Review Paper

ANNALS OF SURGERY OPEN

OPEN

Clinical Applications of Machine Learning

Nadayca Mateussi, PhD,* Michael P. Rogers, MD,† Emily A. Grimsley, MD,† Meagan Read, MD,† Rajavi Parikh, DO,† Ricardo Pietrobon, MD, PhD,* and Paul C. Kuo, MD†

Objective: This review introduces interpretable predictive machine learning approaches, natural language processing, image recognition, and reinforcement learning methodologies to familiarize end users.

Background: As machine learning, artificial intelligence, and generative artificial intelligence become increasingly utilized in clinical medicine, it is imperative that end users understand the underlying methodologies.

Methods: This review describes publicly available datasets that can be used with interpretable predictive approaches, natural language processing, image recognition, and reinforcement learning models, outlines result interpretation, and provides references for in-depth information about each analytical framework.

Results: This review introduces interpretable predictive machine learning models, natural language processing, image recognition, and reinforcement learning methodologies.

Conclusions: Interpretable predictive machine learning models, natural language processing, image recognition, and reinforcement learning are core machine learning methodologies that underlie many of the artificial intelligence methodologies that will drive the future of clinical medicine and surgery. End users must be well versed in the strengths and weaknesses of these tools as they are applied to patient care now and in the future.

Keywords: machine learning, interpretable predictive machine learning, natural language processing, image recognition, reinforcement learning

INTRODUCTION

Machine learning (ML), artificial intelligence (AI), and notably, generative AI have entered daily clinical and research discussions. These methodologies enable crowdsourcing vast amounts of data and subjecting these to new open-source analytic techniques powered by modern computational methods to explore unanswered questions or generate hypotheses that were heretofore unimaginable. In this regard, research questions are largely driven by available analytical methods. For example, using a clinical trial framework, the research question will question the efficacy of intervention A versus B on outcome Y for patient group Z. With ML and AI becoming increasingly common, it is essential for users to understand their fundamentals and implications. In the words of Albert Einstein, "If you can't explain it simply, you don't understand it well enough."

Therefore, this article aims to review ML models and their applicable clinical questions, discuss publicly available datasets for these models, outline result interpretation, and provide references for in-depth information about each analytical framework. Despite the expanding repertoire of ML methods, this review concentrates on interpretable predictive ML models, natural language processing (NLP), image recognition, and reinforcement learning.

"Black-box" vs. "White-box" Models

In the context of interpretability, ML models are classified as "black-box" and "white-box" (interpretable) models. Compared to white-box models, black-box models are more complex regarding mathematical functions and can capture intricate patterns in data, often yielding higher accuracy. However, this increased complexity makes the predictions and internal workings of black-box models more challenging to explain and understand, particularly by domain experts. The term "black box" denotes their opaque decision-making process.[1] Notably, deep learning neural networks exemplify this, which automatically learn input features undergoing several layers of nonlinear transformations, rendering them noninterpretable to end-users.[2]

White-box models, on the other hand, are inherently interpretable. They rely on clear patterns, rules, or decision structures, balancing accuracy and explainability. Interpretability is not limited to white-box models, as black-box models can also be made interpretable through some external approaches. However, white-box models eliminate the need for additional interpretation models to gain insight into their decision-making process.[3] For instance, linear regression and decision tree models are straightforward to interpret for experts who can understand how inputs are mathematically transformed into outputs. The weight of the covariates in linear regression models can be used to assess the relative importance of the variables in the predictions made.[4]

Interpreting black-box models requires using a trained model as input and extracting information about the relationships learned by the model. This process is known as post hoc interpretability and occurs during the post hoc analysis stage through techniques such as local interpretable model-agnostic explanations (LIME) or Shapley additive explanations.[1] However, it is important to note that interpretations may not always perfectly

*From the Sporedata, Durham, NC; and †Onetomap Analytics, Department of Surgery, University of South Florida, Tampa, FL.

Nadayca Mateussi and Michael P. Rogers contributed equally to the study.

Disclosure: The authors declare that they have nothing to disclose.

SDC Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

Reprints: Paul C. Kuo, MD, MS, MBA, Department of Surgery, University of South Florida, 2 TGH Circle, Rm 7009, Tampa, FL 33606. E-mail: paulkuo@usf.edu.

Annals of Surgery Open (2024) 2:e423

Received: 6 July 2023; Accepted 20 March 2024

Published online 18 April 2024

DOI: 10.1097/AS9.0000000000000423

represent the relationships learned by a model. This can be especially challenging for complex black-box models such as neural networks, which encode nonlinear relationships less transparently.[5]

Interpretable Machine Learning or Outcomes Prediction

Interpretable ML can be defined as the extraction of relevant knowledge from an ML model about relationships contained in data or learned by the model. In other words, interpretable ML not only provides prediction to clinicians but also clarifies the contributing predictors, enhancing clinical decision-making by revealing factors affecting event probabilities.

Two fundamental tasks come to the forefront of outcome prediction: classification and regression, which are crucial for comprehending and forecasting the future states of structured data. Classification categorizes outcomes into discrete classes or labels, essential for applications such as disease diagnosis, assessing treatment success, and identifying high-risk patients. The choice between classification and regression depends on the nature of the outcome variable and the specific objectives of the predictive modeling process. For example, classification is applied when dealing with categorical outcome variables. Regression, on the other hand, estimates continuous values or scores, crucial for predicting survival rates, drug dosage optimization, and length of hospital stay, where the aim is to predict specific numerical values. In essence, classification and regression offer tailored solutions for diverse clinical predictive challenges.[6] Supplemental Table 1, http://links.lww.com/AOSO/A329 summarizes these ML models for structured outcomes data, providing insights into their applications, and highlighting their potential advantages and disadvantages, thus aiding in selecting the most suitable approach for specific predictive tasks.

