Deep Science Publishing
https://doi.org/10.70593/978-81-988918-5-3_3

DeepScience
Open Access Books

# Chapter 3: Developing machine learning algorithms for improved diagnosis and prognosis

## 3.1. Introduction

Machine learning research has advanced quickly within the last decade, utilizing the availability of large data sets and data storage and processing advancements to develop state-of-the-art algorithms that rival and often outperform traditional statistical methods. Yet, while ML has been successfully applied in many fields of research from varying disciplines, including neuroscience, politics, criminology, ecology, and remote sensing, among many others, few biomedical researchers have explored the use of ML for improved disease diagnosis and prognosis. This is surprising, considering the fascinating and complex nature of disease, as well as the generalizability and flexibility of ML algorithms. In this chapter, we elucidate some key ML concepts in an effort to empower and excite the biomedical researcher community to further investigate the potential of utilizing ML techniques in diagnostics and prognostics (Bui & Zorzi, 2011; Belle et al., 2015; Ristevski & Chen, 2018).

As health care professionals know, disease diagnosis and prognosis is at the heart of most medical care. Accurate diagnoses guide treatment plans. Disease prognosis was traditionally based on the time to event estimates from curves or models built from a few sample characteristics. However, these techniques are limited, mostly due to the assumption of proportional hazards among different sample characteristics or the subjectivity of sample group prioritization and estimated cutoff values for grouping. For this reason, modern-day prognoses have increasingly turned to immune response – based gene expression signatures as observations from the extracellular components of the tumor microenvironment by high-throughput sequencing are shown to be more powerful than the traditional clinical and pathological variables.

The numerous benefits of accelerating the integration of ML into clinical operations encompass robust model assessment and ready availability of models in real-world clinical use. Nevertheless, these tools alone cannot improve the safety, quality, and efficiency of care without appropriate development, deployment, and translation into the clinical workflow of healthcare professionals. Current efforts have primarily focused on prediction models in evaluating prognosis of disease and clinical outcomes (Zhang & Liu, 2010; Vaquero & Rodero-Merino, 2014).

**Fig 3.1:** Developing Machine Learning Algorithms for Improved Diagnosis and Prognosis

### 3.1.1. Background and Significance

Machine learning is transforming medicine across subspecialties and teaching the next generation of providers about healthcare technology, supporting their objective of improving patient safety, outcomes, and experience while also addressing the value of care provided. A large enterprise ML effort involving a cross-departmental collaboration has been embarked upon to advance medicine and to improve patient care through ML products, some in routine operational use, not just proof-of-concept projects, with the approval of the Clinical Oversight Body for projects that impact patient care. These efforts facilitate the integration of data science and ML into hospital operations with the

potential to develop timely, cost-effective, and value-added ML predictive tools centered on diagnosis and prognosis of disease to automate the routine and support the providers across all.

Over the previous decades, the availability of massive volumes of increasing rich data in biomedical informatics that includes longitudinal electronic health records, clinical omics research data via high-throughput technologies is increasing exponentially. ML paradigms have proven to be powerful tools for data-driven knowledge discovery, biomarker identification of disease, diagnosis, prediction of outcome/prognosis, rehabilitation, and treatment selection across all medical and healthcare domains while addressing both emerging and unmet clinical and biomedical needs.

### 3.2. Background

Machine Learning (ML) has transformed many domains since its inception, allowing for the automation of processes, uncovering and modeling patterns in data, and allowing expert-level provision of services. In Healthcare, these capabilities are leveraged specifically to improve services such as diagnosis, prognosis, and personalization. Modern computer-aided algorithms rely strongly on pattern recognition engines driven by ML, analyzing often massive health datasets. These health datasets typically include images of tissues or organs, clinicians' notes, medical histories, genomics, proteomics, and many more sources. Such huge datasets bring with them both challenges of management and challenges of pattern exploration.

