Scientific Publications

Current Research in Public Health, 2020, 1, 1355
www.scipublications.com/journal/index.phpo/crph
DOI: 10.31586/crph.2020.1355

Review Article

Predictive Modeling and Machine Learning Frameworks for Early Disease Detection in Healthcare Data Systems

Sateesh Kumar Rongali [1,*]

[1] Independent Researcher, USA

*Correspondence: Sateesh Kumar Rongali (sateesh.kumar.rongali.research@gmail.com)

Abstract: Predictive modeling, supported by machine learning technology, aims to analyze data in order to guide decision-making towards actions generating desired values in the future. It encompasses the set of techniques used to build models that estimate the value of a certain variable predicting a forthcoming event from the past or current values of relevant attributes. In predictive healthcare modeling, the built models represent the relationship among the data concerning customer, provider, production, and other aspects of the healthcare environment in order to assist the decision processes in the prevention of diseases and in the planning of preventive actions by detection of high-risk patients. Contrary to trend analysis, whose goal is to describe past events, predictive models aim to provide useful indications regarding future events and changes. Predictive healthcare modeling supports actions that try to prevent the manifestation of diseases in healthy individuals or try to diagnose as early as possible the incidence of a disease in patients at risk. A sound predictive analysis encompasses not only the model-training task, but also the aspects of data quality, preprocessing, and fusion during its entire implementation lifecycle to ensure appropriate input data preparation. The robustness of the predictive model and its results depends highly on data quality. Due to the variety of data sources in healthcare environments, it becomes essential to use preprocessing in order to remove noise and inconsistencies. The increasing number of endorsable data exchange standards makes each data exchange achievable, but it demands the implementation of a data-governance program. In addition, the influence of the hospital-database architect on the architecture of an early-diagnosis model is important to guarantee appropriate input-formatting modularity.

Keywords: Predictive Modeling, Machine Learning, Healthcare Analytics, Preventive Actions, High-Risk Patient Detection, Early Diagnosis, Data Quality, Data Preprocessing, Data Fusion, Healthcare Data Sources, Noise Reduction, Data Consistency, Data Exchange Standards, Data Governance, Hospital Database Architecture, Input Formatting Modularity, Model Robustness, Future Event Prediction, Decision Support, Disease Prevention

How to cite this paper:
Rongali, S. K. (2020). Predictive Modeling and Machine Learning Frameworks for Early Disease Detection in Healthcare Data Systems. *Current Research in Public Health, 1*(1), 1–15.
DOI: 10.31586/crph.2020.1355

Received: August 22, 2020
Revised: October 26, 2020
Accepted: November 29, 2020
Published: December 2, 2020

1. Introduction

The hybridization of traditional research methodologies with data-centric approaches has led to the emergence of yet another branch of research that has important real-world applications. Predictive modeling in healthcare data systems is perceived as an advancement of conventional research methodology supported by modern data science tools, techniques, and frameworks. Computational intelligence-based frameworks, particularly predictive modeling methodologies based on supervised machine learning, occupy a prominent role in the advancement of early disease detection in healthcare environments [1]. Accumulating data is one of the principal by-products of the enormity of patient health data generated daily throughout the world. Previous research shows that early detection of health problems can improve the quality of patient care, lower healthcare costs, and save lives. Supervisory early diagnosis frameworks are built on the data-mining principle that supervised learning employs training data that are correctly labeled (diseased or not diseased) and attempts to learn the relationship between the predicting variables with the target variable output. Predictive modeling relies on current and historical data to assist healthcare professionals in detecting illnesses early, potentially lowering mortality rates.

1.1. Overview and Significance of Predictive Modeling in Healthcare

The exploratory analysis of health records has become increasingly appealing and significant due to accrued complexity and volume. Entirely novel challenges arise from the richness and volume of the data, inherent difficulties characteristic of this domain, and practical consequences involved in early detection of diseases and conditions. Establishing a data exploration strategy adaptable to conditions and applied to a specific context becomes fundamental as it allows assessing feasibility [2]. Evaluating the proposed strategies on the domain health records has proven a healthy scope for research. It requires the development of predictive models and data exploration techniques aiming for early detection of diseases and conditions in particular using temporal modeling of sequential data, including clinical and laboratory measures, as well as employing neural networks in search of optimal prediction of time until onset of events.

2. Fundamentals of Predictive Modeling in Healthcare

The characteristics of data generated in the healthcare domain present unique requirements and challenges for modeling, providing an area of rich investigation within the frameworks of predictive modeling and machine learning. The primary focus lies with data contained within electronic healthcare record systems, which use technology to consolidate patients' medical history and other clinical data into structured records [3]. Healthcare data form an exhaustive description of patients throughout the course of their lifespan and across various diagnoses, procedures, and treatments at multiple institutions. Quantities of data extending back many years or even decades make it highly suitable for supervised ML learning tasks aimed at early detection and intervention of disease and the duplication of results obtained with traditional statistical methods. A key consideration in predictive modeling and ML applied to healthcare systems is the choice of evaluation metrics, particularly given the serious consequences of failing to detect danger signals early enough. The severity and expense of the political and economic distortions associated with a protracted pandemic, together with the strong public appeal of institutions like the UK National Health Service, have opened many of these models to public scrutiny and debate [4]. Proposed solutions generally include the estimation of the present state of the system, based on filtering or smoothing, together with its future extension via prediction or forecasting methods, but there is little literature directly addressing the problem of early prediction of serious events in the healthcare context.

Continuous Adaptation & Knowledge Extraction
Increased Data Complexity & & Volume
Establishing a Data Exploration Strategy
Research Scope & Predictive Modeling
Clinical & Societal impact
Adaptable to Conditions
Assessing Feasibility
Early Detection Focus
Neural Networks for Optimal for Optimal Prediction
Enhanced Patient Healthcare Costs
Key Techniques

Figure 1. Exploratory Analysis of Health Records

Equation 1: Early Detection Prediction Score (EDPS)

For any decision threshold *t*, define counts TP, FP, TN, FN.

