www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 10 Issue 12 December 2021, Page No. 25709-25730
ISSN: 2319-7242 DOI: 10.18535/ijecs/v10i12.4678

Machine Learning Models for Early Disease Detection: A Big Data Approach to Personalized Healthcare

[1]Karthik Chava, [2]Chaitran Chakilam, [3]Mahesh Recharla

1. Devops Engineer, ORCID ID: 0009-0007-6001-4526
2. Validation Engineer, ORCID ID : 0009-0008-3625-4754
3. Sr. Oracle EBS Developer, ORCID ID: 0009-0008-3636-4320

Abstract

Healthcare is one of the vast and most crucial industries in this world. In this era, the healthcare industry is producing a huge amount of data every day. Everybody's life is captured by mobile devices/smart gadgets and smartwatches. Many records of human beings such as activities, heartbeat monitoring, running count, calorie count, number of steps, sleep patterns, daily mood, health status, stress monitoring, symptoms of diseases, blood pressure, temperature, and heart conditions are stored in the form of data. Predicting health issues automatically before noticing other symptoms is the need for the industry. A person's health record data is analyzed using machine learning models to predict the chances of being affected by diseases automatically in a few seconds. The healthcare datasets that are used to build the machine learning models are completely clean, and no preprocessing is required.

In today's world, automated prediction of diseases in an individual is progressing at a faster pace. Everyone is busy with their life and other needs; nobody has time to visit the hospital for a check-up every day. Therefore, every person needs a system that is capable of predicting diseases automatically in a matter of seconds in this busy world. ML techniques are used to analyze datasets containing a person's health records and other necessary information based on which disease can be predicted. To predict diseases in an individual, supervised machine learning models such as logistic regression, decision tree, random forest, support vector machine, and naive bayes are applied. In many countries, mobile applications are developed for the prediction of diseases based on symptoms in an individual. Healthcare experts are consulted for prediction in these applications. By contrast, the dashboard predicted in this paper using machine learning models automates disease prediction completely with the highest accuracy and performance compared to the existing applications.

Machine learning has evolved into a miraculous tool for the healthcare industry. Many organizations are dedicated to the prediction of diseases and their diagnosis such as diabetes, liver diseases, lung cancer, heart disease, breast cancer, and more diseases using machine learning models. The performance of early disease diagnosis is getting better every day, and researchers are focused on studying people's health records and enabling the system to predict diseases automatically. Compared to sharp models, high-performing deep learning models for personalized healthcare are not used widely as many healthcare industries have a huge number of unattended data. Elderly individuals and the common public have little or no technical knowledge and knowledge regarding the latest technologies, and ML applications may not specify.

Keywords: Early Disease Detection, Machine Learning in Healthcare, Personalized Medicine, Big Data Analytics, Predictive Modeling, AI-Driven Diagnostics, Healthcare Data Mining, Precision Health, Risk Stratification, Biomarker Discovery, Clinical Data Analysis, Patient Risk Profiling, Real-Time Health Monitoring, Supervised Learning Models, Healthcare Predictive Analytics.

1. Introduction

The world is exposed to an increasing amount of data every moment of its existence, which has led to various discoveries in all possible areas whether it is economics, politics, medicine, etc. However, while there are a lot of things that improve with the advent of data, on the contrary, there are also some downsides, as it makes a human less aware of their surrounding environment. In 1950, the concept of machine learning was invented. This concept of machine learning captures as much data as possible and didn't leave behind the Healthcare sector either. The healthcare sector produces huge amounts of data every minute, which, when captured, can be used for identifying many patterns.

These patterns can further be used to predict what will happen next. Machine learning is the continued development of artificial intelligence. Artificial Intelligence (AI) is a computer science technique that aims to mimic human cognitive functions. Artificial intelligence is bringing a paradigmatic shift to healthcare today, powered by the increasing availability of big healthcare data. As a result of the rapid growth of healthcare data, many machine learning algorithms for diagnostic prediction are being developed for personalized medicine. Machine learning models are being developed for the automated diagnosis of various diseases such as pneumonia, brain tumour, Alzheimer, breast cancer, etc.

This paper presents an efficient automated disease diagnosis model using various machine learning models such as a random forest, logistic regression, decision tree, extra trees, and gradient-boosting classifier. Three critical diseases, that is, coronavirus, heart disease, and diabetes, and their identification have been selected as the study context. An end-to-end process is designed and developed a new with the required interfaces where the users must enter their details in the mobile application and submit the data. Then the trained parameters of the model are stored in a real-time database, and based on the entered values, prediction is finally done in real-time.

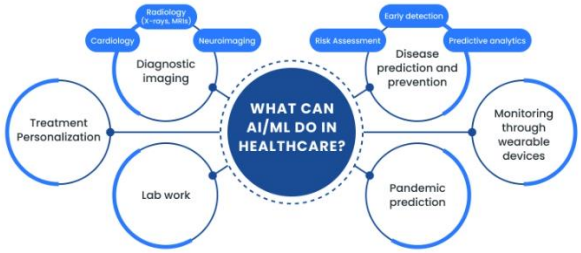

Fig 1: AI/ML Algorithms for Early Disease Detection and Diagnosis.

1. Background and Significance

Personalized healthcare is an emerging approach that uses technologies to provide patients with more individualized and periodic healthcare instead of one-size-fits-all healthcare. With increasing healthcare resources and expenditures, there is an urgent need to utilize technologies to improve the healthcare system and discover actionable knowledge from the data. Personalized healthcare is formed based on the preventive healthcare model. In the traditional model of preventive healthcare, physicians analyze patient information together with epidemiology knowledge to assess a patient's risk for a specific disease. With improper assessment, patients may miss timely preventive interventions. The personalized healthcare model relies on the observation of a patient's physiological conditions to assess the future health status of that patient. Taking patient physiologies into consideration can lead to more precise prediction than other relevant factors such as demographic information. However, it is not feasible to manually analyze such an enormous

amount of data to discover actionable knowledge. With reference to this problem, personalized healthcare is reformulated as a time series forecasting problem. Recent developments in machine learning technologies and the availability of electronic health records have stimulated the interest of many researchers in developing intelligent models to solve it.

The healthcare sector generates tremendous amounts of healthcare data. To make full use of big healthcare data, developing effective models is an indispensable yet challenging step. There has been a growing research interest in big healthcare data. Risk factors ranging from patient-level variables such as demographic and clinical variables to system-level variables such as community screening programs are explored. However, the enormous data volume and multiplicity make it infeasible to manually discover actionable knowledge. As a result, artificial intelligence technologies need to be employed. With a well-trained model given new data, predictions will be obtained within seconds. Advances in machine learning technologies contribute to medical predictions. Conventional machine learning models such as regression trees, support vector machines, and Gaussian classifiers are widely adopted models. Recently developed deep neural networks using massive healthcare data have achieved state-of-the-art performances. More importantly, DNNs can automatically discover informative features and build predictive models accordingly. This can reduce the efforts required in employing conventional models. Specifically, research efforts of utilizing DNNs to model time series healthcare data can be roughly categorized into three threads correspondingly.

2. Overview of Machine Learning in Healthcare

Healthcare is currently experiencing transformational shifts due to the growth of Artificial Intelligence (AI), new data mining and information retrieval algorithms, and cloud computing. This convergence of technology enables physician and patient empowerment with knowledge where it was very limited previously. All of this new technology has begun affecting many areas of healthcare and is slated to become more common in the future as healthcare systems are forced to adopt Big Data storage and communications due to value-based reimbursement requirements. On top of all this, the massive increase in the variety and volume of data available in healthcare systems is beginning to impact a variety of health-related information where it was very limited previously.