The unique societal vision of healthcare to improve and maintain the well-being of all citizens mandates trust in the systems that deliver this service. As a unique domain with its own science-based methodology and structure of rewards and penalties, healthcare now fosters successful collaboration for interlocking areas. Such collaboration among computer science research, and clinical specialty research and industrialization, tends to focus and broaden competing engineering solutions to ML tool processes, architectures, and evaluation. Such dynamics hold the promise that healthcare will lead in the inception and retention of a trusted paradigm, rather than wildly attempting automatic direct ML replacement of people that leads in so many other fields.

Implementation of computer-aided diagnosis systems began in earnest when computing resources became affordable. With the first commercial systems in the early 1990s, clinical impact was modest, but both academia and industry pursued the potential of ML-based research tools to discover treatment solutions unavailable to human experts using classical methods. After the initial hectic rush, a clarion call sounded around the turn of the millennium – external validation would be required for selection or endorsement of any final “diagnostic” ML tool, before doctors would trust any such assistance. Trust

here means decades-long familiarity that ML-based systems, provided through clinical trials, achieved superior or similar performance to clinical judgments. During the past two decades, trust has slowly increased through the first successful clinical candidates in radiology and dermatology, and has expanded further through informatics to prediction in other clinical specialties.

### 3.2.1. History of Machine Learning in Healthcare

Machine Learning (ML) is an artificial intelligence (AI) area focused on the creation of algorithms that learn in an incremental way from a stream of data. ML has been used since the mid-1950s for various applications, and its use in different fields have grown every year, becoming a powerful tool to solve complex problems. Examples of some key historical events in ML development include the following. In 1957, Frank Rosenblatt proposed the Perception model, the first neural network implementation capable of classifying data in two classes. In 1967, the nearest-neighbors algorithm was created, which associates incoming objects to the closest object in a previously labeled training dataset. In 1970, the Backpropagation method was proposed for training multilayer perceptrons neural networks. In 1986, a model for backpropagation was published, which brought back interest in neural networks. In 1996, a dataset for training and evaluating visual object detection systems was created. In 2012, a deep convolutional neural network that created a breakthrough in image recognition was proposed. The increasing availability of high-performance computing platforms and large-scale labeled databases from web services in recent years have enabled the huge growth of neural networks as a powerful ML technique.

The application of several ML techniques has become a widely used tool for the healthcare area in recent years. To create ML applications, two main challenges and roadblocks commonly encountered are the size of the healthcare datasets and their high dimensionality. Machine learning techniques have been used for developing algorithms that support the assessment, diagnosis, and monitoring of patients and possible maladies; support trained doctors with insights into a patient's diagnosis based on the patient's clinical information; and prognose the risk of developing certain diseases, as well as the risk of mortality.

### 3.2.2. Current Trends in Medical Diagnostics

Translational research is a relatively new paradigm that was crafted to reduce the time from basic scientific discovery to the implementation of new medical technologies. The idea of translating research findings to healthcare solutions is still novel and few medical industries have seen much impact from the actual translation process. Subsequently, few

clinical experts have been exposed to the incremental implementation of new technologies discovered through translational research; this leads to hesitancy on the part of clinicians when faced with applying results generated by this new process. However, the potential impact is unimaginable; with rudimentary technologies such as sensor networks, wireless networking, and machine learning, translational research can drastically reduce the time it takes to recognize a particular disease, monitor its progress, and recognize that it is contained or becoming debilitating.