Which Questions Can This Approach Address?

In a clinical context, interpretable ML is valuable for outcome prediction in both longitudinal and cross-sectional studies, where it leverages historical or current variable sets, respectively.[7] Also, the methods can be applied in clinical decision support systems (CDSS) used for enhancing information management and medical decision-making by offering faster, data-driven recommendations than traditional methods.[8,9] ML-powered AI methods are increasingly applied in the form of CDSSs to assist healthcare professionals in predicting patient outcomes[9,10] as well as to evaluate the implementation of CDSSs.[8]

Publicly Available Datasets

Interpretable ML methods can be performed using any dataset with a categorical outcome and multiple predictors. In clinical research, key datasets are frequently used, including:

- Economic Innovation Group Distressed Communities Index data (EIG DCI)[11]: encompasses regional demographics and socioeconomic data.
- NORC, Kaiser Permanente Research Bank (KP),[12] and National Trauma Data Bank (NTDB)[13]: Provide vital health outcome information.
- American Hospital Association Annual Survey database[14]: offers crucial insights into health facilities.
- Healthcare cost and utilization project[15]: a vital repository of United States hospital care data, alongside Optum and AllPayors.
- Other notable datasets include MedPar/Carrier, All of Us, N3C (National COVID Cohort Collaborative), metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP), National Surgical Quality Improvement Program (NSQIP), Veterans Affairs Surgical Quality Improvement Program (VASQIP), Surveillance, Epidemiology, and End Results—Medicare (SEER-medicare), society of thoracic surgeons national database (STS), agency for healthcare research and quality (AHRQ), veterans affairs corporate data warehouse (VA CDW), centers for Medicare and Medicaid Services (CMS), and Chesapeake regional information system for our patients (CRISP).

These datasets enable researchers to address various healthcare questions, improving quality and accessibility.

Interpretation

Achieving interpretability is crucial from several perspectives, with multiple methods available to accomplish it.[16] The first is the conventional method, which considers a feature's significance for the model as a whole. It then delves into how specific features affect predictions, considering conditional expectation curves, partial dependence plots, and compounded local effects. Also, one can consider surrogate trees, which use a short decision tree to approximate the underlying model for clearer understanding. Finally, explanations for personalized predictions, such as individual patients, look at how the value of a feature (predictor) for a given patient affects the prediction.[16,17] Supplemental Table 2, http://links.lww.com/AOSO/A329 outlines various interpretable ML approaches, offering a foundational overview of their key advantages and drawbacks. When selecting an ML model for a study, these insights are valuable for informed decision-making. However, it is essential to recognize that the specific context and dataset can influence the suitability of each method in use.

A case study in high-risk surgery utilized ML to enhance risk calculator predictions, and elucidate individual features and their contributions to mortality prediction.[18] After using a series of ML methods, including gradient boosting machine models, generalized linear models, random forest, and deep neural networks, the resultant modeling features were explored using a LIME approach.[19] LIME is a method that uses an interpretable model to explain the predictions of a regressor by approximating it locally. It modifies a single data sample by altering feature values to observe their impact on the outcome, providing local interpretability. As a result, the LIME approach allowed the identification of patient-specific factors influencing mortality and determining their weight in favor of or against patient survival, besides providing insights into personalized features and their impact on survival probabilities and model accuracy.[18]

To demonstrate the LIME technique, Supplemental Figures 1, http://links.lww.com/AOSO/A321 and Supplemental Figures 2, http://links.lww.com/AOSO/A322 use the BreastCancer database from the R package mlbench, containing 699 observations across 9 biopsy features. The plot in Supplemental Figure 1, http://links.lww.com/AOSO/A321 presents a predictor importance estimation. It indicates which variables the model considers the most important for cancer prediction. Features that display a dot on the right correspond to the most important variables for this model. The plot below corresponds to a binary classification model using the Ranger algorithm, a variant of the Random Forest. This model is employed here as an example to classify cancer as malignant or benign. In this analysis, features such as bare nuclei emerged as crucial predictors (Supplemental Figure 1, http://links.lww.com/AOSO/A321). LIME was then used to create an explainer object, trained just like the model and fit to the data, so new predictions for individuals could be made. Supplemental Figure 2, http://links.lww.com/AOSO/A322 displays the LIME explanation for 4 individuals, named cases 45, 537, 644, and 683 in this example. For each individual, the plot presents the original outcome (e.g., true for case 45 and false for case 537), the predicted outcome (true for case 45 and false for case 537), and the predicted probability (97.2% for case 45 and 100% for case 537). It also approximates how