Machine learning (ML) is a subset of artificial intelligence that refers to a system's ability to learn from experience without programming function by function. The remarkable capabilities of ML systems are driven by new algorithms and computing power that enable vast amounts of hidden knowledge to be extracted from the masses of readily available, unstructured data. The applications of ML in every field have now reached systems that either have human-like performance or match human performance in specific domains.

In healthcare, common ML advances have been evolving for years. ML systems are now more inventively and fruitfully being used in healthcare for kinds of AI applications. The application of AI has the capacity to assist with case triage and diagnoses, enhance image scanning and segmentation, support decision making, and predict the risk of disease. This review focused on ML application to healthcare in the varied fields of electronic health records, medical imaging, and genetic engineering. These areas represent healthcare's BIG data, or the structured and unstructured data of the field, and have shown significant promise in relation to clinical applications.

2.1. Research Design

Research and data mining are solidly grounded on research design and planning. The appropriate design arrangement helps carry out solid and reproducible research. The procedure scope depends on the nature of the project, outcome expectations, and involved participants. To define the research project, the

context, research questions, methodology, and prospective respondents must first be explained.

The study and data mining procedures must also be suitable for the inherent project characteristics. Different data collection and processing methods, such as surveys, interviews, and automated monitoring/logging, must be considered. Automated methods are generally preferred due to efficiency and low human involvement because humans are prone to generate noise, errors, and various problems that reduce data quality and reliability. Accurate and systematic data collection, processing, and storage methods are also required to build a strong foundation for future data mining tasks. Finally, the presented methods must ensure that data processing does not violate expectations and regulations.

The provided research design and procedures are automated, which means that the methods must be fully specified and strictly followed. The aim is to minimize researcher involvement in data processing and improve outcome quality. Many problems can be avoided this way, and if documented properly, the method can be reproduced even after years with substantially different hardware. Thus, the research design and methods can ensure a sound basis for follow-up analysis or knowledge discovery. A comprehensive setting covering various aspects of the presented applications and studies is provided in the following. Understanding the problems raised by medical applications, data mining contexts, and research design essentials can help inspire future contemplation on various topics and explore additional currently undiscovered or disregarded possibilities and contexts.

Equ 1: Logistic Regression for Binary Disease Prediction.

- **Where:**

  $\mathbf{x}$ = patient feature vector (e.g., age, symptoms, lab results)

  $\mathbf{w}$ = model weights

  b = bias term

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$

3. Big Data in Healthcare

Healthcare data is one of the most complex topic areas from the big data landscape. Such data complexity arises from a number of sources, including its diversity, dynamism, the differing formats used for storage and treatment, and the potential presence of missing values and noisy data. Big Data in Healthcare can be classified into two broad categories: clinical data or clinical knowledge bases and experimental, administrative, or operational data. The former refers mostly to electronic Health Records (eHR) sources, i.e., information generated from the day-to-day activities in hospitals, including information about patients, prescriptions, rendered treatments, acquired devices, and monitored vitals, among others. These records can be regarded as clinical knowledge bases. Experimental, administrative, or operational data includes data resulting from medical laboratory analysis, including optical and magnetic resonance images, audio frequencies and spectrograms, pathology slides, and operation room interventions, among others. To ensure best data quality levels, all data sources must be transformed and aggregated by means of the appropriate ETL procedures. Transformations may include the standardization of representations and the imposition of integrity constraints. Missing and noisy data need to be cleaned by either imputation or filtering strategies. Finally, data may need to be aggregated to the appropriate levels to ensure the efficiency of analysis procedures. With the changed emphasis from aggregation to a more personalized approach to analysis, such as the patient-centered laying of Hadoop infrastructures, there is a need to use data mining tools that allow searching for data and transactional patterns, in addition to the high-performance processing of huge data volumes and complexity. There is also a need for data-intensive clustering and classification procedures also working on scalable architectures. Performance and scaling of data mining techniques is still poorly known and needs further studies. On the conceptual level, important subjects of study are the comparative assessment of retention policies, such as the

sensitivity of patients' profiles to the data source hierarchies used or the sensitivity of patient characterization/behavioral profile to the aggregation of data sources. Such public authorities may also provide common access to an anonymized part of their repositories, held on a data lake.

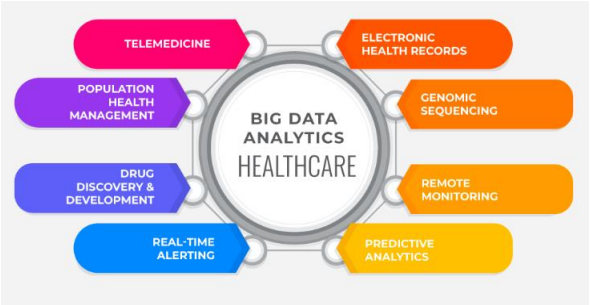

Fig 2: Big Data in Healthcare.

4. Types of Machine Learning Models

The field of medical research has largely benefited from machine learning and artificial intelligence technologies. It can be utilized for a variety of tasks, including early patient screening, patient health record lookup and prediction, disease early diagnosis determination, inquiry systems for patient consultation, and so on. Machine learning models are popular for disease detection and diagnosis, which is the detection of disease with the help of data from test results, medical history, demographic characteristics, etc. Automated disease diagnosis is completed first through disease prediction modelling in a Big Data approach. A machine learning model is first created by encoding, and the model is then trained on the dataset. Details of a patient including test result analysis results, medical symptoms, etc., must be entered onto a web or mobile data entry application. The prediction process is completed using the trained parameters of the model stored on the server, and then the disease name with the condition is depicted to the user.

Many machine learning models have been developed in the literature for the automated diagnosis of various diseases. Most of them are statistical models that have the limitation of providing very bad interpretations of the results. The objective of the proposed models is to correctly detect a person's disease early based on the examination of laboratory test results. Overall small-sized primary datasets are used in the proposed models. Most of the models cannot analyze continuous attributes on categorical labels, and only one or two models are presented in a paper. As a result, the approaches presented in the literature are good, and several new extensions for improvement are possible.

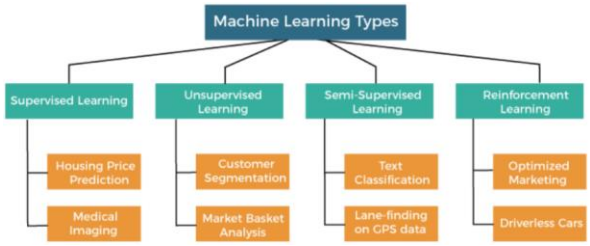

Fig 3: Types of Machine Learning Models.

4.1. Supervised Learning

To compare various classifiers and the effectiveness of supervised learning methods for disease detection, the proposed system aims at seven machine learning algorithms: Naïve Bayes, decision tree, random forest, KNN, support vector machine, logistic regression, and Adaboost. Analysts can analyze the data associated with the model quickly. The proposed strategy evaluates and compares the efficiency of various algorithms applied to medical datasets. Precision, accuracy, and F1-score are used to determine how well a classifier performs after model training has been completed. Classification algorithms have been trained using data from the COVID-19, heart disease, and diabetes dataset.

A healthcare analysis was used in an android application to assess the risk of COVID-19, heart disease, and diabetes. Models with the best skill on test data 6, and two models with mediocre performance are available for deployment. The ultimate goal of this scenario is to transfer everything from offline to online manipulation, advise this intermediary phase, and leave the more complex phases for later. The decision tree algorithm recursively partitions the input space into simpler domains according to rules inferred from the user data. Given an input sample, the rescanned space determines the class label to assign. A decision tree easily presents the model, allowing for alternative decisions and modes of interaction. The analysis of

multidomain medical systems face many challenges for external data due to the collaborative and federated nature of the platforms. The scientific interest on federated systems led to the creation of frameworks for secure multi-party training, continual learning, and analysis. Privacy constraints eliminate the risks of using extensive behavior profiles to the prevailing benchmarks used outside the medical domain. Possible class labels are represented at each candidate input by receiving and executing the patient data streaming processes. Ensemble methods have been used to combine multiple classifiers based on various algorithms in a hybrid model of classifiers. Random forests or decision trees are an example of ensemble methods based on their randomized development process and have exhibited remarkable success on several benchmark datasets.