Currently, presented diagnostic methods are starting to garner interest from the healthcare community due to the availability of the necessary technological platforms. The first interesting trend is a move towards collecting rich clinical data collections. Innovators are starting to think differently and ask: "Why not build applications that take advantage of the input from the patient", "Why not have skilled annotators confirm and augment the data", and "Why not build and maintain a rich data set of ground truth disease states linking to data for the last few years?" Because tools are becoming available, the door is gradually being opened to gather rich, data-heavy diagnostic systems at considerable efficiencies and cost. Electronic health records, direct information input systems, incentives for data collection, specifying disease treatments on a broad rather than targeted scale, increased access to expectant mothers and infants, the encouragement of patient self-monitoring portals, the entrepreneurial focus on health technologies, increasingly lower-cost high-velocity sensors and processing platforms, and incentives to lower healthcare costs are all contributing to this trend.\

### 3.3. Machine Learning Fundamentals

Machine learning is a branch of artificial intelligence that studies the design and development of algorithms that enable computers to perform tasks that typically require human intelligence. These algorithms use machine learning models to evaluate different possibilities in the internal search for the best option, which usually involves a function approximation for error minimization. Machine Learning draws skills from statistics, computer science, neural science, inferring theory, algorithm complexity theory, optimization, knowledge representation, and decision theory. Statistical analysis aids to measure uncertainty, while computational system attention is needed in the implementation of realistic problems. Neural or cognitive science contributes to a better understanding of the learning process. Decision theory helps to model types of decision-making problems, and optimization study seeks the most efficient solution to the problems.

Types of Machine Learning

Machine Learning has been classically divided into three types, according to the kind of supervision retrieved from the problem. In supervised learning, the training data is made up of pairs, where the first element is the input and the second is the expected output or called label. Then, the goal is to learn a function from the training data so that the expected output corresponds to the input for all elements in the dataset. In supervised classification, all outputs are discrete and the training data must be representative enough so that the function is capable of describing the output value for any input in the feature space. In supervised regression, outputs may take continuous values. In unsupervised learning, the training data is made up of only inputs, with no expected outputs. Then, the goal is to find patterns in the data, and to imply different data internal structures. Finally, in reinforcement learning, a model learns to interact with an environment to optimize its outcome minimization or maximization. It can be seen as the generalization of supervised learning to some cases where not all action inputs are available during the training stage, as the agent is trying to learn the best model to choose the best possible action for each situation.


**Fig 3.2:** Machine Learning Fundamentals

### 3.3.1. Types of Machine Learning

Machine Learning is a method based on a computer training method capable of self-discovery signals, rules, and predictive patterns derived from large amounts of data via

intelligent algorithms. Supervised learning is the technique that consists of creating a prediction model from a sample of records, in which the variables that have to be predicted are known with certainty. The values of the variables that have to be predicted are called labels, thus creating a two-fold problem: a classification problem, when the number of possible labels is finite, and a regression problem, when the number of possible labels is uncountable. In the case of classification problems, the labels may be unbalanced; that is, the number of instances of some labels is much larger than the number of instances of the remaining labels, which is a frequent problem in many domains. Besides being unbalanced, classes can be overlapping or not. The task of the created model is to predict the most probable label associated with a new data instance.

In unsupervised learning, the identification of the predictive structure or model of the data set is not guided by labeled instances like in supervised learning. The inference process in unsupervised learning tries to discover a mapping that describes the structure of the observed data. Semi-supervised learning is a middle ground between supervised and unsupervised machine learning tasks, which is increasingly commonly used in practice because of the small number of labeled instances. Typically, in semi-supervised learning, a small number of labeled records are accompanied by a large number of unlabeled records when both records are the same. The purpose of semi-supervised learning is to better predict the unlabeled instances.

### 3.3.2. Key Algorithms and Techniques

Machine learning applications can be distinguished by the way knowledge is produced. Supervised learning has a teacher who exerts control by providing an understanding of how to recognize instances of a signal class or of several classes. These supervise existing models that have been previously trained, providing the labels needed to have a good recognition and to prevent overfitting. Unsupervised learning does not have a teacher telling it what features are to be extracted. It examines the instances available and tries to detect the underlying structure, establishing an internal model. It is believed that both supervised and unsupervised learning have their roots in behavioristic psychology. Semi-supervised learning is a third type of learning, albeit less frequently used, which has the features of both previous types of learning. For example, there are situations in which part of the data comes labeled with class information while most of the data is unlabeled.

