Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241570

The application of machine learning in the field of biomedical science

Zijing Li

The Department of Internet of Things Engineering, Changsha University, Changsha, 410022, China

lizijing@ldy.edu.rs

Abstract. The application of Machine Learning (ML) in the field of biomedical science has been rapidly evolving, providing novel solutions for complex challenges such as disease prediction, drug design, and personalized medicine. This paper presents an overview of ML applications in biomedicine, focusing on three main areas: cancer prediction, common disease prediction, and drug design. The introduction to machine learning workflow is briefly discussed, highlighting the essential steps involved in training ML models with biomedical data. In the realm of cancer prediction, ML algorithms are used to analyze large datasets containing patient information, genetic profiles, and clinical variables to predict cancer susceptibility and progression. Similarly, in predicting common diseases, ML techniques are employed to identify patterns and risk factors associated with conditions like diabetes, cardiovascular diseases, and respiratory illnesses. Furthermore, ML plays a critical role in drug design by accelerating the discovery process, optimizing pharmacokinetic properties, and predicting drug responses. The discussion section delves into the potential impacts and limitations of ML in these areas, emphasizing the need for accurate data, robust validation methods, and interdisciplinary collaborations. In conclusion, the integration of ML in biomedical science offers transformative opportunities for improving healthcare outcomes. However, careful consideration of ethical, legal, and social implications is necessary as these technologies continue to advance.

Keywords: Machine Learning, Biomedical Science, Cancer Prediction, Disease Prediction.

1. Introduction

The application of machine learning in the biomedical field refers to the use of machine learning algorithms and techniques to analyze and interpret biomedical data, thereby uncovering biological mechanisms, assisting in disease diagnosis, predicting disease progression, optimizing treatment regimens, discovering biomarkers, and accelerating drug development. With the popularization of high-throughput technologies such as genetic sequencing, proteomics, and metabolomics, the biomedical field has generated a massive amount of data. This data contains precious information about disease mechanisms, drug actions, and individual differences. However, traditional data processing methods are unable to effectively analyze these complex datasets. Machine learning can assist in identifying new biomarkers, which can be used for early diagnosis of diseases, treatment monitoring, and prognosis assessment. Precision medicine aims to customize treatment plans based on

DOI: 10.54254/2755-2721/87/20241570

an individual's genetic information, lifestyle, and environmental factors. Machine learning is capable of processing and analyzing multidimensional data, helping doctors make more accurate diagnoses and treatment decisions. It can detect early signs of diseases from health data, which is beneficial for early intervention to prevent the onset of diseases or to reduce their severity.

Multivariate analysis and Machine Learning (ML) methods have been used to analyze these spectral datasets. Reduction and clustering are two forms of "unsupervised" machine learning. Dimensionality reduction algorithms project data into a lower dimensional (usually two or three dimensional) space to preserve as much of the original information as possible [1]. Common algorithms include principal component analysis (PCA), T-distributed Random neighborhood Embedding (tGSNE), and Unified Manifold Approximation and Projection (UMAP) [2]. The clustering algorithm clearly divides each observation into discrete groups based on their similarity to each other, which increases visualization and facilitates interpretation of high-dimensional data. Common algorithms such as K-means clustering. Distinct from the processes of dimensionality reduction and cluster analysis, predictive modeling engages in supervised learning, where it establishes a connection between the measured attribute linked to every data instance in the dataset and the corresponding target value it aims to predict. Traditional supervisory models include linear discriminant analysis (LDA), partial least-squares discriminant Analysis (PLSGDA), support vector machine (SVM), K-nearest neighbor algorithm (KNN), and decision tree-based models - random forest (RF) and regression tree (CRT) [3]. In addition, deep learning methods using complex artificial neuron structures enable advanced feature and pattern recognition. Among them, convolutional neural networks (CNNS) use shared weight filters and pooling layers in their architecture, showing higher specificity and sensitivity. When evaluating the model, in the case that the sample size is not enough to form a large number of test sets, cross-validation can evaluate the model performance by omitting a validation set during training [4], and constructing multiple permutations of the training set and validation set, such as K-fold multiple cross-validation method, leave one method, etc. The use of cross-validation must be handled carefully, selecting representative features (variables) or increasing large sample sizes based on the complexity of the features, otherwise, it will be easy to generate overfitted models, that is, high performance on the training set and poor performance on the test set or validation set. In addition to its direct application as a forecasting tool, many monitoring models can perform the function of "feature selection", in which the model ranks all the features in the input in order of importance [5], and only the most important features used to predict a particular outcome are identified and included in the final model [6]. Typically, these selected features encapsulate biological insights that correspond to potential therapeutic targets, the molecular mechanisms underlying disease, or serve as biomarkers in the context of diagnosing or tracking specific cancers.

