additionally, ethical issues such as determining the responsibility for decisions made based on ML recommendations can arise. Transparent and explainable ML models are crucial to address these concerns. Healthcare providers and patients must understand how ML tools make decisions to foster trust and ensure ethical usage. The development of clear guidelines and standards by regulatory bodies will be essential as these technologies become more integrated into healthcare [15].

4.2 Solutions

ML models are only as good as the data they are trained on and must be continually updated to reflect new medical research and changing population health trends. Ensuring that models do not become outdated requires regular retraining and refinement. This ongoing monitoring must be built into the deployment of ML systems to ensure they remain accurate and effective over time. Furthermore, continuous monitoring can identify when models start to perform poorly due to changes in underlying data patterns, prompting timely updates [16].

Maintaining performance and accuracy becomes challenging as ML solutions are scaled to handle larger populations and more complex datasets. Efficient algorithms and scalable infrastructure are necessary to manage the increased computational load. Cloud-based solutions and advancements in distributed computing can help manage these demands, ensuring that ML systems can scale up without losing functionality [14].

ML models must undergo rigorous validation to prove their accuracy and reliability to be truly effective in clinical settings. This involves extensive real-world testing to ensure that the models perform well across various scenarios and patient populations. Gaining the trust and buy-in of healthcare professionals is also crucial for adoption. This requires demonstrating ML systems' clinical relevance and benefits, backed by robust clinical trials and peer-reviewed research [13].

Addressing these challenges requires a coordinated effort among data scientists, healthcare professionals, regulatory bodies, and technology developers. By tackling these issues head-on, the integration of ML in healthcare, particularly in areas like stroke prediction and treatment, can lead to more efficient, personalized, and proactive medical care.

5 Conclusion

This study aims to extensively utilize machine learning (ML) techniques to predict the risk of stroke, employing methodologies like Truncated Singular Value Decomposition for feature selection and Random Forest algorithms for model training. These methods were chosen for their robustness in handling large datasets and their ability to reduce dimensionality and manage feature importance, respectively effectively. The study focused on preprocessing, feature engineering, and rigorous model validation to ensure accurate predictions and high reliability of the findings.

The findings reveal significant insights into the modifiable and non-modifiable risk factors for stroke, emphasizing the potential of ML in personalized medicine and public health. The study identified key risk factors by analyzing extensive datasets and demonstrated how ML could tailor prevention strategies to individual needs, potentially revolutionizing stroke prevention and management. The use of cross-validation and various performance metrics like accuracy, sensitivity, and AUC further underscored the model's effectiveness in real-world settings.

The significance of this research lies in its potential to facilitate early intervention strategies and improve the allocation of healthcare resources, ultimately leading to better health outcomes. Ongoing advancements in ML could enhance the precision of risk predictions and broaden the scope of personalized healthcare, making it a cornerstone of modern medical practices.

References

1. WSO, Global Stroke Fact Sheet 2022 ,viewed April 21, (2024) https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf, last accessed 2024/4/21.
2. WHO, The top 10 causes of death, viewed April 21, 2024, (2020)https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, last accessed 2024/4/21.
3. Girotra T. et al, A contemporary and comprehensive analysis of the costs of stroke in the United States, Journal of the Neurological Sciences, ScienceDirect, viewed April 21, (2024).
4. Wolf, P. A., D'Agostino, R. B., Belanger, A. J., & Kannel, W. B.: Probability of stroke: a risk profile from the Framingham Study. Stroke, 22(3), 312-318 (1991).
5. Beam, A. L., & Kohane, I. S.: Big Data and Machine Learning in Health Care. JAMA, 319(13), 1317-1318 (2018)
6. Rajkomar, A., Dean, J., & Kohane, I. , Machine learning in medicine. The New England Journal of Medicine, 380(14), 1347-1358(2019).
7. O'Donnell, M. J., Chin, S. L., Rangarajan, S., et al.: Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study, Lancet, 388(10046), 761-775 (2016).
8. Goyal, M., Demchuk, A. M., Menon, B. K., et al.: Randomized assessment of rapid endovascular treatment of ischemic stroke. The New England Journal of Medicine, 372(11), 1019-1030 (2016).
9. Alpaydin, E. : Introduction to machine learning (4th ed.). MIT Press (2020).
10. James, G., Witten, D., Hastie, T., & Tibshirani, R. : An introduction to statistical learning with applications in R (2nd ed.). Springer (2021).
11. Shalev-Shwartz, S., & Ben-David, S. : Understanding machine learning: From theory to algorithms. Cambridge University Press (2014).
12. Kohavi, R., & Provost, F. : Glossary of terms. Machine Learning, 30(2-3), 271-274 (1998).
13. Liu, Y., Chen, P. H., Krause, J., & Peng, L. : How to read articles that use machine learning: Users' guides to the medical literature. JAMA, 322(18), 1806-1816 (2019).
14. Luxton, D. D. : Artificial intelligence in behavioral and mental health care. Academic Press (2014).
15. Obermeyer, Z., & Emanuel, E. J. : Predicting the future - Big data and machine learning in health care. The New England Journal of Medicine, 376, 1216-1219 (2016).
16. Topol, E. J. :Deep medicine: How artificial intelligence can make healthcare human again. Basic Books (2019).
17. Zou, J., & Schiebinger, L. :AI can be sexist and racist - it's time to make it fair. Nature, 559, 324-326 (2018).