Precision $P(t)=TP+FPTP, RecallR(t)=TP+FNTP$

Emphasize recall (costly false-negatives) via the F-measure with $\beta>1$: $F\beta(t)=\beta P(t)+R(t)(1+\beta^2)P(t)R(t), \beta=2$ (used in plots).

Reward earlier true-positive alerts. For each truly positive case with predicted lead time *L* (days) define a concave, bounded utility

$$g(L)=1-e^{-kL}\in[0,1], \tag{1}$$

and average it over the true positives flagged at threshold τ_t: $\overline{g}(t)$

Let cohort prevalence be π. The EDPS is

$$EDPS(t)=\pi F\beta(t)\overline{g}(t) \tag{2}$$

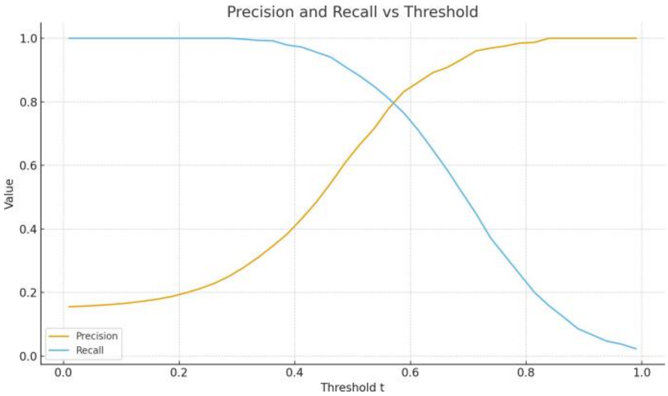

Figure 2. Precision and Recall vs Threshold

2.1. Data Characteristics in Healthcare Systems

Healthcare Predictive Modeling is primarily concerned with inferring the future or conditional class label based on prior observations and is usually built for early detection of diseases. Therefore, the most important property of healthcare data is that it is composed of multiple different disease class labels and is heavily imbalanced since most patients do not have most of these diseases. Disease labels are usually encoded as one-hot vectors where each entry denotes the absence or presence of the disease. Modeling early diagnosis of such diseases and conditions as a multi-label classification problem often yields the best results [5]. The high dimensionality of healthcare data is another characteristic trend. Often, there are thousands of possible attributes risk factors per patient but only a few patients available. Consequently, there are many more parameters than observations and models really perform well if the number of patients available for training is much greater than the number of parameters. However, most healthcare datasets have sparse population distributions and hence the early prediction of health problems is a challenging task due to limited observations in the data when investigated in a multi-label framework where a patient is having multiple and heterogeneous medical conditions, impacting the prediction accuracy. To address these important issues using machine learning, temporal and sequential models are required that can well learn from the data to identify the onset of various diseases. The ultimate goal is to predict the risk of a set of diseases in an individual patient so that preventive care can be initiated before the onset of diseases [6].

2.2. *Evaluation Metrics for Early Detection*

Prediction quality is important in all applications. In the healthcare domain, however, it is even more critical since the costs of any failures can be immense: The “cost” for wrong assumptions, false negatives (N) and false positives (P) are not balanced. Early detection of diseases is crucial (in many situations the disease is not curable) and hence not detecting a disease at the stage of initial development is very important. False negative predictions will invariably result in more severe treatment and are hence predisposed to be associated with higher costs [7]. Conversely, false positive (FP) predictions will typically not increase the monetary cost of the healthcare system (although other costs still may incur, e.g., concern and anxiety for the patients subject to more examination). It is hence expected that healthcare data systems for early prediction will typically target the minimization of false negatives. The objective of the setting can therefore be stated mathematically: N should be as small as possible and N is more important than P. This requirement can be mapped from the general area under the ROC curve (which considers: the trade-off between true positive rate and false positive rate) to balance better with the desire for low false negatives predictions. The metric preferred in this context is the area under the precision-recall curve [8].

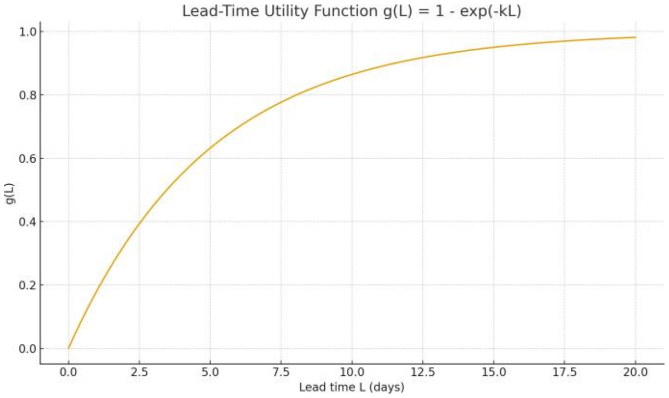

Figure 3. Lead-Time Utility Function g(L) = 1 - exp(-kL)

Equation 2: Clinical Risk Classification Index (CRCI)

Build the precision–recall (PR) curve from sorted scores; compute **Average Precision (AP):**

$AP = \int_0^1 P(R)\ dR$ (area under the PR curve)

Penalize miscalibration with Expected Calibration Error (ECE) using B bins:

$$ECE \ = \ b \ = \ 1\sum_{b=1}^{B} \frac{N_b}{N} \left| empirical\ y - rate_{acc_b} \ - \ mean\ score_{con\ f_b} \right| \tag{3}$$