The research objective of this article is to elucidate the extensive data analysis and processing capabilities of machine learning in the biomedical field. This article summarizes a considerable body of research and methods in the biomedical field from multiple perspectives, providing an overview of the machine data processing process and methods and introducing new approaches.

2. Method

2.1. Introduction to machine learning workflow

Machine learning, as a branch of artificial intelligence, aims to relate problems learned from data samples to general concepts of reasoning. The learning process of machine learning shown in Figure 1 consists of two stages: first, learning from a given data set to an unknown pattern. Second, use the learned patterns to predict new outputs based on the inputs. Machine learning can be primarily categorized into two types: supervised learning and unsupervised learning. In supervised learning, a dataset with labeled examples is utilized in the initial phase of the learning process, as opposed to unsupervised learning, where data without labels is examined. Supervised learning is usually divided into classification problems and regression problems. Classification problem refers to the process of

DOI: 10.54254/2755-2721/87/20241570

learning to divide input data into a finite set of classes. Regression is the problem of learning to map the input data to real values.

When applying machine learning methods, data samples form the basic components. Each sample can be described by several attributes, each consisting of a different type of value. In addition, knowing the specific type of data used in advance allows for the right choice of tools and techniques for analysis. Of course, in order to more accurately realize the effect of data analysis in machine learning, some other issues related to data should be handled well, including data quality improvement and data preprocessing steps, so that the used data is more suitable for training. Data quality challenges encompass various issues such as noise, anomalies (outliers), incomplete or missing data, and data that lacks representativeness. When the quality of the data is improved, the quality of the results is usually improved accordingly. In addition, in order to make the raw data more suitable for further analysis, the data preprocessing step should be adopted, focusing on the transformation of the data. Data preprocessing can use many different techniques and strategies with the goal of transforming the data to better fit a particular approach. Among these techniques, the most important methods include dimensionality reduction, feature extraction and feature selection.

After data preprocessing is completed, the model is trained using the data. In the process of model training, there will be training errors and testing errors. The former is the classification error of training data, and the latter is the error of test data. A good classification model should be able to fit the training set well and be able to classify all instances accurately. If the test error of the model starts to increase while the training error is decreasing, the model overfitting phenomenon will occur.

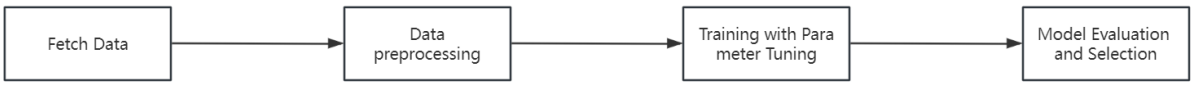

Figure 1. Machine learning workflow diagram (Photo/Picture credit: Original).

2.2. Cancer prediction

2.2.1. Prediction of lung cancer

In recent years, deep learning has made significant progress in the field of image recognition, opening up new avenues for the diagnosis and prediction of lung cancer A new predictive model of lung cancer based on convolutional neural networks (CNN) combined with attention mechanism was proposed. The model is able to automatically extract useful features from lung CT images and improve the accuracy of predictions by focusing on areas most likely to contain lesions through attention mechanisms. Arun Kumar Rana et al. discussed the evolution of machine learning in biomedical engineering, which includes applications in cancer prediction [1]. Lele Ye et al. identified potential N6-methyladenosine effector-related lncRNA biomarkers for serous ovarian carcinoma using machine learning methods [7].

2.2.2. Prediction of brain cancer

In addition to lung cancer, machine learning techniques have also been widely used in brain cancer prediction. By analyzing the patient's MRI or CT image data, combined with the patient's clinical information, researchers can build machine learning models to predict the probability of brain cancer. This predictive method can help doctors assess patients' conditions more accurately and develop more reasonable treatment plans for patients. Muhammad Shahid Shamim et al. explored the role of artificial intelligence and machine learning in medical education, including their use in understanding and predicting diseases like cancer [8].

