

Healthcare Disease Prediction Using Machine Learning: A Comprehensive Review of Prediction Models, Empirical Evidence, and Clinical Decision-Support Challenges

Pratik Shankarrao Shirke, Prof. A. P. Bangar, Prof. S. B. Bhosale, Prof. V. A. Karad

Department of Computer Engineering
Jaihind College of Engineering, Kuran, Maharashtra, India

Abstract — Chronic and life-threatening illnesses are frequently found late, and assessing risk by hand is slow and uneven across large groups of patients. This paper offers a focused narrative review that brings together five studies on machine-learning-based disease prediction published by IEEE and Elsevier: a convolutional-network risk model built on regional hospital data; a hybrid random-forest and linear model for heart disease; a preprocessing-and-ensemble framework for diabetes; a smart monitoring system that combines sensor data and medical records through ensemble deep learning; and a study of breast tumour classification. The review follows three connected themes: the data sources and prediction models used, how strong and how comparable the reported evidence is, and the conditions that matter when prediction is carried into a clinical decision-support pipeline. Only one study could be read in full. It reports an area under the ROC curve (AUC) of 0.950, with sensitivity 0.789 and specificity 0.934, for a weighted ensemble on the PIMA Indians Diabetes dataset. The other four were assessed from their published abstracts and report, among other results, accuracies of 94.8%, 88.7% and 98.5%; these figures involve different diseases, data and metrics, so they cannot be ranked against one another. The review contributes a comparative table, a conceptual architecture for the proposed disease prediction system (not experimentally validated), and a prioritised agenda of research gaps.

Keywords— Disease prediction, machine learning, healthcare analytics, heart disease, diabetes, breast cancer, ensemble learning, clinical decision support.

I. INTRODUCTION

Chronic diseases such as heart disease and diabetes put a heavy load on patients and health systems, and early prediction is the shared goal of the reviewed studies. Mohan et al. call heart disease one of the leading causes of death and describe its prediction as a major challenge in clinical data analysis [2]. Hasan et al. point out that the risk and severity of diabetes can be cut considerably if early prediction is accurate [3], and Chen et al. state that careful analysis of medical data supports early detection, patient care and community services [1]. Manual diagnosis and risk assessment, in contrast, take time and may be inconsistent across large patient populations, as the seminar synopsis observes.

The project behind this review is a Healthcare Disease Prediction system based on machine learning. According to the seminar synopsis, it is a pipeline that takes in patient demographics, clinical test results, lifestyle information, symptoms and medical features from electronic health records (EHRs) or biopsy datasets; cleans and normalises them; selects

relevant features; divides the data into training and testing sets; trains models such as SVM, Random Forest and Neural Networks; evaluates them with suitable metrics; and predicts the likelihood or class of a disease such as diabetes, heart disease, or a malignant versus benign tumour. Its stated aims are earlier and more accurate diagnosis, data-driven support for doctors, scalable prediction of several diseases, and better management of hospital resources. Any technology proposed in this setting therefore has two jobs: it must be accurate on the cases that matter, and it must present its output in a form that a clinician can interpret and act on.

This paper reviews five studies that together cover a convolutional-network risk model on hospital data [1], a hybrid random-forest model for heart disease [2], a preprocessing-and-ensemble framework for diabetes [3], an ensemble deep-learning monitoring system with feature fusion [4], and a breast-cancer classification study [5]. Three questions guide the review. RQ1: Which data sources and prediction models are used? RQ2: What performance and trade-offs do the studies report, and how dependable is the evidence? RQ3: What do the

studies imply for a pipeline that moves from prediction to decision support? The paper closes with a comparative table, a conceptual architecture and a list of research gaps.

II. REVIEW METHODOLOGY

This is a narrative review of five papers chosen for the study, not a systematic database search, so no claim of exhaustive coverage is made. The papers are three IEEE Access journal articles [1]–[3], one article from the Elsevier journal Information Fusion [4], and one IEEE conference paper [5]. They were selected to span heart disease, diabetes, breast cancer and a multi-source regional chronic-disease setting, in line with the disease areas named in the seminar synopsis. For each paper we extracted and compared the aim, method, data, reported results and stated limitations.