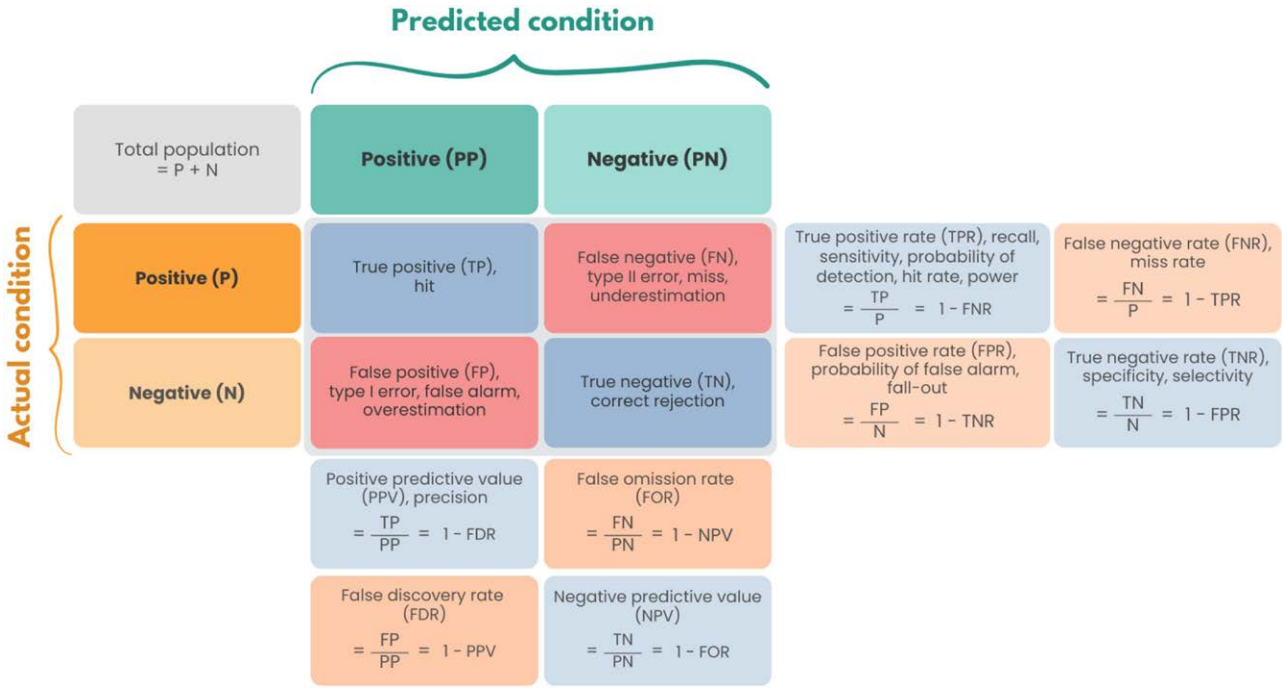

Figure 1. Confusion Matrix. A confusion matrix is a visual tool used to assess the performance of a classification model. It displays the number of correct and incorrect classifications for each class.

much and in which direction each variable contributed to a prediction for each individual. Blue bars represent the amount by which the variable increases the risk of the outcome, while red bars present the risk in the opposite direction. For example, the plot indicates that epithelial cell size, cell shape, margin adhesion, blood chromatin, and cell size were the variables contributing the most to model predictions for case 45.

Model Validation

In model validation, it is crucial to evaluate the predictive model's performance using a comprehensive set of metrics that offer insights into the model's overall discriminatory power and a more holistic evaluation of model effectiveness.[20] Apart from evaluating overall performance, considering bias in the training data becomes crucial. A model that performs well on training data but fails to generalize to new data may result from data biases. Cross-validation can ensure the model's applicability to new data. Therefore, this section discusses the importance of utilizing metrics that provide distinct perspectives on the model's performance.[21]

In classification metrics, accuracy is a fundamental gauge of overall model performance, indicating how many classes were correctly classified without distinction between categories. For instance, if a model accurately classifies 950 out of 1000 instances, it boasts an accuracy of 95%, reflecting its general correctness. However, accuracy can be problematic in unbalanced problems since it can yield misleadingly high scores by neglecting the minority class. On the other hand, precision is a common and essential metric that focuses on a model's precision in identifying a specific class. It is determined by the ratio of true positives to the total number of predicted positives, quantifying the model's adeptness at avoiding false positives. For example, if a model predicts 100 positive cases and correctly identifies 90, its precision would be 90%. Another essential metric is recall, which measures the ability of a model to correctly identify all relevant instances within a dataset. Moreover, the precision-recall curve is a valuable tool for evaluating the trade-off between precision and recall, providing insights into a model's performance, particularly in contexts of imbalanced datasets. However, since it provides a single numerical performance value, interpretation may be less intuitive. The F1 Score is a harmonious mean of precision and recall, offering a balanced perspective when their values need reconciliation. For instance, if precision and recall stand at 0.80 and 0.70, respectively, the F1 score calculates to 0.74, encapsulating the model's ability to balance classification tasks. However, it does not specify which type of errors are more frequent, making it less helpful in certain situations.

A confusion matrix is a visual tool used to assess the performance of a classification model. It displays the number of correct and incorrect classifications for each class, as shown in Figure 1. Sensitivity, or the true positive rate, quantifies the likelihood of correctly identifying positive instances, a metric of paramount importance in applications such as medical diagnostics, where the ability to detect cases accurately can be critical. Specificity, the true negative rate, assesses the model's accuracy in identifying negative instances, complementing sensitivity in providing a fuller picture of model performance.

The area under the receiver operating characteristic (AUROC) is a single metric ranging from 0 to 1, indicating the overall performance of a model in distinguishing positive and negative classes, with higher values indicating superior performance. The area under the precision-recall curve, particularly relevant for classification tasks in imbalanced datasets, offers a different perspective on performance, especially in cases with rare positive instances. Notably, the prevalence of the dependent variable within the dataset profoundly influences the interpretation of these metrics, underscoring the need to contextualize model performance against the backdrop of what could be achieved through random chance or using naive classifiers. For instance, when the prevalence of dependent variable is extremely low, a high AUROC may be misleading, as the model might achieve a high true negative rate by simply classifying the majority of instances as negative. Supplemental Figure 3, http://links.lww.com/AOSO/A323 shows an example of a receiver operating characteristic curve with an area under the curve of 0.9.

In the assessment of regression models, crucial metrics such as the R-squared and the adjusted R-squared play significant roles in evaluating the model's explanatory power regarding the variance in the data, with adjustments made for the number of

predictors to avoid overfitting. Regression metrics include the mean squared error (MSE), which measures predictive accuracy by calculating the average of squared prediction errors. Higher MSE indicates decreased model accuracy. Visual representations, such as plots of squared errors, furnish valuable insights in a graphical manner. The root mean squared error (RMSE), an essential variation of MSE, corrects for units, making the metric more interpretable. RMSE retains the same punishing characteristics for predictions far from actual values.