Equ 2: Cost Function for Training a Disease Detection Model (Binary Cross-Entropy).

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right]$$

- **Where:**

  m = number of patient samples

  $y^{(i)}$ = true label for sample i

  $\hat{y}^{(i)}$ = predicted probability

4.2. Unsupervised Learning

Drastic success of big healthcare data and precise diagnosis treat demands has granted great consideration to use increased proficiency machine learning models to discover new prospects and lessen the financial burden of clinical data. But huge heterogeneous medical data handling and superior DC challenges raised strong speculation. Consequently, a valuable body of related projects and accomplishments become interesting points. Their presented/working portions of work will be briefly surveyed here through subareas to fulfill every aspect of this work. Although massive healthcare data continues increasing, the utmost US population does not have high-quality health access or insurance for this healthcare.

For diverse challenges of big health data mass and diversity, large-scale machine learning methods for health risk prediction with big population scale offer important analytics for diagnosis treatment management. This research proposed a semi-supervised collaborative filtering model, RL-CF and enhanced past collaborative filtering recommendation methods with multi-factor explanation, one of challenges being discovering informative factors causing rating. However, nearly all collaborative filtering-based recommendation methods did not take other features' diversity SCM into account and systems' side information. Price models of systematic asset failures are developed with the Euler scheme coupled with naive and full schemes for the popular Pareto. For testing the models, systemic asset returns are simulated and compared against the ones from the standard SF within the Markov Chain Monte Carlo approach.

The representation of interactive systems collectively is modeled with K-M, DBScan, and t-SNE clustering techniques. Collectivity for interconverted cultured populations, and mutual transformation between collective perception and collective behavior are revealed with K-M data structure transformation. On the other hand, patient recluster analysis regarding ongoing health evidence information or health events delivered by massive EHR data is essential for proposing new services and urgent response strategies by healthcare providers in a timely manner. Here, a novel deep aging assessment model is presented for identifying patients' aging stages and assisting doctors in preventive diagnosis treatment analysis strategies.

4.3. Reinforcement Learning

The investigation of ML applications in clinical areas has recently gained attention from scientists. Health-related big data provide the meaning, context, and association between data and clinical information, which ML can take advantage of. Considering the vital role that prediction plays in

providing treatment, scientists have developed deep learning models for the diagnosis and prediction of clinical conditions using EHRs. This algorithm used both structured data obtained from EHR and unstructured data contained in progress and diagnosis notes. The inclusion of unstructured data within the model resulted in significant improvements in all the baseline accuracy measures, indicating the versatility and robustness of such algorithms. On the other hand, natural experiments from the observational data of these complex systems are exploited to estimate causal relationships or treatment effects. Social media usually consists of natural experiments. A reinforcement learning framework that integrates the structural causal model into the actor-critic reinforcement learning model to learn from the observational data with real-world recommendation systems has been proposed. A model to predict post-stroke pneumonia within 14-day periods, providing a highly accurate model predicting pneumonia following a stroke, has also been built. Several ML-based models have also been implemented to predict mortality in ICU patients. Great ability to predict mortality in paralytic ileus patients using EHRs has been shown, which were developed into XGBoost models. Providing patients and practitioners with predicted mortality, through the use of EHR prediction algorithms, can allow them to make more educated clinical treatment decisions.

5. Data Collection and Preprocessing

This section fundamentally focuses on the collection of different datasets used in the work and their preprocessing to make them usable in training the models, as well as their performance evaluations. Information on the datasets used, along with the storage location of the datasets, is given in the first part. The second part provides information on preprocessing techniques such as data cleaning, data normalization, and estimation of missing data and outliers. The third part represents the architecture of seven different models constructed for this work. The right-side figure represents the architecture of the deep learning model constructed for this work; different layers of the model specified in the text box are also drawn in the figure. The last part provides information regarding how performance statistics such as accuracy, precision, sensitivity, etc. are calculated.

Four different public datasets are collected from the UCI repository. One of the reasons for picking this goal is that by building a comprehensive big-data architecture prototype, with the availability of jewelry health data now being collected and stored by the IoT networks, this work brings recommendations for endless innovative opportunities for researchers, doctors, and the jewelry industry to support consumer health while operating cost-effectively. It provides a technique for explaining the methods of building the comprehensive architecture to (1) deliver effective disease prediction and prevention (2) generate the rules for identifying the health status and predicting future health states of consumers (3) map the IoT data networks for aggregating and collecting data and monitoring services (4) serve the recommendations and information from independent resources to users within the predictions (5) integrate connective intelligence into its predictive model. Blood donation is a high-priority responsibility for medical and health departments in every part of the world. Blood donation safety and health problems emerged as a huge responsibility, research, and business priority primarily due to rising demand from patients for quality-of-life improvement and life-saving operations. Collection of blood donors on a regular basis, whilst ensuring their required health standards before and after blood donation, in conjunction with the health regulations of the medical board, has been flagged as a primary business priority, in consequence, driven by the exciting Machine Learning model developed and proposed for this work.

5.1. Data Sources

The efficacy of prediction models in practice is contingent on the model's setting and medical use.

The current literature primarily evaluates prediction accuracy by comparing the area under the ROC curve using offline predictions with thresholds fixed at the same value across the different prediction tasks. While accuracy is a crucial success criterion, it cannot carry the entire burden when building a prediction model. Different threshold values are reasonable for diverse tasks that exhibit varying clinical and operational implications. For example, consider a prediction model that estimates the likelihood of a readmission within 30 days. The healthcare system could respond to a positive prediction by reinforcing postdischarge contacts and monitoring, which would demand more resources than being notified of a possible need for an intensive care unit (ICU) admission 24 hours in advance. In addition, the magnitude of estimated probabilities can be informative. Low probabilities indicate that the patient does not share the characteristics of similar patients with the outcome, which could trigger alternative clinical pathways. In general, more insight is gained with considerable variances in probabilities than when almost all probabilities are clustered around 0 or 1. Probabilities in the middle range seem less informative because they are ambiguous regarding clinical action, leading to additional uncertainty.

Previously, this work evaluated a departed stochastic prediction model via retrospective analysis and simulation of current ICU prediction practice. The state of a patient at model departure initiates the prediction window. Selected prediction conditions are met if the input data contains the patient's state at that instant along with the essential history. Nevertheless, assessing forecast performance by real predictions in deployed systems entails evaluating the likelihood that the patient had the outcome during the predicted time given the input data used to generate the prediction. Inter-relations complicate checking input conditions for predictions that arrive after model departure. Target tasks change with time, leading to cascading changes in modeling being just one of multiple changes applying jointly. Moreover, newly generated data maintains some prior properties (e.g. frequency of outcome). Situations may arise where models are no longer viable, yet they continue to produce predictions.

5.2. Data Cleaning Techniques

Data cleaning is a crucial process in data mining and knowledge discovery. It involves preparing data for the next step of data processing and analysis, including transforming the data into a format suitable for creating knowledge by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Without meticulous and thoughtful preparation of massive datasets, huge computation time and resources may be wasted, while unreliable results may be inferred erroneously.

There are numerous data cleaning techniques suitable for diverse and complex datasets in disciplines such as health, clinical, and biomedicine. Functions of data cleaning comprise: (1) Data Cleaning for Missing Values, (2) Data Cleaning for Noisy Data, and (3) Data Cleaning for Inconsistent Data.