The depth of access differs between papers, and we state it here so that nothing is overstated. Paper [3] was read in full, and its bibliographic details are those of the IEEE-published version. Papers [1], [2], [4] and [5] were assessed only from their published abstracts and bibliographic records; details such as dataset size, hyperparameters, validation protocol or limitations that these abstracts leave out are marked “not reported in the abstract” and not guessed. All figures quoted are those given by the original authors. Studies cited inside [3] are described only as [3] describes them and were not re-read. Because the evidence is uneven, conclusions about clinical deployment are inferences and are labelled as such.

III. TECHNICAL FOUNDATIONS

1. Data sources and input modalities

The five studies draw on quite different kinds of data. Chen et al. use real hospital data gathered in central China during 2013–2015, combining structured records with unstructured hospital text and using a latent factor model to rebuild missing data [1]. Hasan et al. use the public PIMA Indians Diabetes dataset of 768 female patients (268 diabetic and 500 non-diabetic) with eight attributes [3]. Ali et al. merge features taken from sensor data and electronic medical records to form the input for heart disease prediction [4]. Mim et al. gather cell features from breast cancer patients and from normal persons in order to label tumours as malignant or benign [5]. The abstract of [2] does not name its dataset. These observations support a design that treats structured clinical data as the base input and adds unstructured or sensor data as an optional extension; that design choice belongs to the project and is not a finding of the reviewed papers.

2. Prediction models

The five studies cover classical, ensemble and deep approaches. Paper [3] compares classical classifiers (k-nearest neighbour, decision tree, random forest, naive Bayes) with the boosting methods AdaBoost and XGBoost and a multilayer perceptron (MLP). Mohan et al. join random forest with a linear model into a hybrid method (HRFLM) after a search for significant features [2]. Chen et al. propose a convolutional neural network for multimodal disease risk prediction (CNN-MDRP) and compare it with a unimodal variant, CNN-UDRP [1]. Ali et al. train an ensemble deep-learning model on fused features [4]. The seminar synopsis names SVM, Random Forest and Neural Networks; support vector machines are not evaluated in [3] and are not named in the abstracts of [1], [2], [4] or [5], so the project will have to test them directly.

3. Preprocessing, validation and evaluation metrics

Data preparation is where the studies differ most. Hasan et al. reject outliers with an interquartile-range rule, fill missing values with the mean, standardise the data, select features (PCA, ICA or correlation-based) and use stratified 5-fold cross-validation with grid-search tuning in an inner loop [3]. Chen et al. deal with incomplete data through a latent factor model [1]. The metrics differ too. Hasan et al. report sensitivity, specificity, precision, false omission rate (FOR), diagnostic odds ratio (DOR) and AUC [3], whereas the abstracts of [1], [2] and [4] report accuracy. Accuracy depends on class balance: in the PIMA data, 500 of 768 patients (about 65%) are non-diabetic, so a model that always answered “no diabetes” would score roughly 65% accuracy (the reviewer’s own calculation from the class counts in [3]). Sensitivity, specificity and AUC are therefore more informative for screening, and they cannot be swapped for accuracy.

IV. PREDICTION MODELS AND SYSTEM ELEMENTS

1. Models compared in the reviewed papers

Chen et al. experiment with modified prediction models on hospital data and propose CNN-MDRP, which uses both structured and unstructured data; to the best of their knowledge, no earlier work in medical big data analytics used both types, and the model reaches 94.8% accuracy with faster convergence than CNN-UDRP [1]. Mohan et al. propose HRFLM, a hybrid of random forest and a linear model meant to raise accuracy through significant features [2]. Hasan et al. test k-NN, decision tree, random forest, AdaBoost, naive Bayes, XGBoost and an MLP, and propose an ensemble in which each model’s vote is weighted by its AUC; the best combination was AdaBoost with XGBoost [3]. Ali et al. propose a smart healthcare system built

on ensemble deep learning and feature fusion [4]. Mim et al. apply supervised classifiers to tell malignant tumours from benign ones [5]. The abstract of [5] does not say which classifier did best.