Mean absolute error (MAE) provides a gentler approach to measuring errors by calculating the average absolute differences between predicted and actual values. Unlike MSE and RMSE, MAE is less influenced by outliers, ensuring a more balanced assessment. Additionally, a plot of absolute errors can visually represent MAE in action, making it easier to understand and interpret. Finally, mean absolute percentage error expresses errors as a percentage, simplifying the evaluation of relative accuracy. Particularly valuable in financial forecasting, a lower mean absolute percentage error reflects more precise predictions. Displaying percentage errors graphically significantly aids in comprehending the model's accuracy.

Evaluating a model with comprehensive metrics and visuals enriches our understanding of performance, aiding informed decisions on effectiveness. Additionally, ensuring model fairness is an essential and ethical concern in developing and deploying ML models. One critical aspect of fairness is subgroup equity analysis, which focuses on identifying and addressing disparities in model performance across various subgroups within the dataset, such as gender, race, age, or other relevant attributes, suggesting biases that require correction. Addressing potential bias and discrimination in ML models is crucial since biased models can perpetuate and even exacerbate societal inequalities, which are morally and legally unacceptable,[22] and also lead to unfair outcomes and erode trust in AI systems, potentially resulting in reputational damage and legal consequences for organizations. Remediation may involve data re-sampling, algorithm adjustments, or feature and label modifications. Additionally, fairness-aware ML methods, including adversarial training and fairness constraints, aim to minimize bias by discouraging models from learning features associated with sensitive attributes.[23]

Additional Resources

As clinical tools and research increasingly incorporate ML methods, clinicians are urged to scrutinize these methods' accuracy and significance, much like traditional diagnostic or prognostic tools. The "user guide to medical literature" provides an ML overview and advice on assessing the published literature, outlining the use of ML-based tools to establish medical diagnoses.[21] Additionally, the "guide for making black box models explainable" intends to make complex ML models and their decisions more "interpretable" by exploring concepts of interpretability.[17] ML methods are subject to bias, such as missing data, patients not identified by algorithms, sample size, underestimation, and misclassification and measurement error when using electronic health record (EHR) data. Therefore, it is worth exploring the available literature on recognizing and potentially solving such biases.[22]

Furthermore, transparent and accurate reporting is vital in research and medical imaging. To achieve this, researchers have developed powerful tools and guidelines to uphold their work's reliability and quality. In this context, reporting frameworks are essential for researchers and practitioners, allowing them to create, validate, and communicate their work effectively. [24]PROBAST (Prediction model Risk Of Bias ASsessment Tool),[25] is a valuable tool for creating, validating, and improving multivariable prediction models used in diagnosis and prognosis. It helps determine the probability of individuals experiencing specific outcomes based on age, biomarkers, and symptoms. While invaluable for systematic reviews and evaluating prediction models, PROBAST is unsuitable for predictor discovery studies or comparative evaluations.[25]

The "Guidelines for developing and reporting machine learning predictive models in biomedical research"[26] were created through expert interviews and aimed to ensure transparent reporting of ML models. They offer comprehensive reporting guidelines, including a flowchart for data validation. These guidelines promote using ML in biomedical research to distinguish accurate findings from chance.[26]

Finally, the CLAIM sets standards to ensure transparent and high-quality communication about AI applications in medical imaging. It should be used throughout the research process, from project initiation to publication, to meet ethical and regulatory standards. While comprehensive, CLAIM necessitates adaptation to specific projects and evolving AI and regulatory landscapes, highlighting the importance of contextual flexibility in application.[24]

Natural Language Processing

NLP should be utilized for analyzing human-generated text or speech. Its healthcare applications include data extraction, classification, and sentiment analysis. In a clinical context, NLP can extract, classify, and automate information expressed in a natural language, converting the free text from clinical narratives into columns in a traditional dataset. Examples of medical free text include radiology and pathology reports, admission and discharge summaries, surgical reports, and reports containing laboratory test results. Therefore, NLP can be applied in disease recognition using EHR and patient-experienced or reported events to detect adverse events and postoperative complications from physician documentation, on studies aiming to identify toxicity, hate, or abuse, identification of failures in communication among healthcare teams, generation of summaries compliant with international reporting guidelines, supporting clinical trial recruitment, or assisting biomedical literature retrieval and analysis for therapy, facilitating knowledge discovery, and reducing manual search and review.[27,28]

Regarding sentiment analyses, NLP can be used when the goal is to infer whether the individuals writing the text were positive or negative about that topic, to assess healthcare providers' conditions (such as burnout) through EHR messages, and to evaluate the public perception of a certain condition, treatment, healthcare policy, or other health-related constructs.[29]

Table 1 provides an overview of NLP models that have contributed significantly to the field. These models have played a pivotal role in advancing the capabilities of text understanding, generation, and processing.

Publicly Available Datasets

Extraction and classification NLP methods can be performed using any dataset that might have free text containing medical information, such as medical information mart for intensive care-IV, Metabolic and MBSAQIP, and NSQIP, which contain health outcomes data. Sentiment analysis can be performed using any source of free text written by individuals in the population where researchers would like to evaluate positive or negative feelings, such as social media (e.g., Twitter), satisfaction surveys, and reviews. Table 2 presents significant datasets that play a crucial role in healthcare text analysis.

Interpretation

The application of NLP ranges from processing clinical notes to detecting phenotypes for cohort construction and detecting the occurrence of events pertinent to a medical visit (diagnoses, procedures, medications, etc.). There are several pathways to achieve this, from the use of elementary methods, such as pattern matching,[36] to ML models for named entity recognition.[37,38]

TABLE 1.