Define the **CRCI:**

$$CRCI \ = AP \ \cdot \ \left(1 \ - \ ECE\right) \tag{4}$$

3. Machine Learning Frameworks and Architectures

Machine learning (ML) methods that have achieved remarkable performance for supervised learning tasks such as classification and regression are of paramount interest because they enable early prediction of a future event or condition, which is crucial in healthcare. The setting is the early detection of diseases based on periodic health check-testing data such as medical examination records. An individual's health status is usually recorded by binary values indicating whether key indicators (e.g., cholesterol level, blood sugar level, test for pregnancy) are above or below threshold levels [9]. Further, the recorded values are usually unbalanced over classes—individuals are either healthy or sick, and a large majority of the individuals in a population are healthy. Many healthcare problems are temporal in nature because measurements for each subject are acquired over time. Birth-death processes model such phenomena, and models learning the temporal aspects of the data have received increasing attention lately. In such cases, data are presented with timestamps for each of the records and the problem can be described as time-to-event prediction. Although more general than mere classification problems, such a task can nevertheless be solved via supervised learning with suitably adapted architectures and objective functions. As with most other supervised problems, it becomes a simple regression task when the temporal information is removed [10].

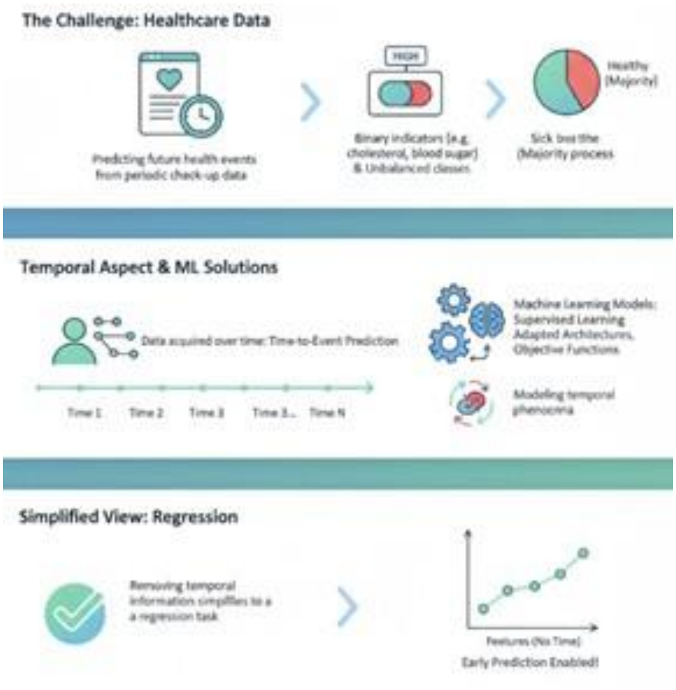
The Challenge: Healthcare Data
HIGH
Healthy (Majority)
Predicting future health events from periodic check-up data
Binary indicators (e.g. cholesterol, blood sugar) & Unbalanced classes
Sick bias the (Majority process
Temporal Aspect & ML Solutions
Data acquired over time: Time-to-Event Prediction
Time 1
Time 2
Time 3
Time 3...
Time N
Machine Learning Models: Supervised Learning Adapted Architectures, Objective Functions
Modeling temporal phenomena
Simplified View: Regression
Removing temporal information simplifies to a a regression task
Features (No Time)
Early Prediction Enabled!

Figure 4. Early Disease Prediction using Machine Learning.

3.1. *Supervised Learning for Early Diagnosis*

For many diseases, the application of supervised learning models to predict disease occurrence with classification accuracy above the chance level is not sufficient. For these outcome classes, it is necessary to apply temporal prediction algorithms that determine the probability of disease occurrence over hourly, daily, or longer time intervals within a specific future time frame. The best trained supervised machine learning algorithm therefore evaluates the probability of disease occurrence within one or more time intervals subject to a certain time constraint. These methods are more appropriate in medical applications than classic classifiers [11]. A necessary condition for successful prediction of early disease occurrence is that the prediction task is solvable at the desired forecasting horizon, which severely limits the application of classical classifiers to a subset of diseases and accidents. Temporal classification is more flexible, yet also riskier, as the algorithms cannot rely on the prior temperature condition of the classifier. Three models are developed for early prediction of surgery: a standard classifier detecting upcoming surgeries, a model predicting any surgery in the upcoming 24h, and a novel temporal classifier pointing in which of the next 12 2-h intervals surgery is expected. Each model is trained and tested on a multi-year dataset. Next, prediction charges are defined, and the desired combinations of classification metrics (sensitivity, precision, etc.) for each use case are empirically determined. The resulting models show that temporal classifiers detecting upcoming surgery are more accurate than standard predictors. Furthermore, reports indicate that for some of the diseases under study, the early-alarm task is actually solvable and superior to standard classifiers [12].

3.2. *Temporal and Sequential Modeling*

Temporal and sequential prediction tasks are found widely in healthcare such as predicting the course of diseases, their problems, progressions, and transitions or predicting comorbidities of patients. A standard supervised classification model would not be applicable as regular data at a fixed point of time do not exist for every patient in the database. The idea is to predict the next event given the history of medical events for a patient, as in next visit prediction, next diagnosis or next procedure with the different treatment stages, or other geo scopes, such as individual patients. A modified recurrent neural network has been applied for next hospital visit prediction [13]. Inverse Reinforcement Learning deals with learning the unknown reward function by observing the behavior of an expert. It also reduces the burden of designing the bias model of the mimicking algorithm by establishing a player who has a clear goal. As health-care data is rich and sequential in nature, it can be ideal for problems in Imitation Learning and Reinforcement Learning procedures like patient treatment protocols. Such treatments are usually not stated clearly by a physician, especially where some parts of the protocol are not procedurally defined but left to the shaping of medical knowledge. Here, the patient provides the game environment and a physician engages with the patient. Hence it is a two-player game, with a patient-character character and a doctor supplement character. The Longitudinal Health data can provide frequent interactions and influence the status of both the players in important sequential health-related decisions. One of the directions taken is to learn the utility function from longitudinal health data and then use it in a reinforcement-learning setup to evolve optimal health-care protocols [14].