In supervised learning, algorithms learn a model from known examples that consist of instances of input data associated with the correct output. This model is then used to predict the output for unknown instances. There are two formats for supervised learning: Classification is used when the output consists of labels taken from a finite set, whereas regression is when the output is a real value. In both cases, it is assumed that the output

is a function of the input processed by a model, so determining which model produces the best results is the aim of supervised learning methods. A model is generally learned through a training process driven by a learning algorithm that uses a training set.

### 3.3.3. Data Preprocessing Methods

Introduction of any system begins with preprocessing. In cellular and molecular biology, data involves structures and relationships that are frequently complex, and that we do not yet fully understand. In healthcare expenses, the mass amounts of data are of various data types and formats. Big data usually involves three Vs: the size of data, the abundance of variety that must be exploited, and the velocity of data. The challenge is found in drawing strong inferences despite the three Vs. The three Vs can heavily increase noise in high-volume, high-dimensional datasets. Because of the resulting high uncertainty in big data trends to limit interference power, preprocessing is essential for obtaining satisfactory results in almost all machine learning studies. The purposes of preprocessing are numerous, and include but are not limited to: (1) reducing the number of features in data, (2) summarizing important data-related relationships for the analysis, making them easier to understand, and visualizing them (3) organizing the unstructured inputs before applying existing or newly-developed machine learning algorithms, (4) accelerating the execution time and improving resources used by the machine learning algorithms, and (5) removing bad quality data points, and compensating for missing values in datasets. When the effects of the three Vs are significantly reduced through preprocessing, it becomes feasible to use machine learning algorithms to extract useful information from datasets. This useful information can consequently help with objective specification and optimization, and automated decision making in healthcare systems. The data preprocessing stage is an unavoidable step in any real-world machine learning pipeline.

## 3.4. Dataset Acquisition

The power, robustness, and adequacy of machine learning algorithms depend on the data characteristics. The accuracy of any diagnostic tool is dependent upon the type and quality of the training data. In this chapter, different sources of medical data and their characteristics are discussed and how this data is acquired with quality and integrity is illustrated. Medical data is heterogeneous, consisting of several data characteristics such as signs and symptoms. There are many types of medical data such as electronic medical records, signal data, image data, various types of medical profiling data, mental health-related data, cardiac data, and many others. This data is widely accessible, and many of these datasets are made publicly available. These datasets can be analyzed to extract

various patterns, which can be helpful for the development of various predictive tools such as decision support systems. There are several works related to healthcare that have been developed around these publicly available datasets to support or assist both the medical professionals as well as patients, which perform well and some of which show accuracy comparable with medical experts.

Data integrity means maintaining and assuring the accuracy and consistency of the data over its entire life-cycle and data quality means the data is exactly what the user is looking for – and serves its intended purpose effectively. Quality determination is a crucial process involved when mining data from various sources, as a dataset with poor quality can lead to severe consequences for the research study. Hence, it is critical to determine data quality before using any datasets. Quality checks can ensure the data is accurate, complete, consistent, timely, trustworthy, and needed for the analysis. In today's world, the explosion of available data has revolutionized the methods by which health services are organized with advances in information technology, leading to a wealth of available health-related data.

### 3.4.1. Sources of Medical Data

In this section, we will discuss the major data sources for machine learning research and how these data sources have been instrumental in making various machine learning algorithms and models. The ability of a machine learning algorithm and model to learn underlying patterns in unseen data is dependent greatly on the data quality and availability. Thus, while these publicly available data repositories might not help with the specific fine-tuning of a specific algorithm for a specific domain, their presence in the academic landscape of machine learning enables the general trend of improvement and advancement in this field. While creating a machine learning model for clinical application, one would encounter the challenge of data availability, given the strict regulatory guidelines and policies around personal medical data usage. In this section, we will focus on the publicly available sources out there and their descriptions.