NLP Models and Frameworks and Their Characteristics

Model	Description
BERT (bidirectional encoder representations from transformers)[30]	A language model known for its groundbreaking approach to pretraining language representations. It excels in NLP tasks due to its ability to comprehend word and phrase context by considering surrounding words simultaneously, represented as high-dimensional vectors.
GPT (generative pre-trained transformer)[31]	It is a series of natural language processing models developed by OpenAI. These models are adept at generating human-like text and performing language-related tasks with high accuracy based on the Transformer architecture.
RoBERTa (A robustly optimized BERT pretraining approach)[32]	It is an NLP model by Facebook AI, building upon BERT's pretraining methods with enhancements like larger scale, longer training, and dynamic masking. It eliminates the Next Sentence Prediction task and refines text processing.
XLNet (generalized autoregressive pretraining for language understanding)[33]	It is another model using the Transformer architecture. It enhances text comprehension with permutation-based training and advanced techniques like two-stream self-attention and relative positional encoding.
ALBERT (A lite BERT for self-supervised learning of language representations)[34]	It was introduced by Google, aiming to improve parameter efficiency through shared parameter factorization. It replaces next sentence prediction with sentence order prediction and is pre-trained for various NLP tasks.
OpenNLP[35]	OpenNLP is a Java-based framework for NLP that allows developers to integrate language processing tasks into Java applications. It is widely used for text analysis, information extraction, and chatbot development. As an open-source project, OpenNLP is continuously improved by the NLP community.

AI indicates artificial intelligence; NLP, natural language processing.

TABLE 2.

NLP Datasets and Their Characteristics

Database	Description
National COVID cohort collaborative (N3C)	N3C is a US collaborative effort to hasten COVID-19 research. It collects de-identified clinical data from diverse sources, standardizes data representation, ensures secure data access, supports public health response efforts, and enables rapid progress in COVID-19 understanding.
PubMed abstracts	PubMed abstracts are used to train NLP models for tasks like text classification, named entity recognition, and information extraction. By automating literature reviews, tracking disease outbreaks, and extracting vital information, they provide structured data for analysis, hypothesis generation, and integration with other biomedical databases.
MIMIC-III and MIMIC-IV	MIMIC-III and MIMIC-IV are comprehensive clinical databases containing de-identified health records of ICU patients at Beth Israel Deaconess Medical Center. The datasets are a valuable resource for NLP research, predictive modeling, clinical decision support, and public health research.
Kaggle datasets	Kaggle is a platform for data science and machine learning. Kaggle Datasets offers NLP datasets for research, analysis, and machine learning. Users can explore, visualize, collaborate, and participate in NLP competitions, driving model training, algorithm benchmarking, application development, and NLP research.
BioNLP shared task datasets	BioNLP shared task datasets offer challenges to extract structured information from unstructured biomedical text in scientific articles, clinical records, and medical literature.
Social media and health forums	Data from social media platforms and health forums represent a valuable source of biomedical text data. These sources often contain patient experiences, health-related discussions, and medical information, making them relevant for biomedical NLP research and applications.

ICU indicates intensive care unit; MIMIC, medical information mart for intensive care; NLP, natural language processing.

Also, generative language models trained with human feedback capable of synthesizing clinical interpretations based on knowledge acquired from clinical texts can be applied.[39] While NLP interpretation is not one-dimensional, as each technique or model addresses unique challenges, several companion methods, including sensitivity analysis, specificity analysis, positive predictive value, and negative predictive value analyses, aid in interpreting NLP-generated data.[40]

An illustrative example involves the use of NLP and ML algorithms to analyze and classify self-reported narratives by patients with migraine and cluster headaches.[41] The study applied chi-square tests to calculate the key tokens for the analysis of lexical diversity discerning the 2 diagnostic categories (Supplemental Figure 4, http://links.lww.com/AOSO/A324) and ML to classify the text into the right diagnosis category, using positive predictive value and sensitivity to evaluate accuracy. The study also presents a lexicon-based sentiment analysis to assess the sentiment expressed in the dataset as positive or negative. The results showed that NLP could detect differences in lexical choices between the 2 groups, and ML algorithms have good potential to classify patients' descriptions of headache attacks, highlighting the relevance of NLP in clinical information extraction and the potential benefits of using digital techniques in analyzing patient-generated text.[41]

Additional Resources

With the availability of Python and other open-source tools, modern text analysis has become easily accessible, enabling individuals to explore textual data analysis and gain insights using NLP and computational linguistics algorithms. A starting point for NLP is learning data cleaning, statistical NLP, and deep learning with natural language and text samples.[42]

Image Recognition

Image recognition methods can extract and classify data from an image. In a clinical context, medical images are analyzed (e.g., magnetic resonance imaging and computerized tomography examinations), for the purposes of diagnosis, prognosis, and response to therapy predictions. Image recognition can be applied to establish a diagnosis of a given condition based on medical images, staging of a given condition based on medical images, screening of candidates for a determined surgery, prognosis based on examinations with a graphical output such as the electrocardiogram, and prognosis of a given condition.[43–45]

We provide an overview of cutting-edge image recognition methods used to extract and classify data from images in Table 3.

TABLE 3.

Image Recognition Methods

Model	Description
ResNet (residual networks)[46]	Deep networks that use residual connections to successfully train deeper networks are widely used in medical image analysis.
UNet[47]	Popular architecture for medical image segmentation is used to segment organs or structures in MRI and CT images.
DenseNet (densely connected networks)[48]	Networks that introduce dense connections between layers saving parameters and improving performance, applied in medical tasks.
Inception (GoogLeNet)[49]	Efficient architecture used in various computer vision applications in healthcare, including medical image analysis and diagnosis.
EfficientNet[50]	A family of architectures known for balancing performance and computational efficiency, used in various computer vision tasks in healthcare.
MobileNet[51]	Architectures designed for use in mobile devices, applied in telemedicine applications and mobile health apps.
SqueezeNet[52]	Efficient architecture successfully applied in medical tasks, useful in scenarios with limited computational resources.
Dilated convolutional networks[53]	Popular in medical image segmentation tasks, allowing for the capture of broader contextual information.
ConvNetX (facebook/meta)	ConvNetX by Facebook/Meta is an advanced image recognition model that excels in object detection, classification, and segmentation. It utilizes cutting-edge techniques in computer vision to solve complex image analysis problems.