5.3. Feature Selection

Feature selection is the process of identifying and selecting a subset of relevant features for a learning task. In ML, feature selection is crucial to reduce the computational cost, enhance the predictive performance, explain the model, and prevent overfitting when the number of features is larger than the number of observations. Feature selection approaches can be classified into filter methods, wrapper methods, embedded methods, and hybrid methods. Filter methods evaluate the relevance of features by intrinsic criteria of the data without involving the learning algorithm, while wrapper methods evaluate subsets of features with the learning algorithm in order to identify a good subset. Embedded methods include model selection as part of the training process, and hybrid methods combine more than one approach to make full use of the advantages of the involved methods.

Feature selection can also be referred to as variable selection, variable subset selection, variable subset optimization, and feature subset selection. In predictive modeling, often the total set of features is not optimal. A small subset of features may lead to models with better performance as well as better interpretability. Several feature selection methods have been developed for knowledge discovery from data. They are based on different strategies, and they may even produce different feature subsets on the same dataset. It is important to compare them in order to make a better choice in practice. The proposed feature selection methods are evaluated on multiple benchmark datasets, and the experimental results reveal that the concise features selected by the proposed method can achieve at least comparable classification performance compared with the models trained on the original feature sets.

6. Model Training and Validation

In the healthcare and medical field, ML is becoming an indispensable technology in making automated decisions. Disease detection using ML models on rare domain data may be an uphill task. The automation of disease detection using ML models is an excellent tool for realization in healthcare systems. Many ML models are implemented on various diseases efficiently using an android application. Modern deep learning-based methods, with the availability of large amounts of data and high computational resources, have achieved remarkable performance in various tasks. Despite the achievements, similar to traditional non-deep learning methods, supervised learning methods remain very dependent on large-scale annotated data. Although supervised learning-based models are achieving high performance, they still suffer from some challenges. Firstly, labeling disease cases can be time-consuming and require expert knowledge. Secondly, the interpreter design is required for deep networks, which brings interpretability challenges that undermine physician trust in the models. Therefore, there is an urgent need to explore unsupervised learning-based methods to address these challenges. Unsupervised learning-based models have become critical research topics for early disease detection. Unsupervised learning can find visual concepts, local image feature detectors, or visual words without labels. Moreover, the models trained using unsupervised learning usually require less labeling effort. Transferring knowledge learned from different but similar tasks is the paradigm of self-supervised pre-training.

To enhance performance, domain adaptation methods leverage auxiliary data from the related source domain and create the adapted model for the targeted domain. It is mainly composed of two types, including data-level adaptation and feature-level adaptation. The data-level adaptation mostly matches the distribution between the target and source domain. Thus, the a priori assumption is attempted to be relaxed. Conversely, the feature-level adaptation approaches assume that similar distributions in source spaces should produce similar outputs. Therefore, the distances between feature spaces across both are attempted to be matched. Nonadditive noise and representation inconsistency issues are not addressed well. Specifically, nonadditive noise makes the input corrupted by heavy-tailed probability distributions. With representation inconsistency issues, the input feature distribution shifts, which happens in many real applications. These areas still need to be explored for further improvement in many fields such as behavior and circuit phase detection.

The conception of newly defined domain adaptation methods is to address domain adaptation challenges from both categories. Specifically, on the one hand, to address the non additive noise issue, given a few labeled instances and a large amount of unlabeled instances with significant data noises obeying heavy-tailed distribution, a multi-granularity hierarchical self-supervised learning paradigm is designed to exploit labeled and unlabeled speaker instances by a speaker embedding network. The channel feature space is robustified by augmenting data at the frequency level, enhancing early disease detection in FAEE. Furthermore, to deal with representation

inconsistency issues, cross-space correspondence is modeled from the witness triple pairs by a packing and expansion strategy. The temporal and image distributions across two domains are aligned through mutual information.

6.1. Training Techniques

Despite advances in modern technology and research, timely disease identification is critical in the healthcare industry. Preventive steps may be taken with early detection of life-threatening illnesses. A machine learning approach is used to solve the big data issue of quick disease detection in this study. For detecting and diagnosing serious and chronic illnesses like diabetes, heart disease, and coronavirus, random forest, SVM, and naive Bayes classifiers are used. Machine learning approaches identify a variety of characteristics. Preprocessing, selection of machinery, and selection of applied studies make up the proposed method. The algorithms are implemented in the big data environment utilizing various options to evaluate the effectiveness of the classifiers. There has been a lot of research in the area of disease prediction. Machine Learning based disease prediction was first introduced in the 1990s. Predicting future cardiovascular disease risk using a variety of health-related factors and attributing risk scores to people using a variety of classification techniques. Problems. Statistics, multicategory decision making, clustering, and association rule mining are just a few of the high workloads, consuming enormous amounts of handling social differentiation health-related applications in big data. The basic methods include temporal representation. Only very minor portions of these applications.

As a medical application in IoT that enables remote diagnosis of patients' health conditions, an enhanced patient health monitoring system construction containing a health data collection module, big data analytics, and an analysis result dissemination module is proposed. Based on the improved collaborative central residual-variant deep learning model, the big data analytics module provides a demographic health data record procedure realtime healthcare analysis that diagnoses health conditions. The analysis model training process speed and health conditions classification accuracy improve using feature decomposition, gradient deviation monitoring, weights consistency retention, and shared weights variance remaining. Coronary heart disease is one of the world's main causes of mortality. Such medical knowledge is difficult to analyze because it is often vast, broad, and numerous, which increases the complexity of velocity in big data analytics. Health evaluations that quickly analyze heart disease risk from demographics and historical diagnosis data obtain and illustrate risk factors that contribute to a patient being at risk of heart disease using a variety of data mining-based classification algorithms.

6.2. Cross-Validation Methods

There are two main approaches: traditional and advanced, as shown in Fig. The first method relies on one model, and after the pooled scores have been obtained, they are aggregated and assessed in a different way. The latter method consists of a more complex ensemble, where several models are trained on the entire dataset, and one of the ways of aggregating their predictions is meant to enhance performance.

The simplest form of cross-validation is K-fold cross-validation. The training and validation datasets are built from the entire pool of individuals. Thus, the distribution of risk factors, training, and testing metrics differs. A more appropriate solution is to condition them not only on the effect but also on the matching features, where the training and validation set contains equal manners of features for a given bucket, and the distributions of effects are allocated such that one bucket has no effects. There are also other methods for evaluating prediction scores compared with the previous method. One can rank samples of unseen score and those of predicted scores and compare them with alternative methods. In all approaches, one can aggregate these scores in a more appropriate manner instead of the correlation.

The only ensemble methods that are rather interpretable are linear models. Still, the advantage of such models is that from them, feature importance can be derived (despite these models being derivative of better performance). The matrices that evaluate prediction scores can be estimated and linked back to individual features, yielding global interpretability.

6.3. Performance Metrics

The evaluation of the performance of the proposed Machine Learning models for early detection of disease approaches is done through standard statistical measures, which are used in the related literature on predictive modeling of clinical events. The early warning system obtains predictions of uncertain events, in the form of probabilities from 0 to 1, where any interpreted value above 0.5 can be regarded as a landmark event. One of the most important measures of predictive performance is the good identification of positive events, i.e. by true positives and false negatives. The performance is expressed by sensitivity and specificity. In addition, it is important to quantify inherent predictive power, expressed by the area under the receiver operating characteristic curve (AUC).

Sensitivity, or true positive rate, is the proportion of patients with the occurrence of an event that are properly classified as positive predicted event cases. In CTD prognosis, sensitivity quantifies the fraction of patients presenting with a CTD event that are predicted within a specified time interval, e.g. one hour or one day. A higher sensitivity value reflects better performance of the model. Acceptable sensitivity values are > 50% or > 75% with a corresponding decrease in false positive rates. Specificity or precision is the proportion of patients without the occurrence of the event, who are correctly classified as negative predicted event cases. The prognostic models were evaluated with respect to 24-hour prediction horizon. AUC, the total area above the curve and below the horizontal line of predicted probabilities of 50%, is used to assess the relative accuracy of the computed probabilities: a higher AUC value indicates better discrimination performance. AUC ranges between 0 and 1: the random guess of the baseline model has an AUC of 0.50; perfect classification corresponds to AUC values of 1.0; AUC values $\leq$ 0.50 imply a non-informative model.