For supervised learning in the clinical or medical domain, usage of labeled data is the cornerstone. The standard pipeline is considered to not just train a model solely once it has proven useful, but continue to return to use it in other circumstances and when it under-performs in that domain, rather than starting from scratch. The knowledge thus gathered through data from years of research on a specific area will make algorithms for that domain more applicable to problems, overcome any semantic gap between technical fundamentals of the model and its local application, and build confidence in its results by trying to solve similar problems and driving research in that direction. As more academic datasets become available, algorithms on those datasets get better refined and presented.

### 3.4.2. Data Quality and Integrity

Exponential advancements in medicine, photonics, digital imaging and computing have made medical data acquisition a rapid and fruitful process. Capable of in-depth investigation of structural and functional attributes, data can be acquired in multiple domains, including genomics, transcriptomics, proteomics, metabolomics, morphomics, radiomics and phenomics. Voluminous datasets in the form of electronic health records, whole slide images, and multi-omics have catalyzed the adoption of AI and ML in clinical investigation.

Despite the increasing availability of medical data, a valid objective trained model is difficult to achieve. One of the contributing factors could be the quality and integrity of datasets which is not frequently recognized. Poor quality datasets in terms of class imbalance, ground truth incorrect labels and overfitting duplicates could result in frequent model overfitting or performance deterioration during validation phase. For example, missing clinical outcomes or incorrect labels when building models for early detection of neurodevelopmental disorders could lead to misdiagnosis. Overfitting duplicates, that tend to overrepresent a certain class, could negatively influence a ML model trained for class prediction and can err in unseen images, leading to negative downstream effects. It has been shown, for various disease LBP-MOs, that radiogenomics are often impractical due to the considerable overlap in the distributions of several radiomic features in selected areas. Furthermore, lack of balance or presence of artifacts could also impair the training of the ML models. Inconsistencies in datasets collected from academic centers versus community healthcare centers are also a factor that needs to be given due consideration as they may result in a 'ML bias' that could affect health care equity.

## 3.5. Feature Selection and Engineering

Feature selection refers to the process of selecting a subset of relevant features for use in model construction. Feature selection helps to narrow down data dimensions which leads to comparatively easier optimization, faster computations, reduced storage space, and less data to examine. In this regard, feature selection can be seen as a process for enhancing the overall accuracy of a developed model by reducing overfitting using pruning techniques. Feature selection can also be employed by algorithms operating on the instance level. An informative feature set typically contains low-dimensional yet meaningful descriptors of relevant parameters. Selecting the wrong features, or having too many features, will reduce the predictive accuracy of the model and increase the computation time. In predictive modeling, feature selection methods can broadly be classified into three categories: Wrapper methods, Embedded methods, Filter methods. Wrapper methods are often considered as being the best all-around feature selection

strategy because they usually result in the most predictive models. However, they also are the most costly feature selection strategy. Wrapper methods repeatedly use the algorithms that model the data to actually evaluate the feature sets.