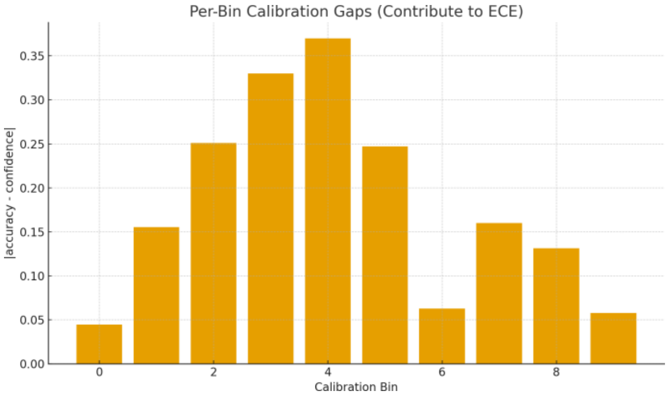

Figure 5. Per-Bin Calibration Gaps (Contribute to ECE)

Equation 3: Model Sensitivity Optimization (MSO)

Choose *t* to maximize recall while meeting clinical constraints:

$$tmax\ s.t.\ R(t)P(t) \geq P_0(minimum\ PPV), \tag{5}$$

$$\overline{g}(t) \geq R_0(minimum\ lead - time\ utility). \tag{6}$$

4. Data Pipelines and Integration in Healthcare Environments

Reliable predictive healthcare systems depend on a high-quality data processing pipeline that rigorously prepares and curates raw input data prior to modeling. Given the sensitive nature of healthcare data and its predominant use for validating datasets, quality checking must be comprehensive. Standardizing data preparation and transformation depersonalizes the data while optimizing integration with the predictive model [15]. Community data systems and Predictive Data Markup Language data management services (e.g. PredML) help streamline these requirements by providing standard protocols and governance while functioning as cloud-ready infrastructures. The integration of Predictive Data Markup Language Datasets with an analytical Modeling Framework, Data Governance, Domain Knowledge Inference Engine (e.g., DRQML), and Quality Grading Infrastructure (e.g., DRQML-QG) supports reliable predictive analytics. Regulation and protocol compliance should therefore be encapsulated into a harmonized open-source software package and locally adopted by healthcare organizations.

4.1. Data Quality and Preprocessing

Healthcare data is often unreliable and thus poses a challenge for predictive modeling. Missing, incongruous, inaccurate, and duplicated values must be detected, corrected, or removed to ensure high-quality data. Preprocessing techniques include, but are not limited to, imputation, outlier detection and correction, normalization, text preprocessing, and even generation [16]. Annotated texts must be relabeled, redundant data preemptively deleted, and dimensions analyzed. Outliers negatively affect a model's predictive power, particularly in supervised learning, while mislabeling may introduce patterns that the model mislearns. Careful attention must also be paid to the selection of data distribution when oversampling techniques are applied. Temporal models often need gap-filling techniques to estimate values for missing time steps or events; imputation and Kalman filtering are two mainstream methods. Large-scale computational infrastructures are also conducive for predictive modeling, as evidenced by frameworks designed specifically for spatiotemporal forecasting. However, any techniques applied must also conform to broad guidelines on temporal data quality [17].

Equation 4: Data Quality Enhancement Score (DQES)

Let baseline error rates b_i and post-preprocessing rates a_i for components $i \in \{missing, label-noise, outliers, drift\}$, with weights $w_i (\sum_i w_i = 1)$.

$$gaini = bi / bi - ai,\ DQES = \sum_i w_i gaini \tag{7}$$

4.2. Interoperability Standards and Data Governance

To generate an effective predictive modeling framework for healthcare, data may be collected from many disparate sources. Data governance frameworks are important to ensure that each data source can be understood, manipulated and trusted by those in the organization, through the imposition of standards for documentation, quality and accuracy. Data governance frameworks leverage guidelines, policies and templates to provide structure for generating data catalogues and ensuring that data profiling occurs at the source of each data set. Interoperability standards must also be assessed at this integration step to ensure the quality and accuracy of models produced by the integration of many data sources [18]. To ensure that heterogeneous data sources can be combined, an idealized reference model for the data is first developed. Data format and semantics for the variables in the model are then established, and varables labelled with ontologies that provide this information. The first step may be the selection of a simple population, for example pediatric patients with health data for example in the European region. Standard data formats are then applied to all sources (e.g. FHIR in the health domain), and these requirements communicated to owners and creators of the data. The second step entails collecting administrative data such as demographics, socio-economics, behaviours and indicators of the living and care environment and defining common variables. At the same time, healthcare services and outcomes (diagnoses, therapies, visits, hospitalizations), lifestyle and health support factors variables defined during the first step of the study are collected, together with labels for mental and physical health. These are supported by a specific, simplified ontology. Resources allowing the recognition of information in various languages are used for the definition of the languages and the conversion to a semantic approach via the explicit labelling by the underlying e-health ontology, elements that makes accessible the definition of the health conditions' labels in the regions involved in the study [19].