DOI: 10.54254/2755-2721/87/20241570

2.3. Prediction of common diseases

2.3.1. Prediction of heart disease

Heart disease remains one of the leading causes of mortality worldwide, necessitating novel approaches for early detection and prevention. Machine learning has become a formidable instrument within this sphere, capable of processing extensive datasets and uncovering subtle patterns that might escape the detection of conventional statistical approaches. This section delves into the application of machine learning techniques in predicting heart disease, shedding light on the various algorithms used, the features or biomarkers identified, and the performance metrics achieved.

Machine learning algorithms utilized for heart disease prediction run the gamut from simple logistic regression models to complex deep learning architectures. One of the critical steps in developing these predictive models is feature selection, which involves choosing the most relevant biomarkers from a patient's medical history and current health profile. Characteristics commonly associated with this condition encompass factors such as age, gender, blood pressure readings, cholesterol levels, smoking habits, and a family history of heart disease, among others. These data points are subsequently input into predictive models to estimate the probability of an individual being at risk for heart disease in the future. Wu Hang et al. developed a novel causal inference algorithm for personalized biomedical causal graph learning, which can be applied to the prediction of common diseases [3].

Various studies have used these metrics to compare the performance of different machine learning models. For instance, a meta-analysis conducted by compared the efficacy of neural networks, decision trees, and logistic regression models across multiple heart disease prediction datasets. The analysis revealed that while neural networks had the highest average AUC, decision trees had the best balance between sensitivity and specificity. These comparisons are vital for determining the most suitable model for a particular clinical setting or population.

Moreover, the integration of machine learning models into clinical practice requires attention to practical considerations such as interpretability and computational efficiency. While deep learning models may offer higher accuracy, their "black box" nature can be a hindrance in healthcare settings where explainability is paramount. Conversely, models such as logistic regression, while more transparent, might not offer the sophisticated insights provided by more intricate model. Balancing these factors is crucial for the successful deployment of machine learning in heart disease prediction.

2.3.2. Stroke prediction

Stroke, much like heart disease, is a major health concern with significant morbidity and mortality rates. The ability to predict strokes before they occur could greatly improve patient outcomes by enabling early intervention. Machine learning has been at the forefront of efforts to predict stroke risk, leveraging a variety of algorithms, biomarkers, and performance metrics to achieve accurate predictions. Huan Jia Ming et al. constructed a biomedical knowledge graph of symptom phenotype in coronary artery plaque, aiding in the prediction and analysis of common cardiovascular diseases [4].

The algorithms employed for stroke prediction in machine learning are similar to those used for heart disease, encompassing techniques such as support vector machines, random forests, gradient boosting machines, and neural networks. However, the selection of features or biomarkers often differs, reflecting the unique physiological aspects of stroke. Common features include blood pressure, body mass index (BMI), history of diabetes, smoking and alcohol consumption habits, prior instances of transient ischemic attack (TIA), and the presence of carotid plaque, among others.

Performance metrics for stroke prediction are similarly focused on accuracy, sensitivity, specificity, and AUC. However, given the time-sensitive nature of stroke interventions, models are also evaluated based on their speed and computational efficiency. Rapid prediction is crucial for providing timely treatment to patients who may be experiencing a stroke. Therefore, models must not only be accurate but also capable of generating predictions quickly enough to make a difference in emergency

DOI: 10.54254/2755-2721/87/20241570

situations. Ensemble methods have likewise been explored in stroke prediction, aiming to combine the strengths of multiple models.

The interpretability issue is even more pertinent in stroke prediction due to the immediate consequences of misdiagnosis. Researchers are thus exploring techniques such as local interpretable model-agnostic explanations (LIME) and SHAP values to explain model predictions in understandable terms. These methods help clinicians understand why a model made a particular prediction, which is essential for gaining trust and promoting the adoption of machine learning in clinical practice. Jason Nan et al. used personalized machine learning-based prediction to assess wellbeing and empathy in healthcare professionals, which can be associated with their ability to predict and respond to common diseases [9].

In conclusion, machine learning has shown immense promise in predicting both heart disease and stroke, two of the most significant health concerns globally. Through a variety of algorithms, biomarkers, and performance metrics, researchers have made strides in improving prediction accuracy and efficiency. However, challenges remain, particularly in balancing model complexity with interpretability and speed. As machine learning continues to evolve, so too will its applications in predicting and preventing these life-threatening conditions.