2. Data and hardware

Only [3] describes its data and hardware in detail. It uses the PIMA Indians Diabetes dataset of 768 female patients from the Pima Indian population near Phoenix, Arizona, with eight attributes. The models were written in Python with Keras and run on a Windows 10 machine with an Intel Core i7-7700HQ processor, 16 GB of RAM and a GeForce GTX 1060 GPU with 6 GB of memory. The MLP was trained for 200 epochs with a learning rate of 0.001, a batch size of 8 and 60% dropout. Chen et al. report hospital data from central China for 2013–2015 and an experiment on the regional chronic disease cerebral infarction [1]. For [2], [4] and [5], dataset size, train–test split and hardware are not reported in the abstracts.

3. Cautions

Several cautions come up repeatedly in [3]. Data quality matters: for the MLP, rejecting outliers together with mean imputation raised AUC by 7.1% over the raw data, whereas rejecting outliers alone did not help because too many samples were lost. Standardisation did not improve the tree-based models and could discard information in a small dataset. Once outliers were removed and missing values filled, correlation-based feature selection beat PCA and ICA for the classical models, and diastolic blood pressure and the diabetes pedigree function were judged to carry less information. Deeper networks did not help: the best MLP had three hidden layers (16, 64 and 64 neurons) and reached AUC 0.902, below XGBoost at 0.946 ± 0.020 , a limit the authors put down to the small sample. Combining all six classical models was not better by AUC than the AdaBoost-plus-XGBoost pair. Finally, a sensitivity of 0.789 means that about one in five diabetic patients (roughly 21%, the reviewer's calculation as $1 - 0.789$) would still be missed. These are failure modes that a clinical decision-support system should anticipate and should surface for human review, not act on automatically.

V. EVIDENCE OF PERFORMANCE

1. Reported results

In [3], the AUC-weighted ensemble of AdaBoost and XGBoost gave sensitivity 0.789, specificity 0.934, FOR 0.092, DOR 66.234 and AUC 0.950, which the authors say beats state-of-the-art results on the same dataset by 2.00% in AUC. The best single model was XGBoost with AUC 0.946 ± 0.020 , so the ensemble improved XGBoost's AUC by only 0.6%; the

ensemble's precision was 84.2% [3]. Chen et al. report 94.8% accuracy for CNN-MDRP [1]. Mohan et al. report 88.7% accuracy for HRFLM [2]. Ali et al. report 98.5% accuracy for their ensemble deep-learning system and describe it as higher than existing systems [4]. The abstract of [5] says classifiers were applied but gives no performance figures.

2. Accuracy versus complexity

The studies show a trade-off between model complexity, data needs and the size of the gain. In [3], a carefully preprocessed boosting model reached AUC 0.946, a two-model ensemble added only 0.6%, and the deeper neural network was weaker at 0.902, so extra complexity did not buy accuracy on 768 records. Chen et al. report faster convergence for the multimodal CNN than for the unimodal one, alongside the accuracy figure [1]. Ali et al. pair feature fusion with ensemble deep learning to cope with high-dimensional data from several sources [4]; their abstract does not report training cost. The lesson for the proposed system is that accuracy, sensitivity, training cost and data requirements must be reported together.

3. Comparability and quality of the evidence

The 98.5% accuracy in [4] is higher than the 94.8% in [1], the 88.7% in [2] and the AUC of 0.950 in [3], but this does not show that [4] is the best method. The studies address different diseases (heart disease, cerebral infarction, diabetes, breast cancer), use different data and report different metrics, and an accuracy and an AUC are not the same kind of number. Whether [1], [2] and [4] were evaluated on held-out or externally collected patients is not stated in their abstracts. Only [1] uses real hospital records from a specific region and period, and its authors note that regions have distinct disease characteristics, so the results may not carry over elsewhere. Paper [3] uses a public dataset from a single population, so generalisation to other populations is unknown (an inference, not a finding of [3]). Paper [5] reports no figures in its abstract, so its strength cannot be judged.

VI. FROM PREDICTION TO CLINICAL DECISION SUPPORT: THE COMMUNITY-DATA AND MONITORING STUDIES

Of the five papers, [1] and [4] come closest to a system that takes prediction toward a care setting. Paper [1] targets disease risk and outbreak prediction in disease-frequent communities using real hospital data, and [4] proposes a smart healthcare monitoring system that fuses sensor data with medical records. Paper [3] offers the most transparent training pipeline. The abstracts of [1] and [4] do not describe how a prediction reaches

a clinician, so conclusions about decision support are inferences and are labelled as such.