5. Model Development Lifecycle and Validation

Progressing through the machine learning model development lifecycle involves partitioning datasets for training and evaluation, implementing cross-validation and calibration routines, ensuring a focus on model explainability and interpretability, and recognizing distinct aspects of performance evaluation concerning early detection [20]. Evaluating selected prediction models requires estimating performance on previously unseen data. A revolutions in computational capability and data availability has rendered cross-validation a common practice when datasets allow many training-testing divisions. The choice of data problematic is however significant, and reserved testing sets exist for such supervision-leaning tasks, while prediction of missing data as a method of cross-validation is particularly representative and widely used. Additionally, the particular nature of predictive modeling in healthcare being frequently concerned with rare events, calibration instead of standard ROC analyses, or further adjustments to assess performance for temporal or sequential prediction with imbalanced data, are beneficial [21]. Considerations governing data splits and performance problems that appear unique to predictive modeling have been often a focus of informatics-oriented studies. In contrast, and possibly because predictive tasks are currently less in vogue in the machine learning community, a less stressed is the importance assigned to explainability and interpretability. Yet tests and the quest for explainable models should not be neglected in the present age of deep learning. Rather, the difficulties these cannot always be answered in supervised-learned models or explained for neural network-like endeavor should be recognized, and more it should be re-emphasized that validated, still-attractive data-collections and data-analyses for supervised learned prediction task remain sizeable [22].

5.1. Data Splitting, Cross-Validation, and Calibration

Data splitting enables the validation and subsequent application of the model to unseen data. In early detection scenarios, testing the model on data from a forward time period (i.e., hold out for a later date) is particularly important. Early detection models are usually trained on temporally ordered data (e.g., information for patients available at time t) in order to prevent information leakage [23]. Data leakage can lead to overoptimistic predictions and unreliable results. Cross-validation is broadly applied to assess the model to avoid overfitting. The common approach is K-fold cross-validation, where data are partitioned into K equally sized segments and the model is trained K times. In each iteration, K-1 segments are used for training and the unseen segment for validation and the performance for each fold is aggregated. When class imbalance is pronounced, stratified K-cross-validation is employed, where the proportion of classes is preserved within each fold [24]. For classification tasks, Brier score, area under the receiver operating characteristic curve (AUC), area under the precision/recall curve (AP) or top K errors are commonly used metrics. The aforementioned evaluation measures will point in different directions if classes are incompatible, are partial (i.e., for some test cases an event will be too late to be treated) or predictions are imbalanced. Hence it is recommended to consider complementary evaluation measures (e.g., Brier score paired with AUC). Recalibration is an additional task recommended for classification model application. Without recalibration, the predictive probabilities during deployment deviate from the true risk. Calibrated predictors during deployment ensure reliable risk estimates [25].

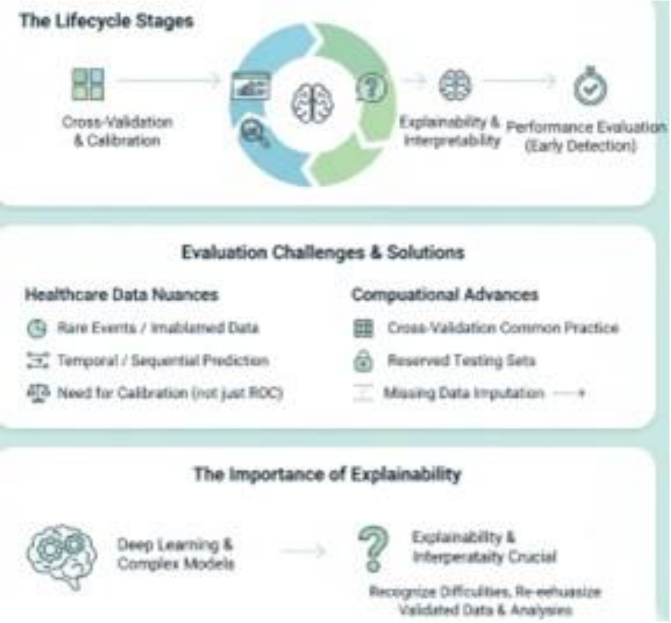
The Lifecycle Stages
Cross-Validation & Calibration
Explainability & Interpretability
Performance Evaluation (Early Detection)
Evaluation Challenges & Solutions
Healthcare Data Nuances
Rare Events / Imablamed Data
Temporal / Sequential Prediction
Need for Calibration (not just ROC)
Compuational Advances
Cross-Validation Common Practice
Reserved Testing Sets
Missing Data Imputation
The Importance of Explainability
Deep Learning & Complex Models
Explainability & Interperatiaity Crucial
Recognize Difficulties, Re-ehuasize Validated Data & Analyses

Figure 6. ML Model Development for Healthcare Prediction

5.2. Explainability and Interpretability

A prediction model is a black box that produces a score and/or a decision for a combination of patient characteristics. Understanding these processes can considerably improve predictive classification. Predictive medicine typically requires addressing the question "Why?" in addition to "What?". Explainability studies the conditions under which a prediction obtained from a complex predictive model, based on perhaps thousands of patient characteristics, can be reconstructed by a simpler predictive model —such as a linear discriminant function— with only a few characteristics. The simpler model often includes automatically selected interaction terms. Such an approach makes it easier to answer both questions: "What is likely to happen?" and "Why?" [26]. Interpretability studies the relation between predictor importance as measured by a predictive model and clinical significance. It investigates how such insights may help physicians, for example, to understand the important predictor regions for high and low mortality probabilities associated with the most accurate predictive model covering a persistent-set danger in cardiovascular surgery.