Given 10 scoring functions, performance of the ML models distributed over 7 different Hospital versions, during 104 different competitions per hospital, was categorized using the Mann-Whitney U-test. A receiver operating characteristic (ROC) curve analysis was performed to assess true positive rates (sensitivity) versus false positive rates (1-specificity) at various cutoff values. This analysis was performed to gain insight into the performance of the different ml models for each hospital. The area under the curve (AUC) statistic was computed to derive an aggregate measure of performance and account for all possible differences in classifications as done previously in prediction of short-term ICU mortality.

7. Case Studies in Early Disease Detection

The increasing availability of personal application wearables and continuous real-time data collection methods has fueled the interest of researchers in developing tools for effective early disease detection. These longitudinal datasets present unique challenges that differ from traditional data classification approaches, such as the need for proactive prediction. This means predicting the likelihood of disease occurrence for a given future time frame is unknown at time t, instead of the more common task of predicting outcomes for the next immediate time frame of the dataset. The difficulty is further compounded by the deluge of temporal sequences with many data streams across various patients. Predictive systems are critical in the modern healthcare landscape to minimize healthcare burdens and provide for the aging world population to maintain a high standard of life. Separately, the emergence of machine learning techniques allows building sophisticated predictive models to mine existing electronic health records and administrative claims data for valuable insights. In this field, a few

promising research efforts have demonstrated the feasibility of early disease detection with non-negligible clinical value. The task formulation and proposed prediction models are of high innovation, and the latest simulation results run on massive datasets stimulate continuous exploration of the paradigm.

Two case studies are presented as notable success stories on the recent research topic of early disease detection from electronic health records. The first case study provides a time-to-event prediction model derived from clinical notes to predict future acute diseases, which significantly outperforms baseline models. The massive dataset covers a long-term study of more than eight years from over half a million patients, which has witnessed a fundamental step change as many real-world datasets are of modest size, and the feasibility of prediction modeling at large scale remains under-explored. The results demonstrate the clinical feasibility and contribution of progressive predictive risk models via clinical notes. The second case study develops an ensemble model to predict future reimbursement claims of heart failure from multi-source structured data, which can capture the non-linear correlations between multiple predictors using tree ensemble models. Recent study has also been conducted on medical and pharmaceutical claims data for longitudinal health risk assessment against multiple chronic diseases.

7.1. Cardiovascular Diseases

Cardiovascular Diseases (CVD) are the primary reason for death, accounting for one-third of all deaths worldwide. Routine health check-ups are frequently conducted to address this issue. However, the identification of CVD early before the symptoms become evident is a major threat to humans. Machine learning can help to address these concerns and do so early on. This work aims to provide physicians with a way to analyze patients' health data and integrate it with social media to monitor patients' health in real-time. It only takes a single data snapshot to monitor health for CVD, thus real-time access to patient medical records is very valuable for a pensioner patient to improve their health in critical conditions. Social media make it easier to disclose an individual's general background and health monitoring data. There are several types of diseases, but the early detection of chronic diseases can lessen damage to patients and prevent premature mortality. Out of all the chronic diseases, CVD are highly complex and major diseases because 80% of heart diseases and strokes can be avoided by changing unhealthy lifestyles and through early detection. The early identification of CVD is an open challenge to researchers. It includes heart complexities such as artery blockage, irregular heartbeat, blood flow stopping, and fat structure in veins. The patients' health data in datasets consist of features that are vital for CVD prediction. In this paper CVD prediction is done by selecting features with various standards so that model accuracy increases. A large patient database with over fifty-five attributes is searched for the best set features for model training. The attributes are demographic characteristics, symptom and examination features, ECG based features, laboratory features, and echocardiography features. From the original selected features, discriminatory features are chosen based on LASSO, Tree based algorithms, Chi-Square, and Recursive Feature Elimination. The features selected with various standards are trained with seven classifiers that were optimized for the dataset. Overall accuracy for patients suffering from CVD was the criterion for selecting the best model. A Support Vector Machine (SVM) classifier was utilized along with 15 chosen features which perform better than other selected feature sets with overall accuracy. A dataset with 56 attributes and 303 occurrences for the CVD patients was collected from a veteran hospital. Out of 303 patients, 87 were healthy and the remainder had factors conducive to CVD. Outliers were removed using the K-means clustering method. Feature reduction was applied to retain only no more than 13 features to improve the performance of the classifiers. This part of the research particularly focused on seven different

classifiers, and applicable measures explaining the model performance of selected classifiers. These classifiers were chosen after gaining an understanding of the CVD dataset, emerging training data preprocessing, and feature extraction.

7.2. Diabetes

Diabetes mellitus is a group of metabolic conditions characterized by chronic hyperglycemia. Insulin ineffectiveness or inadequacy is mainly responsible for the onset of this chronic disease. Diabetes is classified into two forms: type 1 diabetes (T1D), which generally occurs in children and arises from autoimmune destruction of insulin-producing beta cells, and type 2 diabetes (T2D), which is characterized by insulin resistance often resulting from dietary factors, lifestyle, and genetic predisposition. Both types can lead to long-term complications, such as heart disease, kidney failure, loss of sight, and amputation. According to the International Diabetes Federation, more than 537 million adults (20–79 years) worldwide had diabetes in 2021, with the number projected to rise to 643 million by 2030 and 783 million by 2045 at a cost of USD 966 billion. In the absence of intervention, there may be more than 197 million undiagnosed cases in both developing and newly industrialized countries.

Recent investigation has shown that early diabetes diagnosis helps prevent further damage. Healthcare organizations are actively working to decrease the diabetes burden by continuously monitoring several health data. However, healthcare health professionals struggle with the increasing volume of data, lack of technological background, and inefficient tools to assist them. Therefore, researchers have paid great attention to the development of artificial intelligence (AI)-based tools and methods suitable for chronic conditions monitoring and control. Specifically, machine learning (ML) models have been widely utilized to quantify the risk of a disease occurrence. However, most works focused on predicting the shot-down of a chronic condition, such as asthma, chronic obstructive pulmonary disease (COPD), heart disease or failure, mental health or depression, etc. The majority of the studies employed health and biodata such as activity, sleep and vital signs, besides traditional static patient factors such as age, BMI, gender, hypertension, etc., with a focus on ML methods proper analysis and well-designed data collection but not novel AI-based classifiers.

7.3. Cancer Detection

The research interest around the application of deep learning in the investigation of cancer is gradually growing. At present, a wider range of research work is done with a focus on lung cancer than the other types of cancer, even breast cancer. This helps to understand that the ratio of lung and breast cancer is too high. In the previous works, several demerits were found. Specifically, it was found during correlation analysis that due to the use of inaccurate methods of processing or hyperparameters, the accuracy of some models was found to be too poor. On the other hand, some models used complex models but were hardly robust, as this study collects several parameters. Most of the collective works consider only one or two parameters for the exploration, which is too small to check the accuracy and robustness of the model.

In this work, the idea of a statistical biopsy is introduced for early cancer detection, analogous to tissue biopsy and liquid biopsy. As an initial application of this approach, two neural networks were trained to predict cancer risk for 17 different cancers from health exam features in the Prostate, Lung, Colorectal, and Ovarian studies. Using a test set from UK Biobank shows that it generalizes beyond the training distribution and provides convincing evidence that this model is not overfitting the PLCO data on the training set. Testing the model on an independent and external dataset from UK Biobank shows high performance as well on the test set, providing convincing evidence of robustness. There remain places where the model does not perform well, such that on biliary cancer, liver cancer, and breast cancer in males, the model does not generalize at all. Furthermore, it was observed

that in almost all cancers, the generalizability of the male model is worse than that of the female model. The goal is also to further test the importance of features that are only for females/male and where there are other features that should be added.