CT indicates computed tomography; MRI, magnetic resonance imaging.

TABLE 4.

Image Recognition Datasets

Database	Description
CheXpert	Chest X-ray dataset with detailed clinical annotations.
MIMIC-CXR	MIMIC-related dataset with chest X-ray images and annotations.
NIH chest X-ray database	A collection of chest X-ray images used in pulmonary disease research.
ChestX-ray8	Dataset of chest X-ray images for the detection of various conditions.
RSNA pneumonia detection challenge	Challenge dataset for pneumonia detection in chest X-ray images.
The cancer imaging archive (TCIA)	A platform with multiple medical image datasets, including CT and MRI.
Medical ImageNet	Medical image dataset inspired by ImageNet.
BraTS (brain tumor segmentation)	Brain MRI dataset for brain tumor segmentation.

CT indicates computed tomography; MIMIC, medical information mart for intensive care; MRI, magnetic resonance imaging; RSNA, radiological society of North America.

Publicly Available Datasets

Image recognition methods can extract and classify data from any dataset with medical images, such as examination results (e.g., radiographs, magnetic resonance imaging, computerized tomography, electrocardiogram, mammography, and microscopy) or photographs, such as hospitals' EHR, the National Cancer Institute's Genomic Data Commons Data Portal, and the National Center for Tumor Diseases Biobank and University Medical Center Mannheim pathology archive.[54,55] Table 4 highlights the importance of these datasets in this field.

Interpretation

Automated analysis techniques are increasingly applicable because medical images can be digitized. Initially, low-level pixel processing (edge and line detector filters and region growing) and mathematical modeling (fitting lines, circles, and ellipses) were used to create compound rule-based systems for specific tasks. Then, supervised techniques that utilize training data to construct a system (active shape models, atlas approaches, feature extraction, and statistical classifiers) became more prevalent. However, these techniques rely on human researchers to extract discriminant features from the images.[56] Recently, deep learning methods, which use neural networks with multiple layers to convert input data into outputs, have become increasingly popular for image detection and classification. One of the most widely used algorithms for this purpose is neural networks. The neural network has several stages, including forward propagation, total error calculation, gradient (derivative) calculation, gradient checking, and updating weights. Hyperparameters, which include characteristics such as the number of layers, nodes, learning rate, weight values, bias or offset value, and hidden layers, can be modified by the modeler to improve the model's performance.[57]

Understanding the results of deep learning models is a crucial task that involves comprehending the model's architecture, hyperparameters, and input data. Evaluating the performance of a deep learning model is a primary goal, which is achieved by assessing its accuracy and error rate. To evaluate a deep learning model, several performance metrics, such as accuracy, precision, recall, and F1-score, are commonly used.

Unbalanced data, which occurs when instances in one class greatly exceed instances in another, is yet another significant issue in clinical settings that requires special attention. This leads to models biased toward the majority class and poor performance on the minority class. Techniques such as oversampling the minority class, undersampling the majority class, or class weighting can balance the classes to address this issue.

For example, convolutional neural networks (CNN) are used in the development of computer-aided diagnosis systems. Supplemental Figure 5, http://links.lww.com/AOSO/A325[57] illustrates a system capable of detecting radiological abnormalities by analyzing chest X-ray images. It provides the probability of an underlying condition and generates a heatmap, highlighting the regions in the image most indicative of the input pathology. The aim is to assist doctors and radiologists in the interpretation and classification of pulmonary diseases.[57] As the example presented in Supplemental Figure 5, http://links.lww.com/AOSO/A325, the diagnosis system takes an X-ray image as input. It outputs a heatmap highlighting the regions in the image most indicative of pathology, in this case, tuberculosis. The heatmap indicates the areas of the image that receive more attention. In this case, the cavitary lesion is indicated by small arrows, while the larger arrows highlight airspace opacities. These abnormalities are accurately localized by the base-CNNs, as depicted by the red-colored areas.

Another example is fully automated algorithms for segmenting the abdomen from computerized tomography scans using CNN to quantify body composition, implying better information for individual care and decreasing, if not excluding, time as a limiting factor in studying body composition metrics and their influence on many clinical outcomes (Supplemental Figure 6, http://links.lww.com/AOSO/A326).[58] The images demonstrate body composition segmentation through deep learning, depicting various body compartments such as subcutaneous adipose tissue, muscle, visceral adipose tissue, visceral organs, and bone. They compare segmentation from a semi-automated method and U-Net predictions, with arrows marking areas of disagreement.[58]

Deep learning algorithms using cropped images from angiographies can be used for more accurate predictions. For example, in a comparison among the ability of humans, angiographic parameters, and deep learning to predict the lesion that would be responsible for a future myocardial infarction in a population of patients with nonsignificant coronary artery disease at baseline, deep learning outperformed human visual assessment and established angiographic parameters in the prediction of future culprit lesions (Supplemental Figure 7, http://links.lww.com/AOSO/A327).[59] The performance of the deep learning model and the models based on angiographic parameters such as diameter stenosis, area stenosis, and quantitative flow ratio were evaluated using receiver operating characteristic curves. The corresponding area under the curve values were obtained for each model.[59]

Additional Resources

Successful neural networks for image recognition typically comprise multiple analysis layers. To facilitate an understanding of how these neural networks work, an analogy of written language can be applied.[60] Also, a comprehensive overview of the methods belonging to the category of spectral-spatial classification, along with guidelines for the design of new approaches, can help to clarify the mechanism behind existing classification methods.