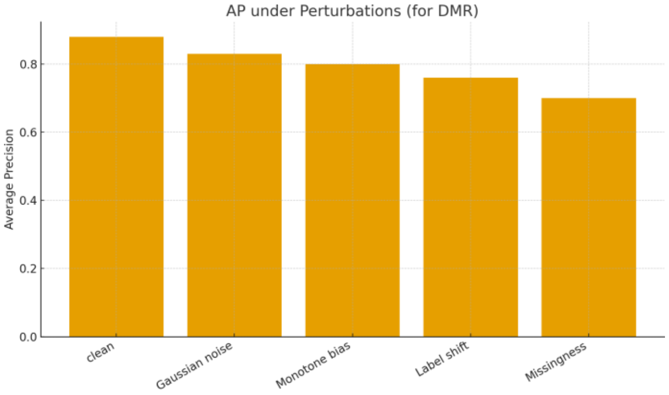

Figure 7. AP under Perturbations (for DMR)

Equation 5: Diagnostic Model Robustness (DMR)

Compute AP on clean data (APcleanAPclean) and under a set of perturbations S (e.g., Gaussian noise, monotone bias, label-shift emulation). Let

$$\Delta rel = max(0, AP_{clean} - \min_{s \in S} AP_s), \tag{8}$$

$$DMR = AP_{clean}(1 - \Delta rel) \tag{9}$$

6. Deployment Considerations in Healthcare Data Systems

Deployment of predictive models in healthcare data systems must consider a broad range of issues that are not normally addressed in other application domains. The architecture of model deployment, monitoring schemes to detect both performance degradation and concept drift of a model, and the ethical and legal regulations related to model development and deployment are all necessities for real-world deployment in healthcare. Ignoring these issues may lead to an untrustworthy model that is of no use in making predictions for unseen instances in the healthcare domain. A predictive model in healthcare is said to be normal only when it satisfies the expectations of healthcare workers as well as the end users [27]. Monitoring the performance of deployed predictive models is critical since their performance can decrease over time because of a change in the underlying data distribution, termed concept drift. Concept drift monitoring means keeping track of the prediction performance of deployed models over time. Change detection strategies can be used to warn when concept drift is present, so that techniques such as retraining or updating can be performed. In addition to concept drift monitoring, models deployed in healthcare data systems must also be transparent and explainable, as model predictions can have an influence on human lives. Therefore, monitoring schemes that track the prediction output of models over time and notify when a model becomes unreliable are also relevant for applications in the healthcare domain [28].

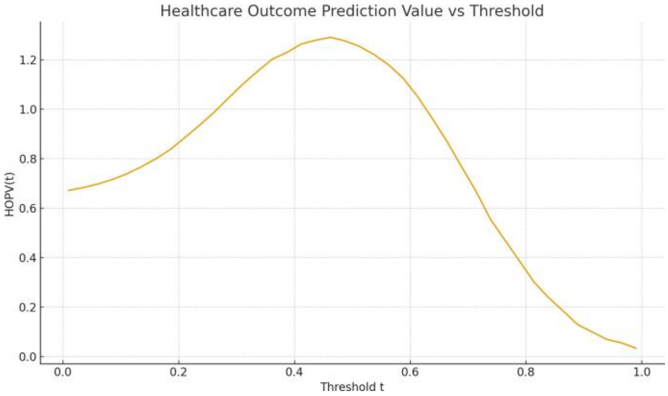

Figure 8. Healthcare Outcome Prediction Value vs Threshold

6.1. Deployment Architectures and Monitoring

Two principal deployment architectures may be considered when data pipelines and predictive models for early disease detection are implemented in a healthcare data system. The first option supports ingestion of new data, with both the predictive model and its related functions running in the data pipeline. This approach enables early predictions in streaming mode as new data records become available. Such forecasts may trigger alerts for patient clinicians or care teams, warning them of potential incidents worthy of further investigation (e.g., signs of a sepsis-episode onset such that a patient effectively seizes) [29]. In contrast, the second deployment architecture actuates a post-processing mode. With trained predictive models made available, these functions can run periodically to query the healthcare data warehouse. Any forthcoming historical records can be ranked by their predicted probabilities for the target diseases, enabling the alerting of care teams about patients in elevation-prone status so that such individuals might undergo extra monitoring or preventive actions. These two methods can also be merged in real-world deployments. An important aspect of deployed predictive solutions is their monitoring. Data drifting throughout time may lead to diminishing performance of the classifiers, thus requiring periodical re-training. Also, new diseases or incidents might arise and become important as the time advances, requiring the inclusion of distinct predictive models [30].

Equation 6: Healthcare Outcome Prediction Value (HOPV)

At threshold t, with benefit per true case B and downstream cost per false-positive C, prevalence π: Recall πR(t), False-Positive Rate πFPR(t).
Per-patient expected utility:

$$HOPV\left(t\right)=\pi R\left(t\right)B-(1-\pi)FPR\left(t\right)C. \tag{10}$$

6.2. Regulatory Compliance and Ethical Implications

Health care monitoring systems use machine learning models to assess the probability of an event. These monitored events may have implications on health, finance, safety, or security. Such events need not be the same across different deploying agents, which may have different fulfillment tasks. For both the monitoring agents and those fulfilling the requests, the deployments are usually in the form of cloud services, enabling scale-up of resources depending on demand. Since events with an impact on health usually require a special consideration, the monitoring agents need to be particularly aware of the ethical implications of deploying ML models. Indeed, in these cases, non-compliance with data governance regulations can initiate either legal or fiscal penalties, or both. Regulatory supervision may not necessarily involve data governance processes, but the established compliance plans provide a roadmap that increases the probability of a successful enforcement [31]. For example, if an event is highly improbable, operation plans usually state that more than a given percentage of false positives is strictly prohibited. The main legal text regulating the production of automatic decisions in health care settings is the General Data Protection Regulation (GDPR) of the European Union. The main principles of the GDPR for automatic decisions are transparency, fairness, and non-discrimination. Transparency translates into the obligation by the deploying agent to inform the individual about the monitored event and collect the consent. The monitoring agent faces the transparency principle twice: when setting up the monitoring entity and when fulfilling the request compliance. Fairness implies that the model shall not discriminate against any individual or group. Non-discrimination more specifically states that the monitoring agent must set up the model so that the predictions produce no unjustified margin of error for each sub-group important for the monitored event [32]. Usually, but not necessarily, these sub-groups are defined by the values of the sensitive variables present in the data, such as age, gender, ethnicity, and other related aspects.