2.4. Drug design

Drug design is one of the important research directions in biomedical science. Traditional approaches to drug design require a lot of time and resources, and have a low success rate. Machine learning techniques can build predictive models to predict the biological activity of new compounds by analyzing known data on compound structure and activity. This prediction method can greatly accelerate the process of drug design and improve the efficiency of new drug research and development. In addition, machine learning techniques can be used to optimize the structure and properties of drug molecules for better efficacy and lower side effects. Hyung Eun An et al. developed a two-layer machine learning model for the classification of legal and illegal poppy varieties, which has potential applications in drug design and development [10]. Diwei Zhou et al. emphasized the combination of data-driven machine learning approaches with prior knowledge for robust medical image processing and analysis, crucial for drug design and testing phases [11].

3. Discussion

In the modelling process, Convolutional Neural Networks (CNNs) were employed for diabetes risk prediction. Originally designed for image analysis, CNNs were adapted to handle time-series data such as fluctuations in blood sugar levels. Additionally, Random Forest and Support Vector Machines (SVMs) were integrated into the framework, with their predictions evaluated alongside CNNs. The outcomes were presented in a graphical user interface, facilitating a visual understanding of the diabetes risk. The objective was to leverage multiple models for enhanced accuracy in risk assessment through comparative analysis and visualization. Medical data is often faced with problems such as high dimension of data features, redundant features and irrelevant features. For some specific machine learning algorithms, it is unknown which features are effective for the model. In order to reduce redundant information in the data, improve the efficiency of model construction, increase the interpretability of the model and improve the model generalization performance, it is necessary to select the beneficial features of the model from all the features of the data set. Common feature extraction methods include Principal Component Analysis (PCA), mutual information, etc. Through the visual analysis of the diabetes data set in this paper, it is known that the cause of diabetes is determined by multiple factors, and the mutual information method only examines the influence of a single feature on the target variable when selecting features. Thus, effective features cannot be extracted; PCA method is used for dimensionality reduction, but not all the data in diabetes data set conform to normal distribution, so the extracted principal components are not optimal. Moreover, the data in this dataset has many features with small variance, and the correlation between the features is not taken into account, so PCA and mutual information method cannot satisfy the selection of data

DOI: 10.54254/2755-2721/87/20241570

features in this dataset. Since XGBoost algorithm has the function of helping researchers to clarify the influence degree of a specific feature on the label, and has excellent classification effect and robustness on data sets with small sample size, recursive feature elimination method based on XGBoost feature importance, namely RFE-XGBoost, is used in feature selection. The optimal feature subset is selected.

In the health management of diabetic patients, patients upload and store health indicators through a prediction system to help individuals adjust their health status and lifestyle. Diabetes prediction is based on personal health data. With the development of big data technology, data collection is more convenient and the accuracy is greatly improved. Intelligent blood glucose meter, electronic scale, health bracelet and other devices can dynamically obtain health data, and diabetes prediction model can be used to achieve prediction.

First, all the features of the pre-processed data set are input into the model, and the XGBoost algorithm model is used to sort the feature importance, and the whole data set is fitted to complete the preliminary feature screening. When XGBoost algorithm is used to sort the features of the pre-processed data, GridSearchCV (grid search method) is used to find the optimal parameters of the model. First, the learning_rate, n_estimators and max_depth of the tree, three main parameters that determine the model performance, are put into the dictionary param_grid variable as keys. The value of the key is set by arange() in the numpy module, and then the required parameters are put into the Grid Search CV() function, where the value of the parameter cv is set to 10. Then best_params, best_score_, best_estimator_ and best_index_ are used to output the value of the optimal parameter, the score of the model under the optimal parameter, the model under the optimal parameter and the index under the optimal parameter. Finally, the feature importance of the data set is obtained. The plot_importance module in XGBoost library is used to check the feature importance and order.

The XGBoost Special importance score is determined by the sum of the number of splits per tree for a particular feature. For example, if the feature splits once in the first tree, twice in the second tree, three times in the third tree, and so on, then the score of the feature is the sum of the number of times the feature splits on all trees. According to the ranking of feature importance in XGBoost model, the features with zero scores are eliminated in this paper, and the number of remaining features after screening is 20. In the way of recursive elimination, the first n features of feature importance ranking are selected each time to form a feature subset. According to the evaluation index AUC performance of the classifier, the best feature subset of the n feature subsets is selected at last.