8. Personalized Healthcare Approaches

Personalized healthcare aims to provide individualized interventions by taking into account various risk factors, highlighting the great importance of accurate biomarkers and robust risk prediction models. Early detection of complex diseases, such as cardiovascular disease (CVD), cancer, and Alzheimer's disease (AD), is crucial for successful intervention. PulseWave, the presented computational pipeline, combines rigorous statistical methods, gene expression and genotype information, and an innovative application of multiple machine learning (ML) algorithms. Using PulseWave, it has been shown that gene variants (V) and expression (E) data, exclusively and integratively, can identify known (B) and novel (B') biomarkers and achieve high accuracy in CVD risk prediction.

In this study, complementary aspects of various ML approaches show great potential for personalized healthcare applications. The capacity of gene variants and expression data and novel multi-step methodology hold great promise for public health. Future work will focus on further refining PulseWave, evaluating it on complex diseases with detailed genomic data. Other avenues of research include adaptations of the proposed approach for diseases like breast cancer, diabetes, and AD, optimizing intervention and treatment options. Additionally, the development of a user-friendly platform is paramount for quicker adoption by medical professionals sharing common goals of improving health interventions in vulnerable patient populations.

Fig 4: Personalized healthcare strategies.

Moreover, personalized medicine is the personalized treatment of diseases based on biological characteristics, achieving the best outcome with the least harm. The study of the human genome has helped biomedical research to gain insight into syndromes, screening populations, and identifying disease-causing genes. Next-generation genomic sequencing (NGS) has generated both opportunities and challenges in clinical practice, raising concerns about downstream bioinformatics and translation of findings into meaningful clinical interpretations. The clinical output of DNA sequencing needs to be combined with other modalities of –omics, including RNA, proteins, metabolites, and clinical data collections, creating a huge amount of data with great complexity. Data mining and machine learning (ML) have shown great potential in analyzing and mining medical big data for clinical impact, as they can automatically construct finite models from the data.

8.1. Tailored Treatment Plans

Recent medical advances using wearable sensors mean that billions of user wearables are being on the market. Wearables can measure many different types of physiological signals at scale, such as heart rate, heart rate variability, heart conduction quality, blood pressure, electrocardiogram, glucose, etc. Estimating

clinical parameters, predicting chronic diseases, and determining whether examination is needed are all critical healthcare tasks. However, modeling temporal sequences rather than single values is critical for analysis on wearable data streams. This work models chronic disease prediction as a binary classification task on high-dimensional sequential data, where advanced one-dimensional convolutional neural networks are built. Models treat wearable streams as time series signals in spectrogram format for effective feature extraction. A pilot study demonstrates robust feasibility and high predictive accuracy on a 15-fold cross-validation test.

Wearable devices can measure many physiological signals, and with their rapid acceptance, their market has reached billions of users worldwide. However, wearable data streams differ from classical tabular data, making it critical to develop appropriate models. This work investigates treating wearable data sequences as a spectrogram for one-dimensional convolutional neural network modeling, investigating its utility for predicting a candidate clinical diagnosis. However, appropriate designs need to satisfy different needs, and the correct algorithms largely depend on careful effort and tuning. Good choices regarding the key hyperparameters may yield different outcomes. Analyzing the model's metrics is nontrivial when predicting potentially relatively rare conditions. Advanced ensembles of hard classifiers don't achieve increased performance on this source data. Clinical experts' knowledge in physics needs to be better exploited.

This work proposed a novel framework to plan treatment processes. It evaluated the framework from the viewpoint of methodology using a synthetic dataset. Accordingly, it is applied to an actual health checkup dataset to plan actionable treatment processes for improving blood pressure values. The computed treatment processes have been confirmed to be actionable and consistent with clinical knowledge about lowering blood pressure. A pilot study combined treatment planning with a surrogate Bayesian optimization approach to health improvement planning based on health checkup results and personal characteristics. The treatment plans computed are adapted to the individual and improve health at a population level. It opens new opportunities for developing realistic and practical health improvement plans, which have become a crucial component of precision medicine.

8.2. Patient-Centric Models

To counteract the growing burden of disease and the inequities in healthcare brought about by a global ageing population, healthcare systems may be reformed by shifting focus from treatment to prevention. For this, patient-centric, risk-stratified healthcare, where each patient is continuously assessed both in terms of their medical risk and their engagement with healthcare services to understand that risk, must be spearheaded by population health management (PHM). Attaining this vision is difficult and will require a revolutionary approach to the modelling of how health evolves and the wider epidemiology of disease over time. This modelling must be patient-centric in that it must translate patient data into patient-level outcomes that affect a larger population over the timescales of health intervention, typically weeks, months or years from detection. Foundational to this healthy information system is a scalable approach to learning disease outcomes from electronic health records (EHRs).

The great potential of this paradigm has yet to be fully harnessed. Modeling low-dimensional embeddings of clinical notes allows embedding features to be easily derived for other models. The interpretable embeddings trained on structured EHR data from a population-level health intervention were found to generalize both to different geographies with different healthcare systems and to patient-level prediction of individual disease onset and progression over time.

Broadly, population-level health is shaped in large part by primitive healthcare events. Treatment or medical intervention events can only serve to negate the effects of high risk prior event types such as receiving a diagnosis, drug type or procedure type.

Random forest exclusion of the first medical intervention event type removed the vast majority of positive treatment events and preserved a much more tractable dataset for training patient disease outcome models. Recently-failed standalone predictive models derivatively processed both event types into structured embeddings.

Embedding encodings for the event types were created using a deep semantic encoder that outputs high-dimensional annular vectors in which every dimension describes the relational clustering of one of 52 event types. Notably, all set embeddings are similar across event types as learned by a shared temporal attention mechanism, allowing models trained on event type-specific embeddings for one healthcare system to transfer learn models with inductively learned input. Both independence of model parameters from EHR count and training on lower-dimensional representations of clinically-relevant features decrease model size without sacrificing model performance, potentially opening the arena of individual health predictive models to under-resourced healthcare systems.

9. Challenges in Implementation

ML provides solid foundations for health technology challenges but its implementation for (early) disease detection is not straightforward. The challenges include: (1) Data acquisition and sharing, (2) Model selection and validation, (3) Trust and Explainability, (4) Clinical integration, (5) Regulation and audit.

There can be practical concerns with the collection and sharing of recordings, e.g., for lung sound recordings in primary care settings. Data from wearable sensors can also often only be accessed via device or app APIs. This is despite low technical barriers to entry. Whose data is it? Is it possible to stream it all to the cloud for processing? These concerns and more may lead clinics to decide not to participate in a data collection scheme. In such cases, it may be possible to deploy (early) disease detection systems at smaller partner organizations where collection and access issues are less strict and still remain representative in terms of networks. Federated Learning (FL) can be a means to maintain privacy while collaboratively developing models. However, this too can lead to issues with the device manufacturers or central parties holding power over the applications or capturing all the data.

Properly trained AI models with excellent validation results can still go wrong in practice. This is regularly observed in cases of dataset shift where data distribution changes over time, the loss function is miscalibrated in training, or feedback loops cause an elevation of states in a closed system. Therefore, proper assessment efforts should be in line with the expected multiplicity of model applications and key assumptions. Simulated dataset shifts mimicking realistic use-case conditions such as different population age distribution or symptom severity can be investigated in clinical practice, while simulation of application scale digitization footprints in pre-clinical practice environments can be used for testing. If the model's nondominant prediction values are not trusted or anticipated to be incorrect, the models should not make locomotion recommendations and should exit.