7. Conclusion

The analysis presented within this study confirms the importance of correctly applying predictive modeling techniques in healthcare data systems. Important modeling considerations and practical implementations for select machine learning problem types adapted from the disease detection task have been thoroughly examined [33]. Understanding the unique character of the data, defining suitable predictive modeling evaluation metrics, and selecting an appropriate architecture for model development are essential steps if one is to meaningfully contribute to the body of advanced predictive modeling research. As research matures, additional topics will need to be considered to make predictive modeling suitable for deployment in actual healthcare environments. Integrating data pipelines spanning everyday operations requires ensuring consistent high quality during model training and production. Producing data that conform with health-evidence quality standards requires ongoing management. Populations that vary dramatically in size or characteristics over time necessitate methods for identifying requirement shifts, unsafe conditions during deployment, and triggering automatic retraining. Forecasting patients who are becoming diseased is a different task from forecasting diagnostic test results, and maintaining model reliability across these categories requires vigilance. Integrated solutions must, of course, be built to comply with legislative requirements and address the ethical implications of health-related modeling, ideally minimizing human risk and leveraging the positive aspects of both business intelligence and dark knowledge [34].

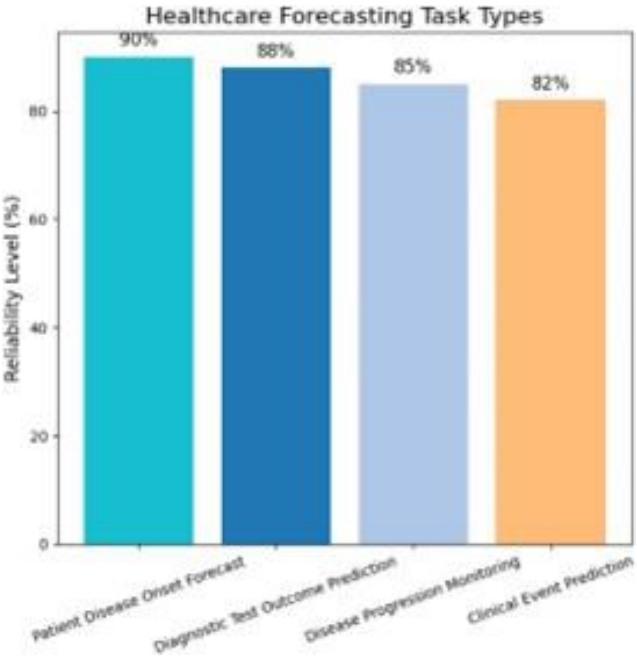

Figure 9. Healthcare Forecasting Task Types

7.1. Summary and Future Directions in Predictive Healthcare Modeling

The application of predictive modeling and machine learning in healthcare research is extensive and dynamically evolving. However, the implementation into clinical practice for real-time decision making is still limited. Furthermore, a large number of healthcare models focus on predicting disease outcomes instead of predicting diseases prior to onset and prior to symptoms. Hence, the application, design, and testing of predictive models for disease conditions that are capable of predicting no disease versus a disease are emphasized here. More specifically, the discussion directs attention to the requirements of predictive models that support early disease detection for successful preventive care, including fundamental characteristics of the underlying data, temporal aspects, and related machine learning frameworks. This discussion also addresses the complete model application lifecycle, from data preparation and exploration, through model development including model evaluation and comparison, to deployment and model use. Particular emphasis is given to data quality checks, explainability, and interpretability. Consequently, predictive models that support the early detection of major diseases in order to enable an effective and motivating preventive care that really benefits patients become more accessible. Implementations of such predictive models now increasingly benefit from mature architectures and technologies, ensuring that a vast number of predictively labeled data sets will become available very soon [35].

References

[1] Kummari, D. N. (2020). Machine Learning Applications in Regulatory Compliance Monitoring for Industrial Operations. Global Research Development (GRD) ISSN: 2455-5703, 5(12), 75-95.

[2] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2020). Internet of Things: Enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, 22(1), 127-176. https://doi.org/10.1109/COMST.2019.2963315

[3] Chakilam, C., Koppolu, H. K. R., Chava, K. C., & Suura, S. R. (2020). Integrating Big Data and AI in Cloud-Based Healthcare Systems for Enhanced Patient Care and Disease Management. Global Research Development (GRD) ISSN: 2455-5703, 5(12), 19-42.

[4] Alharbi, F., & Atkinson, K. (2020). Adoption of FHIR-based interoperability in health information systems. Journal of Biomedical Informatics, 107, 103476. https://doi.org/10.1016/j.jbi.2020.103476

[5] Arpaci, I., Kanat-Maymon, Y., & Baloglu, M. (2020). The impact of security and privacy concerns on cloud computing adoption in healthcare. Computers in Human Behavior, 109, 106349. https://doi.org/10.1016/j.chb.2020.106349

[6] Botlagunta, P. N., & Sheelam, G. K. (2020). Data-Driven Design and Validation Techniques in Advanced Chip Engineering. Global Research Development (GRD) ISSN: 2455-5703, 5(12), 243-260.

[7] Bender, D., & Sandler, J. (2020). HL7 FHIR: An emerging standard for interoperability in healthcare systems. Journal of the American Medical Informatics Association, 27(5), 712-719. https://doi.org/10.1093/jamia/ocaa023

[8] Boonstra, A., Ewout, R., & Broekhuis, M. (2020). Barriers to electronic health record use: A systematic literature review. Health Informatics Journal, 26(3), 1891-1909. https://doi.org/10.1177/1460458219839617

[9] Gadi, A. L. (2020). Evaluating Cloud Adoption Models in Automotive Manufacturing and Global Distribution Networks. Global Research Development (GRD) ISSN: 2455-5703, 5(12), 171-190.