Reinforcement Learning

Reinforcement learning (RL), as a branch of ML, aims to determine a series of steps that maximize the likelihood of reaching a specific objective. It has proven to be effective in various healthcare domains, particularly in situations where sequential decision-making is involved, such as diagnosing patients or establishing treatment regimens. RL addresses these challenges by employing a trial-and-error learning process, mimicking human learning behavior.[61]

RL involves a learning agent interacting over time with its environment and making decisions using a policy to optimize a specified reward function. Beyond the agent and the environment, an RL system comprises 4 main sub-elements: (1) a policy that defines the agent's behavior at a given time; (2) a reward signal that determines the immediate, intrinsic desirability of environmental states and defines the goal of an RL problem; (3) a value function that specifies the long-term desirability of environmental states; and, optionally, (4) a model of the environment that mimics the behavior of the environment or allows for inferences about its behavior.[62]

In a clinical context, RL can be applied in optimizing treatment strategies, clinical decision-making, and risk-based screening policies.[63–65]

Publicly Available Datasets

RL methods can be performed using any dataset that might have EHR data, such as medical information mart for intensive care-IV and the N3C.[66,67]

Interpretation

To interpret RL methods, it is crucial to understand how the different components interact and how they are used to train the agent. There are various approaches to RL, including value-based, policy-based, and actor-critic methods, and each has its own strengths and weaknesses.[62]

For instance, Yu et al.[68] compared the dissimilarities between Soft Actor-Critic (SAC) and conventional Actor-Critic algorithms concerning their effectiveness in dealing with decision-making challenges related to ventilation and sedative dosing in intensive care units. In the SAC approach, an actor suggests the most favorable action (policy optimization). At the same time, a critic evaluates the qualities of the actions by computing the suggested action's quality (assessing the quality of the action). The findings indicated that the SAC algorithm not only focuses on long-term patient recovery but also minimizes the divergence from the strategy employed by medical professionals, leading to enhanced therapeutic outcomes.[68]

In another example, the VentAI algorithm, a computational model using RL, was developed to suggest a dynamically optimized mechanical ventilation regime for critically ill patients, including the three-dimension settings: ideal body weight-adjusted tidal volume, positive end-expiratory pressure levels, and the fraction of inspired oxygen (Supplemental Figure 8, http://links.lww.com/AOSO/A328).[69] To evaluate the differences in performance conservatively, the study compared the 90% lower bound of the VentAI performance return with the 90% upper bound of the clinicians. The best dynamically chosen mechanical ventilation regime by the VentAI algorithm resulted in a 93.64 estimated performance return in validation and 91.98 in the testing dataset, respectively. This represents an improvement of 42.6%, compared to the best performance of the clinicians, based on the learned model, and an improvement of 22.6%, compared to observable clinician behavior.[69]

Additional Resources

For a clear and straightforward view of the main ideas of RL, an introduction aimed at readers in all related disciplines would be of great help.[62] Then, the implementation of RL models can be studied and performed both using R[56] and Python.[34]

DISCUSSION

Interpretable predictive ML models, NLP, image recognition, and RL are core ML methodologies that underlie many of the AI methodologies that will drive the future of clinical medicine and surgery. End users must be well-versed in the strengths and weaknesses of these tools as they are applied to patient care in the future. In this review, we have described publicly available datasets that can be used with these models, outline interpretation of results, and, finally, provide references for in-depth information about each analytical framework. It is our hope that future reviews will focus on a case-driven approach to accelerate the adoption of the most recent developments in artificial intelligence into clinical use.

REFERENCES

1. Murdoch WJ, Singh C, Kumbier K, et al. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci USA*. 2019;116:22071–22080.
2. Learning Deep Architectures for AI | Now Foundations and Trends books. IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/8187120. Accessed November 1, 2023.
3. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Journals & Magazine | IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/8882211. Accessed November 1, 2023.
4. Chakraborty D, Başağaoğlu H, Winterle J. Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Syst Appl*. 2021;170:114498.
5. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1:206–215.
6. Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2:160.
7. Cowling TE, Cromwell DA, Bellot A, et al. Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably. *J Clin Epidemiol*. 2021;133:43–52.
8. Wasylewicz ATM, Scheepers-Hoeks AMJW. Clinical Decision Support Systems. In: Kubben P, Dumontier M, Dekker A, eds. *Fundamentals of*