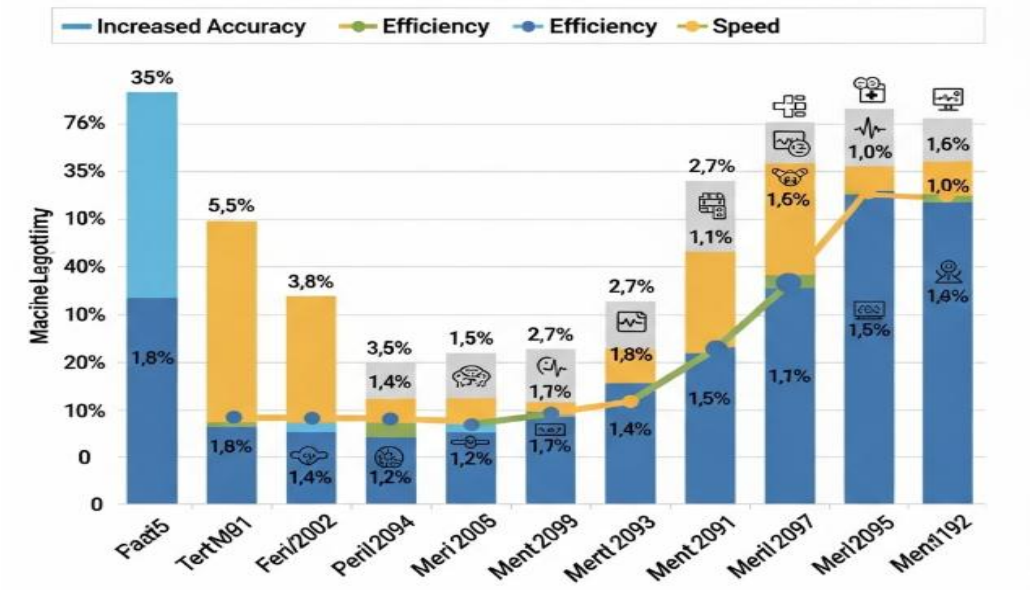


**Fig :** Machine Learning Algorithms for Improved Diagnosis and Prognosis

There are a number of widely-utilized techniques for feature engineering including Univariate feature selection, Recursive Feature Elimination, Feature Importance from Tree-based estimators, SelectFromModel using feature importances, SelectKBest using statistical tests, Parent-Child feature selection with clustering, LASSO regularization, ENET Model, Ridge regularization. Expert knowledge about the datasets is often insufficient and machine learning is frequently required to identify those measurements and features that can be used to build efficient classification and regression models.

### 3.5.1. Importance of Feature Selection

Feature selection is an important component of model building in such a pipeline. There are two key reasons why we would want to reduce the number of features in the model. The first is for model accuracy and the second is for model interpretability. High dimensional datasets can lead to overfitting as the classifier will be fitted to noise in the dataset rather than the signal. An overfit model may source high accuracies in the training set, but it is unlikely that the classifier will generalize to unseen datasets. Unlike the old adage “more is better,” more features is not always the best approach, and, in fact, it can lead to the opposite outcome. Additionally, many machine learning

algorithms are sensitive to the number of features and can take more time if more features are included due to the increased complexity.

Many machine learning algorithms have difficulty with high dimensional spaces given that their decision boundaries and risk surfaces can be complex in shape. Consequently, the challenges imposed require exponentially more data compared to lower dimensional datasets to identify the underlying signal. Although supervised algorithms can have the best predictive accuracy with high dimensional datasets, few generalize well. Furthermore, it is impossible for unsupervised algorithms, which learn without feedback, to model high dimensional data optimally. Other machine learning algorithms highlight that they will not manage high dimensional data well. It has been documented that high dimensionality without proper consideration can yield empty models when evaluating such algorithms.

### 3.5.2. Techniques for Feature Engineering

Feature engineering is the process of transforming raw data into features that better represent the problem to the predictive models, resulting in improved accuracy. Although this step is arguably the most challenging and important part of the development process, it is often overlooked by machine learning practitioners that rely heavily on automatic feature extraction techniques. A good way to create new features is to apply domain knowledge to define functions that model the way new features relate to the predicted target. Domain knowledge is often used to specify new features based on combinations of raw observations, such as applying mathematical functions to individual measurements or groups of measurements, or to enable advanced data presentation. Feature transformations can also be used to increase robustness against noise and enhance the coherence of data that may be randomly distributed, such as applying power transformations to reduce skewness.