9.1. Technical Barriers

Despite major advancements, several technical barriers hinder the widespread adoption of machine learning technologies in the clinical setting. Clarifying the challenges faced by healthcare professionals and data scientists when implementing and interpreting machine learning techniques, such as neural networks and deep learning methods, is essential. How to foster a hybrid collaboration allowing efficient designing and development of interpretable applications and reliable validation, in order to perform the final check on correctness and reliability. Explaining the black-box technique, doctors are more likely to use rule-based systems.

This would lead to the mitigation of cognitive dissonance problems. They would be simpler to validate with great performance weights. Early discussion about the development with doctors. To aid in the understanding and use of the models,

recommendations should be created. Models accounting for borderline instances in a supervised way. Explain the legal limitations of machine learning in clinical practice, the training set sampling problem that fosters biases and unfairness. Explain when and how to generate and use validated historic data. How to assure a trustworthy translation of knowledge in machine learning applications.

A study of the reported literature suggests different areas of improvement. Explainability of trained models and monitoring of compliance with clinical practice. A better formalization of how to convey the knowledge to clinical personnel. A better understanding of model accepted failures and errors, validated by medical experts. Trust-and-agree displays, to allow an interactive analysis of the data and the requested features. Additional model checks, able to assess the performance of the input data in terms of precision and recall, might be created.

Meta-models possibly accounting for the uncertainty of black-box models could also be useful. Better understanding of acceptance and agreement of black-box paths. Methods capable of monitoring the convergence mechanism of the latent states in terms of neural activation and explorative sample requirements. Human-machine interfaces that are capable of testing both the performance plausibility and controlling the training process behaviours on a wider range of datasets could be useful.

Equ 3: Precision and Recall.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

- **Where:**

  TP = true positives (correct disease detection)

  FP = false positives (false alarms)

  FN = false negatives (missed detections)

9.2. Regulatory Issues

In recent years, the rapid advancement of machine learning has allowed healthcare companies to adopt methods that promise greater efficiency, better risk assessment, and more refined data analytics in the analysis of clinical Heath records. However, if these methods become widespread, patients' privacy would be at stake. So far, the issue has been neither sufficiently researched nor is there a consensus about its urgency. This is especially true for emerging fields such as the Internet of Things and machine learning. Issues arising from the analysis of speech files or data triangulation have not yet been properly addressed.

It is a lapse in foresight to consider these issues solely in the light of the current models or technology. The consideration of possible harm should lead to an evaluation of the impact of data breaches on different kinds of patients. Furthermore, regulation should not single out industries, which would inevitably hold back innovation. Instead, it should focus on the micro-level: the ways patients are currently being classified. In order to effectively regulate machine learning on clinical data, this micro-analysis of risk should be the starting point.

In the light of the discourse on risk classification – in a variety of fields from insurance to the Internet – four such ways can be distinguished, and this analysis shall show how such a typology could be conducive to regulation. In abstraction these methods range from the highly aggregative approach of simple classifiers such as the ‘calibrated estimate’, customary in finance, to time-oriented methods, such as survival analysis, common in clinical risk assessment. This is unfortunately all too often followed by the final step of the four-fold typology: the more opaque algorithms of deep learning. Tackling such issues by demolishing black-box models, alterations of clinical practice or insurance rates is theoretically possible. However, it is a last resort that would eliminate arguably articulable and manageable methods while hampering progress in the other disciplines involved.

10. Future Trends in Machine Learning for Healthcare

In general, applications of ML process data from the collection web and generate predictions for learning agents. In many ways, healthcare is a ripe setting for

ML approaches. The increasingly comprehensive collection of behavioral and physiological patient data through wearables, implants and other patient-generated devices as well as medical and home monitoring data has opened the gates for big data approaches to personalized healthcare. Despite some notable success stories, however, the deployment of ML in healthcare remains limited. No statistical IA could outperform doctors in detecting breast cancer in mammograms, and no predictive model has entered the clinical routine in neuroimaging. Points of failure in real-world applications exist at every step of this complex process and they require thorough investigation. Health information systems and their progressive implementation. The improvement of community conduct and the treatment of chronic diseases are greatly aided by health information systems and their progressive implementation. Health professionals' working methods are changing as a result of the gathering, processing, organizing, analyzing, exchanging, and utilizing health and medical data. Healthcare information systems and data management are seen to be agents of change in enhancing knowledge-based practice and treating long-term illnesses such as diabetes, cardiovascular diseases, and hypertension as problems of public health. The outcomes of improved efficiency, the elimination of duplications, and the collaboration of healthcare providers all contribute to higher patient safety and quality. Accidentally or intentionally, a patient may end up with incomplete knowledge while seeking care. As information asymmetry remains important in this arena, the healthcare challenge would be reduced if the agents exercising expertise recognized the health need and efficiently planned care based on continuous health. For providers, this translates into running shifts, optimizing workflows, scheduling, processing cohorts, creating episodes of care, and managing errors with claims reporting.

Fig 5: Future Trends in Machine Learning for Healthcare.

10.1. Advancements in AI Technology

Artificial Intelligence (AI) has enormous potential for analysis of the massive amounts of data generated by the Internet of Things (IoT) systems. AI-based technologies will provide quicker results on millions of signals and smart devices that potentially benefit the Domain Environment and lower their Energy consumption. Earlier, AI technologies were significantly applied only in the Field of Mathematics and Computer Sciences. With the recent advancements in technologies and band-width improvement in internet facility, AI's usage expanded widely to various disciplines like Medicine, Health care, Telecom, Agriculture, Finance, and explaining Cosmic Radiation. Startups providing discovery engines, natural language processing engines are mushrooming due to its immense potential_large data generated from electronic health records in clinics, diagnostics refrigerators, and biosensor control devices_legacy IT structures are wasting resources and do not provide sound decision metrics. Big data analysis from multiple sources will fetch better information mining quality explorations from the data_synthetic controls for electronic health measures, diagnostic process consultancy, and Drug efficacy validation for burgeoning Adverse Drug Events (ADEs)_Big Data from social networks and mobile devices is used for disease monitoring, Human Activity Recognition, and routing optimization. Data science methods and machine learning algorithms have wide implications in Healthcare, Pedagogy, Social Sciences, Bio informatics and mathematical modelling. Hospital beds categorized On Duty, On Inventory; and social load classified as Normal,

Above Normal connect various measurements with deep learning (DL) approach. The domain-relevant data sources Memo Ray, ASKI_KARE and ELSE level three dimensions, varied data and domains, complexities, classification variables adopted. These big data-native learning systems handle business-rule and excursion detection modeling of genomic, proteomic and metabolomic time-series, continuous and human and data sensing diverse-nonstationary scalable raw data.

10.2. Integration with Wearable Devices

Wearable devices have emerged as a turning point in individuals' proactive management of their own healthcare. Wearable devices can facilitate the continuous monitoring of health parameters in a more personalized manner in fields such as 5P-medicine , where a massive amount of data can be collected and analyzed. The objective of Smart Health is to leverage wearable sensors' capabilities by allowing integrated, holistic, and long-term approaches in healthcare. In the context of the big data era, Big Data technologies and related applications will enhance the persistence, structure, type of data collected and behavioral modeling with standard and accepted data formats and structures. This chapter proposes a new approach to patient monitoring leveraging smart technologies, adapted textual and fully interoperable sensor data. The information from patients' sensors will be processed and knowledge will be used to identify individuals and communities that need support. Personalized interventions will be devised adapting both devices gathering smart technologies and information processing involved.