Therefore, the optimal feature subsets of the diabetes data set in this paper were HBA1c, blood glucose value, age, BMI, waist circumference, triglyceride, low-density lipoprotein, urea, uric acid, total cholesterol, and alanine aminotransferase. It can be seen from the features in the selected optimal feature subset that both glycosylated hemoglobin and blood sugar are the most favorable indicators for judging diabetes, which is consistent with the medical rules. However, in the traditional diagnosis process, doctors' personal experience and subjective judgment are often relied on, while the feature selection in this paper ranks the importance of diabetes features. In addition, the selection of diabetes features was quantified, and different features had different degrees of impact on the contribution of the prediction model, so that the risk factors inducing diabetes could be known more scientifically and accurately.

4. Conclusion

In conclusion, the applications of machine learning in the biomedical field have revolutionized the way people approach healthcare, diagnosis, and drug discovery. The capacity of machine learning algorithms to handle extensive datasets and reveal patterns invisible to the human eye has pioneered new territories in precision medicine and personalized therapies. From improving the accuracy of disease diagnosis to optimizing treatment plans, machine learning techniques are poised to transform healthcare delivery and patient outcomes.

Moreover, as the technology continues to evolve and become more sophisticated, scientists expect to see even more innovative applications in the biomedical field. Machine learning's potential in

DOI: 10.54254/2755-2721/87/20241570

predictive modelling, risk assessment, and biomarker discovery is particularly promising, as it holds the key to earlier intervention, prevention, and ultimately, the eradication of many chronic and life-threatening diseases.

However, it is also crucial to recognize the challenges that accompany the use of machine learning in biomedical research, including data privacy, interpretability, and ethical considerations. Moving forward, it is crucial to tackle these challenges ethically and sustainably to maximize the benefits of machine learning while mitigating potential risks.

References

[1] Rana AK, Sharma V, Rana SK & Chaudhary VS 2024 Evolution of Machine Learning and Internet of Things Applications in Biomedical Engineering CRC Press: 2024-06-13

[2] Goel N & Yadav RK 2024 Internet of Things enabled Machine Learning for Biomedical Application CRC Press: 2024-03-30

[3] Wu H, Shi W & Wang MD 2024 Developing a novel causal inference algorithm for personalized biomedical causal graph learning using meta machine learning BMC Medical Informatics and Decision Making vol 24 (1) pp 137-137

[4] Huan JM, Wang XJ, Li Y, Zhang SJ, Hu YL & Li YL 2024 The biomedical knowledge graph of symptom phenotype in coronary artery plaque: machine learning-based analysis of real-world clinical data BioData Mining vol 17 (1) pp 13-13

[5] Prasad A, Santra TS & Jayaganthan R 2024 A Study on Prediction of Size and Morphology of Ag Nanoparticles Using Machine Learning Models for Biomedical Applications Metals vol 14 (5)

[6] Labory J, Fotso EN & Bottini S 2024 Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data Computational and Structural Biotechnology Journal vol 23 pp 1274-1287

[7] Ye L, Tong X, Pan K, Shi X, Xu B, Yao X, Zhuo L, Fang S, Tang S, Jiang Z, Xue X, Lu W & Guo G 2024 Identification of potential novel N6-methyladenosine effector-related lncRNA biomarkers for serous ovarian carcinoma: a machine learning-based exploration in the framework of 3P medicine Frontiers in Pharmacology vol 15

[8] Shamim MS, Zaidi SJA & Rehman A 2024 The Revival of Essay-Type Questions in Medical Education: Harnessing Artificial Intelligence and Machine Learning Journal of the College of Physicians and Surgeons--Pakistan : JCPSP vol 34 (5) pp 595-599

[9] Nan J, Herbert MS, Purpura S, Henneken AN, Ramanathan D & Mishra J 2024 Personalized Machine Learning-Based Prediction of Wellbeing and Empathy in Healthcare Professionals Sensors (Basel, Switzerland) vol 24 (8) pp 2640-2642

[10] An HE, Mun MH, Malik A & Kim CB 2024 Development of a two-layer machine learning model for the forensic application of legal and illegal poppy classification based on sequence data Forensic Science International: Genetics vol 71 p 103061

[11] Zhou D, Duan J, Qin C & Luo G 2024 Editorial: The combination of data-driven machine learning approaches and prior knowledge for robust medical image processing and analysis Frontiers in Medicine vol 11