The recursive feature elimination method is a training set-based approach for feature selection that reduces the data dimensionality by recursively considering smaller and smaller sets of features. This method uses a model that assigns weights to each input feature and is applied to eliminate the lowest weighted feature until a stopping criterion is reached. In applications that use linear regression with L1 regularization or random forests or gradient boosting with tree-based significant features, the model that assigns importance scores to each feature is internally generated as part of the fitting process, and its outcome can be used to prune and select the optimal set of relevant features. For L1-regularized linear regression, the L1 penalty is used to zero out features in the set which are considered less relevant than others. For model-based feature selection, a model is fitted to the complete feature set and the most significant features are selected. Features that are deemed unimportant through significance tests are eliminated.

## 3.6. Conclusion

The use of algorithms in clinical settings is becoming more common for diagnosis and prognosis in a number of widely heterogeneous diseases. In addition, previously unapproached tasks may benefit from such algorithms, including uncertain diagnoses based on clinical symptomatology; prediction of disease severity and risk stratification; discovery of new disease endpoints; monitoring of disease course and treatment response, and help in triage, directing patients to the best healthcare resources available. These algorithms may have clear advantages over current tools, including standardization, higher sensitivity and specificity, especially when working with clinical data that are unapproached or only superficially approached, faster calculation in large data sets, or the ability to integrate heterogeneous data. However, while they may significantly assist clinicians in their work, their function should be complementary to clinical intuition and experience. They also require thorough validation on well-designed datasets representative of clinical practice, and long-term safety evaluations, to ensure that their implementation in clinical practice does not lead to adverse outcomes.

Despite the ciphering effect of algorithms, particularly deep-learning methods, the architecture of these may be fully transparent and explained in lay terms, to facilitate understanding by clinicians. Mechanistic algorithms based on Bayesian approximation are good examples of clear insight into likelihood functions driving responses, and easy computation for clinical applications. It is likely, as more clinical data becomes available, that new architectures will emerge that are made to understand the structure of clinical data contained in score matrices. We expect, and advocate, that mechanistic models should be always tested against data-driven models on progressively larger datasets to ensure that optimum solutions are available as clinicians demand direct support by these technical tools.

### 3.6.1. Emerging Trends

Machine learning is becoming more closely linked to e-health. The burst of interest in machine learning and deep learning propelled by modern digitization is spilling over to prognosis and diagnosis of diseases in healthcare. Enormous efforts are being made to apply machine learning to diverse biomedical data and problems. On the data side, massive repositories of genomic, imaging and electronic health record data are opening up for large-scale innovative discovery, and exciting new machine learning and deep learning methods are developing at a rapid pace, including feature learning, weakly and self-supervised training, reinforcement learning, continuous representation learning, theory-driven learning and creative generative models, to name a few. In addition to conventional supervised learning, unexplained variation learning, contrastive modeling, and various forms of causal regularizations are emerging as powerful techniques. These

innovations may uniquely position machine learning or deep learning to tap onto large biomedical datasets and advance key healthcare goals, including identifying risk factors, diagnosing the disease correctly, estimating natural history of disease progression, predicting serious adverse outcomes and supporting treatment and health management.

Machine learning and deep learning research in healthcare usually focus on traditional supervised learning methods, where annotated samples in the form of matched input-output pairs are used for training.

## References


Bui, N., & Zorzi, M. (2011). Health care applications: A solution based on the Internet of Things and Cloud Computing. Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies, 131–136. https://doi.org/10.1145/2093698.2093829

Zhang, R., & Liu, L. (2010). Security models and requirements for healthcare application clouds. IEEE 3rd International Conference on Cloud Computing, 268–275. https://doi.org/10.1109/CLOUD.2010.40

Vaquero, L. M., & Rodero-Merino, L. (2014). Finding your way in the fog: Towards a comprehensive definition of fog computing. ACM SIGCOMM Computer Communication Review, 44(5), 27–32. https://doi.org/10.1145/2677046.2677052

Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. Journal of Integrative Bioinformatics, 15(3), 1–15. https://doi.org/10.1515/jib-2017-0030

Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. BioMed Research International, 2015. https://doi.org/10.1155/2015/370194