1. Findings

Chen et al. show that incomplete hospital data can be handled by reconstructing missing values, and that using structured and unstructured data together is feasible for risk prediction [1]. Ali et al. show that fusing sensor-derived features with electronic medical records can feed an ensemble deep-learning predictor [4]. Hasan et al. give a pipeline with outlier handling, imputation, feature selection, stratified cross-validation with inner-loop tuning and several complementary metrics [3]. The inner- and outer-loop design in [3] is the closest published counterpart, among the five papers, to the training-and-testing stage of the project's pipeline. It is evaluated on one public dataset, so it is a design principle and not proof of clinical performance.

2. Design issues

Hasan et al. plan to build a web application with a user-friendly interface on top of the trained model and to apply the framework to other medical contexts [3]. Mohan et al. observe that earlier studies give only a glimpse of heart disease prediction with machine learning, and they aim to find significant features [2].

Chen et al. stress that incomplete medical data lowers the accuracy of analysis [1]. None of these abstracts, nor the full text of [3], describes a clinician-facing explanation of a prediction.

3. Limits and transferability

The seminar aims at a decision-support tool for doctors across several diseases, while each reviewed paper targets one disease and reports offline model performance.

The abstracts of [1], [2], [4] and [5], and the full text of [3], report no prospective clinical validation, calibration, clinician time saved, interpretability testing or fairness analysis across patient groups. Whether the results transfer to the hospital data of another region, population or recording standard remains untested.

VII. COMPARATIVE ANALYSIS

Table 1 summarises the five reviewed studies. The findings are not directly comparable because disease, data and metrics differ. Entries marked "abstract only" come from the published abstract.

Table 1. Comparison of the five reviewed studies

Study	Objective and design / data	Main findings and limitations
Chen et al. (2017) [1]	Aim: predict chronic disease outbreaks in disease-frequent communities Design: real hospital data from central China (2013–2015); latent factor model for missing data; CNN-MDRP on structured and unstructured data; cerebral infarction (abstract only)	Findings: accuracy 94.8%; faster convergence than CNN-UDRP Limits: data split, sample size and external validation not given in the abstract
Mohan et al. (2019) [2]	Aim: improve heart disease prediction by finding significant features Design: HRFLM, a hybrid of random forest and a linear model (abstract only)	Findings: accuracy 88.7% Limits: dataset, validation protocol and baselines not given in the abstract
Hasan et al. (2020) [3]	Aim: robust diabetes prediction Design: PIMA data, 768 patients; outlier rejection, mean imputation, feature selection; k-NN, DT, RF, AdaBoost, NB, XGBoost, MLP; AUC-weighted ensemble; stratified 5-fold CV (full text)	Findings: sensitivity 0.789, specificity 0.934, FOR 0.092, DOR 66.234, AUC 0.950; XGBoost 0.946; MLP 0.902 Limits: small single-population dataset; about 21% of diabetic cases missed; no external validation
Ali et al. (2020) [4]	Aim: heart disease prediction within a smart healthcare system Design: feature fusion of sensor data and medical records; ensemble deep learning; compared with traditional classifiers (abstract only)	Findings: accuracy 98.5% Limits: dataset and validation protocol not given in the abstract

Study	Objective and design / data	Main findings and limitations
Mim et al. (2022) [5]	Aim: classify malignant and benign breast tumours Design: supervised classifiers on collected cell features; Bangladesh context (abstract only)	Findings: none stated in the abstract accessed Limits: dataset size and per-classifier results not stated

VIII. PROPOSED CONCEPTUAL ARCHITECTURE

Figure 1 gathers the design implications of the reviewed studies into a staged architecture for the proposed disease prediction system. It is a conceptual proposal derived from the literature and has not been tested experimentally.