Predictive risk modeling is a fashionable area of research. It involves using historical data to predict which individuals are most likely to develop undesirable future states. Matter of concern has to do with proper evaluation of the resulting risk models. Metrics that consider conflict of interest are required, since it concerns health and people's lives. Current practice evaluates models by means of concordance statistics, typically Harrell's c-index. For predictive monitoring, it seems natural to consider similar metrics that account for the timing of events, such as restricted time-dependent concordance statistics. When evaluating risk models trained to monitor a health condition at a high prevalence, it is critical that performance metrics are custom to the point at which decisions are made, which may be adjusted as a function of prevalence.

11. Conclusion

The demand for sufficient funds and improved healthcare facilities is a vital action taken by every nation. Given India’s booming population and limited healthcare facilities, it is imperative to innovate healthcare technologies that can provide better efficiency and accuracy. Over the last decade, there has been a steep rise in mobile phone usage, which greatly increases the possibility of promoting innovative healthcare. Detecting diseases is the first step towards offering medical assistance. Such disease detection can be aided with the advancement of mobile technology and Machine Learning (ML) techniques.

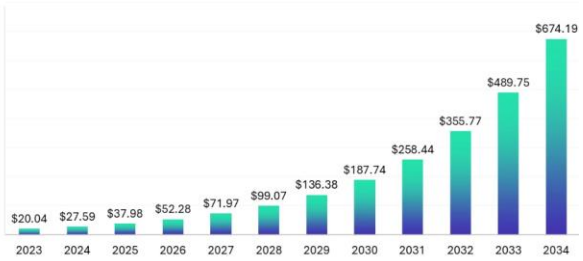

Fig 6: Machine Learning for Early Disease Detection: A Big Data to Personalized Healthcare

The machine learning models proposed and implemented in this study can detect diseases based on the disease parameters given over mobile. Three diseases were chosen for this study: Heart Disease, Type-2 Diabetes, and COVID-19. An interactive User Interface is developed using Android Studio and Java on a smartphone, where a user can enter the parameters for the chosen disease. The model will give the predicting disease results in real time based on the parameters entered. Moreover, the pandemic of COVID-19 has changed human life considerably. There are many things that have changed, like the

way work was handled, meetings were conducted, and assessment of the strength of the people was done, but the main challenge still remains. "The COVID-19 virus affects humans and not machines". There is still a great challenge concerning Human COVID-19 stage prediction. The machine learning models proposed and implemented in this study can help predict the pandemic diagnosis stage based on the COVID-19 patient details provided over mobile. A User Interface is developed using Android Studio and Java on a smartphone, where users can enter details regarding the patient. The model will provide a prediction of the patient's COVID-19 infection stage in real-time, along with the runtime glossary using predefined medical terms and their meanings.

12. References

1. Paleti, S., Singireddy, J., Dodda, A., Burugulla, J. K. R., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures (December 27, 2021).
2. Gadi, A. L., Kannan, S., Nanan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87-100.
3. Someshwar Mashetty. (2020). Affordable Housing Through Smart Mortgage Financing: Technology, Analytics, And Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 99–110. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11581.
4. Sriram, H. K., ADUSUPALLI, B., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks.
5. Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29-41.
6. Just-in-Time Inventory Management Using Reinforcement Learning in Automotive Supply Chains. (2021). International Journal of Engineering and Computer Science, 10(12), 25586-25605. https://doi.org/10.18535/ijecs.v10i12.4666
7. Koppolu, H. K. R. (2021). Leveraging 5G Services for Next-Generation Telecom and Media Innovation. International Journal of Scientific Research and Modern Technology, 89–106. https://doi.org/10.38124/ijsrmt.v1i12.472
8. Adusupalli, B., Singireddy, S., Sriram, H. K., Kaulwar, P. K., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks. Universal Journal of Finance and Economics, 1(1), 101-122.
9. Karthik Chava, "Machine Learning in Modern Healthcare: Leveraging Big Data for Early Disease Detection and Patient Monitoring", International Journal of Science and Research (IJSR), Volume 9 Issue 12, December 2020, pp. 1899-1910, https://www.ijsr.net/getabstract.php?paperid=SR201212164722, DOI: https://www.doi.org/10.21275/SR20121216472

10. AI-Based Financial Advisory Systems: Revolutionizing Personalized Investment Strategies. (2021). International Journal of Engineering and Computer Science, 10(12). https://doi.org/10.18535/ijecs.v10i12.4655
11. Cloud Native Architecture for Scalable Fintech Applications with Real Time Payments. (2021). International Journal of Engineering and Computer Science, 10(12), 25501-25515. https://doi.org/10.18535/ijecs.v10i12.4654
12. Innovations in Spinal Muscular Atrophy: From Gene Therapy to Disease-Modifying Treatments. (2021). International Journal of Engineering and Computer Science, 10(12), 25531-25551. https://doi.org/10.18535/ijecs.v10i12.4659
13. Pallav Kumar Kaulwar. (2021). From Code to Counsel: Deep Learning and Data Engineering Synergy for Intelligent Tax Strategy Generation. Journal of International Crisis and Risk Communication Research , 1–20. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2967
14. Raviteja Meda. (2021). Machine Learning-Based Color Recommendation Engines for Enhanced Customer Personalization. Journal of International Crisis and Risk Communication Research , 124–140. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3018
15. Nuka, S. T., Annapareddy, V. N., Koppolu, H. K. R., & Kannan, S. (2021). Advancements in Smart Medical and Industrial Devices: Enhancing Efficiency and Connectivity with High-Speed Telecom Networks. Open Journal of Medical Sciences, 1(1), 55-72.
16. Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29-41.
17. Kannan, S., Gadi, A. L., Preethish Nanan, B., & Kommaragiri, V. B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.
18. Implementing Infrastructure-as-Code for Telecom Networks: Challenges and Best Practices for Scalable Service Orchestration. (2021). International Journal of Engineering and Computer Science, 10(12), 25631-25650. https://doi.org/10.18535/ijecs.v10i12.4671
19. Srinivasa Rao Challa. (2021). From Data to Decisions: Leveraging Machine Learning and Cloud Computing in Modern Wealth Management. Journal of International Crisis and Risk Communication Research , 102–123. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3017
20. Paleti, S. (2021). Cognitive Core Banking: A Data-Engineered, AI-Infused Architecture for Proactive Risk Compliance Management. AI-Infused Architecture for Proactive Risk Compliance Management (December 21, 2021).
21. Vamsee Pamisetty. (2020). Optimizing Tax Compliance and Fraud Prevention through Intelligent Systems: The Role of Technology in Public Finance Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 111–127. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11582
22. Venkata Bhardwaj Komaragiri. (2021). Machine Learning Models for Predictive Maintenance and Performance Optimization in Telecom Infrastructure. Journal of International Crisis and Risk Communication Research , 141–167. Retrieved from

https://jicrcr.com/index.php/jicrcr/article/view/3019
23. Transforming Renewable Energy and Educational Technologies Through AI, Machine Learning, Big Data Analytics, and Cloud-Based IT Integrations. (2021). International Journal of Engineering and Computer Science, 10(12), 25572-25585. https://doi.org/10.18535/ijecs.v10i12.4665
24. Kommaragiri, V. B. (2021). Enhancing Telecom Security Through Big Data Analytics and Cloud-Based Threat Intelligence. Available at SSRN 5240140.
25. Rao Suura, S. (2021). Personalized Health Care Decisions Powered By Big Data And Generative Artificial Intelligence In Genomic Diagnostics. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v7i3.3558
26. Data Engineering Architectures for Real-Time Quality Monitoring in Paint Production Lines. (2020). International Journal of Engineering and Computer Science, 9(12), 25289-25303. https://doi.org/10.18535/ijecs.v9i12.4587
27. Mandala, V. (2018). From Reactive to Proactive: Employing AI and ML in Automotive Brakes and Parking Systems to Enhance Road Safety. International Journal of Science and Research (IJSR), 7(11), 1992-1996.