[10] Chen, M., Hao, Y., Cao, J., Li, Y., & Zhang, L. (2020). AI-enabled healthcare: Data fusion for smart IoT and cloud platforms. IEEE Network, 34(5), 78-83. https://doi.org/10.1109/MNET.011.1900594

[11] Cruz, E., & Chowdhury, M. (2020). Microservices in healthcare systems: Challenges and opportunities. IEEE Software, 37(3), 56-63. https://doi.org/10.1109/MS.2020.2973351

[12] Pandiri, L., Singireddy, S., & Adusupalli, B. (2020). Digital Transformation of Underwriting Processes through Automation and Data Integration. Global Research Development (GRD) ISSN: 2455-5703, 5(12), 226-242.

[13] Dinh-Le, C., Chuang, R., Chokshi, S., & Diev, S. (2020). The role of API platforms in digital health ecosystems. Digital Health, 6, 205520762090275. https://doi.org/10.1177/2055207620902756

[14] Esposito, C., & De Santis, A. (2020). Blockchain-based secure data management for healthcare. IEEE Transactions on Industrial Informatics, 16(10), 6348-6359. https://doi.org/10.1109/TII.2020.2967449

[15] Lakkarasu, P. (2020). Scalable AI Infrastructure: Architecting Cloud-Native Systems for Intelligent Workloads Scalable AI Infrastructure: Architecting Cloud-Native Systems for Intelligent Workloads. Global Research Development (GRD) ISSN: 2455-5703, 5(12), 133-151.

[16] Fan, K., Gong, Y., Li, H., & Yang, Y. (2020). Lightweight and secure IoT communication for smart healthcare. IEEE Transactions on Industrial Informatics, 16(8), 5565-5574. https://doi.org/10.1109/TII.2019.2952422

[17] Farahani, B., Firouzi, F., Chang, V., Badaroglu, M., Merrett, G., & Saracco, R. (2020). Towards fog-driven IoT healthcare. Future Generation Computer Systems, 109, 1-17. https://doi.org/10.1016/j.future.2020.03.020

[18] Meda, R. (2020). Real-Time Data Pipelines for Demand Forecasting in Retail Paint Distribution Networks. Global Research Development (GRD) ISSN: 2455-5703, 5(12).

[19] Fiaidhi, J., & Khatib, E. (2020). From HL7 to FHIR: Transforming healthcare data exchange. IT Professional, 22(3), 23-31. https://doi.org/10.1109/MITP.2020.2980706

[20] Fomundam, S.,& Tavakoli, N. (2020). DevSecOps metrics for secure API development. Journal of Systems and Software, 167, 110617. https://doi.org/10.1016/j.jss.2020.110617

[21] Somu, B. (2020). Transforming Customer Experience in Digital Banking Through Machine Learning Applications. International Journal Of Engineering And Computer Science, 9(12).

[22] Inala, R. (2020). Big Data-Driven Optimization of Retirement Solutions: Integrating Data Governance and AI for Secure Policy Management. Global Research Development (GRD) ISSN: 2455-5703, 5(12).

[23] Gaur, L., Jain, A., & Kumar, A. (2020). Blockchain and API integration for secure electronic health records. Computer Methods and Programs in Biomedicine, 189, 105341. https://doi.org/10.1016/j.cmpb.2019.105341

[24] Ghosh, S., & Paul, S. (2020). API-led integration strategies for scalable IoT and healthcare systems. IEEE Access, 8, 110204-110215. https://doi.org/10.1109/ACCESS.2020.2998512

[25] Hasanzadeh, A., & Pahl, C. (2020). Containerization for cloud-native applications: A survey. Journal of Cloud Computing, 9(1), 45. https://doi.org/10.1186/s13677-020-00205-3

[26] He, Y., & Zhou, J. (2020). API management and security in cloud-based healthcare systems. IEEE Cloud Computing, 7(4), 45-52. https://doi.org/10.1109/MCC.2020.2994622

[27] Huang, Y., & Liu, X. (2020). FHIR-based API development for patient-centric healthcare applications. Health Informatics Journal, 26(4), 2723-2738. https://doi.org/10.1177/1460458220942583

[28] Iqbal, A., & Ahmad, M. (2020). Secure API gateway design for multi-tenant healthcare systems. Computer Networks, 180, 107393. https://doi.org/10.1016/j.comnet.2020.107393

[29] Jiang, F., & Ruan, Y. (2020). Big data integration and API-driven inter-operability for healthcare analytics. Information Systems, 91, 101488. https://doi.org/10.1016/j.is.2020.101488

[30] Kaur, P., & Singh, M. (2020). Cloud-native healthcare applications using microservices architecture. Journal of Network and Computer Applications, 149, 102467. https://doi.org/10.1016/j.jnca.2019.102467

[31] Kaur, R., & Kaushik, A. (2020). Security frameworks for API-driven cloud integration in healthcare. Computers & Security, 94, 101856. https://doi.org/10.1016/j.cose.2020.101856

[32] Khanna, S., & Kandpal, S. (2020). Hybrid cloud architecture for interoperable EHR systems. IEEE Transactions on Cloud Computing, 8(6), 1575-1587. https://doi.org/10.1109/TCC.2020.2978440

[33] Kumar, N., & Tripathi, R. (2020). Healthcare interoperability through API-based middleware. IEEE Access, 8, 141769-141782. https://doi.org/10.1109/ACCESS.2020.3006018

[34] Li, W., & Tang, X. (2020). Continuous integration and deployment for cloud microservices in healthcare. Software Quality Journal, 28(5), 2311-2331. https://doi.org/10.1007/s11219-020-09525-7

[35] Luo, J., Wu, M., & Zhao, Y. (2020). FHIR-based semantic interoperability for clinical decision support. BMC Medical Informatics and Decision Making, 20(1), 223. https://doi.org/10.1186/s12911-020-01244-1