Fig. 1. Proposed conceptual architecture for the healthcare disease prediction system (conceptual, not experimentally validated)

Stages supported by the reviewed papers. Stage 1 collects structured patient data, as in [3], with optional unstructured text or sensor features as in [1] and [4]. Stage 2 prepares the data; [3] rejected outliers with an interquartile-range rule and filled missing values with attribute means, and [1] rebuilt missing data with a latent factor model. Stage 3 selects features; correlation-based selection worked best for the classical models in [3]. Stage 4 splits the data; [3] used stratified 5-fold cross-validation with inner-loop grid search. Stage 5 trains models; XGBoost, AdaBoost and their AUC-weighted ensemble performed best in [3], while the choice among SVM, random forest and neural networks, all named in the synopsis, should be tested on the project's own data. Stage 6 evaluates with sensitivity, specificity, FOR and AUC as in [3], not accuracy alone.

Stages that are the author's own design recommendations. None of the five studies evaluates stages 7 or 8, nor de-identification in stage 2 or calibration in stage 6. Stage 7 shows the predicted risk, the contributing features and an uncertainty flag to a clinician, because a person should review borderline cases and no reviewed study tests clinical use. Stage 8 records the clinician's decision and the later outcome, and the dashed loop returns reviewed cases for model updating, which the synopsis describes as continuous improvement. A decision threshold that favours sensitivity is motivated by the 0.789 sensitivity in [3].

IX. RESEARCH GAPS AND FUTURE DIRECTIONS

Gaps supported by the reviewed papers. (i) Generalisation: [3] uses a single public dataset of 768 patients from one population and [1] uses hospitals from one region; none of the sources reports testing at other sites. (ii) Missed cases: a sensitivity of 0.789 in [3] shows that a high AUC can coexist with many missed patients. (iii) Inconsistent evaluation: accuracy in [1], [2] and [4], AUC and sensitivity in [3], and no figures in the abstract of [5], so results cannot be ranked [1]–[5]. (iv) Data quality: [1] and [3] both show that missing values and outliers shape results. (v) Multimodal data: only [1] and [4] combine more than one kind of data, and their abstracts give no breakdown of what each source adds. (vi) Model coverage: SVM, named in the synopsis, is not evaluated in [3]. (vii) The prediction-to-care loop: no reviewed paper reports clinician use, calibration or outcome follow-up.

Design recommendations (the author's own, not findings of the papers). Fit imputation and scaling inside each training fold to avoid leakage; report sensitivity, specificity, AUC and a

confidence interval alongside accuracy; test on data from a second source; remove direct identifiers and restrict access to stored records; check performance across age and sex groups; choose the decision threshold together with clinicians; and keep a human review step before any diagnosis or treatment decision. The seminar's claims of multi-disease scalability and continuous improvement depend on these tests.

Directions requiring investigation. Priorities include a retrospective study of diabetes, heart-disease and breast-tumour datasets under one shared protocol; head-to-head tests of SVM, random forest, a neural network and boosting on the same folds and metrics; a study of whether an AUC-weighted ensemble beats the best single boosted model by more than the 0.6% seen in [3]; evaluation of decision thresholds for screening; and external validation on data from a partner hospital.

X. CONCLUSION

Across the five reviewed studies, machine-learning models can predict disease from clinical data with encouraging results. The best-documented result, from [3], shows an AUC of 0.950 with sensitivity 0.789 for an AUC-weighted ensemble on the PIMA diabetes data, and [1], [2] and [4] report accuracies of 94.8%, 88.7% and 98.5% on different data. The evidence is nevertheless uneven: only one paper was read in full, the metrics differ, and none reports prospective clinical use. For the proposed healthcare disease prediction system, this points to a modular pipeline in which preprocessing, feature selection, model training and evaluation can each be replaced independently, with sensitivity-aware evaluation and clinician review of uncertain predictions. Evaluation on data from more than one source, with calibration and subgroup performance measured, is needed before the system can be called dependable.

REFERENCES

1. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.
2. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
3. M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.
4. F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Inf. Fusion*, vol. 63, pp. 208–222, 2020.
5. R. B. Mim, A. Sattar, A. B. Islam, and S. Roy, "Breast cancer detection using machine learning approach," in *Proc. 2022 Int. Conf. Elect., Comput. Energy Technol. (ICECET)*, Prague, Czech Republic, Jul. 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9872893>