Clinical Data Science. Springer; 2019. Available at: http://www.ncbi.nlm.nih.gov/books/NBK543516/. Accessed November 1, 2023.
9. Amann J, Vetter D, Blomberg SN, et al. To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health*. 2022;1:e0000016.
10. Marafino BJ, Schuler A, Liu VX, et al. Predicting preventable hospital readmissions with causal machine learning. *Health Serv Res*. 2020;55:993–1002.
11. Kesler P. Distressed communities. economic innovation group. Available at: https://eig.org/distressed-communities/. Accessed September 19, 2022.
12. Kaiser Permanente Research Bank. Kaiser Permanente. Kaiser Permanente Research Bank. Available at: https://researchbank.kaiserpermanente.org/. Accessed November 2, 2023.
13. NTDB National Trauma Data Bank (TQP Trauma Quality Program). GitHub. Available at: https://github.com/onetomapanalytics/Meta_Data/wiki/NTDB---National-Trauma-Data-Bank-(TQP---Trauma-Quality-Program). Accessed March 15, 2023.
14. AHA Annual Survey DatabaseTM. AHA Data. Available at: https://www.ahadata.com/aha-annual-survey-database. Accessed March 15, 2023.
15. Healthcare Cost and Utilization Project (HCUP). Available at: https://www.ahrq.gov/data/hcup/index.html. Accessed November 2, 2023.
16. R Core Team. Introduction to IML: Interpretable Machine Learning in R. Available at: https://cran.r-project.org/web/packages/iml/vignettes/intro.html. Accessed March 16, 2023.
17. Molnar C. Chapter 6 Model-Agnostic Methods. In: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed.; 2023.
18. Rogers MP, Janjua H, Fishberger G, et al. A machine learning approach to high-risk cardiac surgery risk scoring. *J Card Surg*. 2022;37:4612–4620.
19. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016:1135–1144.
20. Leisch F, Dimitriadou E. mlbench: Machine Learning Benchmark Problems . *R package version 2.1-3.1*; 2021.
21. Mathai N, Chen Y, Kirchmair J. Validation strategies for target prediction methods. *Brief Bioinform*. 2020;21:791–802.
22. Dankers FJWM, Traverso A, Wee L, van Kuijk SMJ. Prediction Modeling Methodology. In: Kubben P, Dumontier M, Dekker A, eds. *Fundamentals of Clinical Data Science*. Springer; 2019. Available at: http://www.ncbi.nlm.nih.gov/books/NBK543534/. Accessed November 2, 2023.
23. Do H, Nandi S, Putzel P, et al. A joint fairness model with applications to risk predictions for underrepresented populations. *Biometrics*. 2023;79:826–840.
24. Mongan J, Moy L, Charles E, et al. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:0200029.
25. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51–58.
26. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323.
27. Liu F, Weng C, Yu H. Natural Language Processing, Electronic Health Records, and Clinical Research. In: Richesson R, Andrews J, eds. *Clinical Research Informatics*. Health Informatics. Springer;2012.
28. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc*. 2005;12:448–457.
29. Baxter SL, Saseendrakumar BR, Cheung M, et al. Association of electronic health record inbasket message characteristics with physician burnout. *JAMA Netw Open*. 2022;5:e2244363–e2244363.
30. Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Prepr*. 2018.
31. Yenduri G, M R, G CS, et al. Generative pre-trained transformer: a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions; 2023.
32. Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach; 2019.
33. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding; 2020.
34. Bilgin E. Mastering Reinforcement Learning with Python: build next-generation, self-learning models using reinforcement learning techniques and best practices. Packt Publishing Ltd; 2020.
35. Apache OpenNLP. Available at: https://opennlp.apache.org/. Accessed November 2, 2023.
36. Couto FM, Couto FM. *Text processing*. Data Text Process Health Life Sci. 2019:45–60.
37. Ward PJ, Young AM, Slavova S, et al. Deep neural networks for fine-grained surveillance of overdose mortality. *Am J Epidemiol*. 2023;192:257–266.
38. Oliwa T, Maron SB, Chase LM, et al. Obtaining knowledge in pathology reports through a natural language processing approach with classification, named-entity recognition, and relation-extraction heuristics. *JCO Clin Cancer Inform*. 2019;3:1–8.
39. Zack T, Dhaliwal G, Geha R, et al. A clinical reasoning-encoded case library developed through natural language processing. *J Gen Intern Med*. 2023;38:5–11.
40. Bice N, Kirby N, Li R, et al. A sensitivity analysis of probability maps in deep-learning-based anatomical segmentation. *J Appl Clin Med Phys*. 2021;22:105–119.
41. Vandenbussche N, Van Hee C, Hoste V, et al. Using natural language processing to automatically classify written self-reported narratives by patients with migraine or cluster headache. *J Headache Pain*. 2022;23:129.
42. Srinivasa-Desikan B. Natural Language Processing and Computational Linguistics: a Practical Guide to Text Analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd; 2018.
43. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211–2223.
44. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394:861–867.
45. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci USA*. 2018;115:E2970–E2979.
46. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition; 2015.
47. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2015:234–241.
48. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks; 2018.
49. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions; 2014.
50. Tan M, Le QV. EfficientNet: Rethinking Model scaling for convolutional neural networks; 2020.
51. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications; 2017.
52. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size; 2016.
53. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions; 2016.
54. Access Data. NCI Genomic Data Commons. Available at: https://gdc.cancer.gov/access-data. Accessed November 2, 2023.
55. Archives of pathology and clinical research. HSPI. Available at: https://www.pathologyresjournal.com/. Accessed November 2, 2023.
56. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
57. Computer-aided diagnostic for classifying chest X-ray images using deep ensemble learning. BMC Medical Imaging|Full Text. Available at: https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-022-00904-4. Accessed November 2, 2023.
58. Weston AD, Korfiatis P, Kline TL, et al. Automated abdominal segmentation of CT scans for body composition analysis using deep learning. *Radiology*. 2019;290:669–679.
59. Mahendiran T, Thanou D, Senouf O, et al. Deep learning-based prediction of future myocardial infarction using invasive coronary angiography: a feasibility study. *Open Heart*. 2023;10:e002237.
60. Carin L, Pencina MJ. On deep learning for medical image analysis. *JAMA*. 2018;320:1192–1193.
61. Datta S, Li Y, Ruppert MM, et al. Reinforcement learning in surgery. *Surgery*. 2021;170:329–332.

62. Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT press; 2018.
63. Komorowski M, Celi LA, Badawi O, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24:1716–1720.
64. Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Med Inform Decis Mak*. 2019;19:111–120.
65. Yala A, Mikhael PG, Lehman C, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat Med*. 2022;28:136–143.
66. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. Mimic-iv. PhysioNet Available Online Httpsphysionet Orgcontentmimiciv10; 2020. Accessed August 23, 2021.
67. National COVID Cohort Collaborative (N3C). National Center for Advancing Translational Sciences. https://ncats.nih.gov/n3c. Accessed April 28, 2023.
68. Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Med Inform Decis Mak*. 2020;20:124.
69. Peine A, Hallawa A, Bickenbach J, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *Npj Digit Med*. 2021;4:1–12.