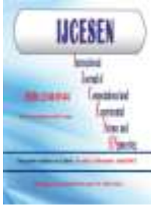
International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)

Vol. 11-No.3 (2025) pp. 4636-4641
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Statistical and AI-Based Hybrid Modeling for Predicting Chronic Disease Progression

Khalid Talib Othman*

Aliraqia University College of Medicine- Iraq
***Coreesponding Author Email:** khalid.t.othman@aliraqia.edu.iq - **ORCID:** 0009-0008-2452-0402

Article Info:

DOI: 10.22399/ijcesen.2791
Received : 11 April 2025
Accepted : 19 June 2025

Keywords

Chronic Diseases
Artificial Intelligence
Statistical Modeling
Supervised Learning
Health Prediction

Abstract:

Non-communicable diseases such as diabetes, hypertension, and cardiovascular diseases are a growing worldwide health concern. Early-stage prediction of disease progression is critical for improving clinical outcomes and reducing healthcare costs. Conventional statistical models may assist in highlighting significant risk predictors, yet cannot suitably model high-dimensional, nonlinear construct the large clinical datasets. Machine learning methods, such as Random Forests, Support Vector Machines and Neural Networks, provide very high prediction power but tend to provide low interpretability.This study aims to develop hybrid model for chronic illness progression forecasting by supervised learning methods using statistical and AI-based approaches. We will use freely available datasets such as the UCI Chronic Disease, NHANES, and MIMIC-III. The study will compare the predictive performance of logistic regression and Cox regression against the existing standard of care (the AI models) but also against a new model that integrates aspects of both approaches.In this study, feature encoding techniques and imbalance data fixing methods such as SMOTE, etc. have been used. The evaluation metrics that will be used to assess model performance include accuracy, precision, recall, F1-score, and ROC-AUC. The result is expected to show that hybrid models can offer better predictive performance while keeping statistical interpretability.This research helps develop transparent and efficient clinical decision-support systems that provide medical staff with realistic opportunities for early detection of risk and the tailored preparation of care.

1. Introduction

1.1 Research Background

National healthcare systems across the world are broken, chronic diseases like diabetes, cardiovascular diseases, and hypertension being some of the hardest to treat and therefore most expensive problems. These conditions may progress slowly and require long-term management, leading to a significant burden on patients and health systems. More accurate predictions of disease progression could enable clinicians to intervene sooner, resulting in fewer hospitalizations and improved overall outcomes for patients.

Statistical modeling has long been considered a cornerstone of epidemiology and clinical research for quantifying the impact of relevant risk factors and relationships between variables. Because of their interpretability and strong theoretical basis, models like logistic regression and Cox proportional hazards are preferred. But it is always linear and independent and has limited power in the diverse and high-dimensional real-life datasets which can disrupt with nonlinear interactions.

A supervised deep learning model thus emerges as a powerful alternative. Many algorithms, such as Random Forests, Support Vector Machines (SVM), and Neural Networks, have demonstrated good predictive power in medicine on various clinical data in multiple fields, specifically, medical diagnosis and prognosis. These models can learn the distribution of the data without assuming a specific överlying structure. However, particular AI models could not be interpretable, which brings up challenges in clinical decision-making.

If done well, this blend of statistical analysis and AI can yield valuable insights in both world. Statistical techniques ensure interpretability and introduction to causal relationships, while artificial intelligence builds prediction power and deals with complex, nonlinear relationships. This hybrid framework is ideal for chronic disease progression predictions where both understanding and accuracy are critical.

We study approaches to bridge the gap between statistical modelling of patient pathways, ‘explainable’ AI based supervised learning of propensity scores, and robust, accurate interpretable tools for the implementation of early detection and optimised management of chronic diseases in healthcare systems.

1.2 Research Problem

Commonly used statistical models in medical research, such as logistic and Cox regression, were widely adopted owing to their interpretability and theoretically-firm basis. However, these models are limited by the assumptions of linearity, normality and independence of the variables. As such they underestimate the performance when being applied to high-dimensional datasets, and when modelling complex nonlinear interactions between clinical, demographics and behavioural factors.

In contrast, AI and supervised machine learning methods such as Random Forests, Support Vector Machines and Neural Networks offer well-established techniques for extracting actionable insights from large and complex data collections. Those models are very good at discovering latent structure and nonlinear relationships. But they are often “black boxes,” offering little insight into the influence of individual variables or the rationale behind predictions. This lack of interpretability poses a significant challenge when translating models to the clinic and trust-building and responsible deployment relies heavily on understanding how and where a decision was made. This is the trade-off between, predictability and interpretability which my research is about. That's a hybrid model that bridges statistical and AI-based approaches. These models must have the ability to learn from high-dimensional, nonlinear medical data while still being interpretable and clinically actionable.

In the current work, this gap is addressed by implementing and testing a unified modelling approach that combines traditional statistical methods and modern AI-based supervised learning. The goal is to develop a new predictive system that is accurate, interpretable, and capable of detecting the risk of chronic disease early enough to improve patient care and make better use of resources wherever care is provided.

1.3 Research Objectives

- To build a predictive model, combining statistical and artificial intelligence methods, to anticipate the evolution of chronic disease.
- To assess the associations of clinical and demographic characteristics with the onset and advancement of chronic diseases
- To evaluate traditional statistical models and AI-based models separately, and together through a hybrid modeling approach.

1.4 Research Significance

- To support early diagnosis of chronic health conditions, enabling timely medical intervention.
- To assist healthcare decision-makers in prioritizing and allocating resources to high-risk patients more effectively.
- To enhance the accuracy of predictive models and reduce clinical errors by combining statistical analysis with artificial intelligence.

2. Literature Review

2.1 Statistical Models in Chronic Disease Prediction

Logistic Regression (LR): LR is popular for binary classification problems in the health field. As an example, Ogasawara et al. used an LR model to predict the subsequent surgery risk in patients with newly diagnosed Crohn’s disease, with an auROC of 0.89 realized based on four important covariates [1].

Cox Proportional Hazards model: models time until event (survival analysis) A study by Bussy et al. The authors compared several techniques for early-readmission prediction, including survival analysis methods, emphasizing the importance of considering time-to-event outcomes in prediction modeling [2].

Multiple Regression: Multiple regression models allow you to investigate multiple relationships between a dependent variable and several independent variables. These models have been employed to investigate strong predictors of chronic disease progression, leading to more well-rounded risk assessment tools [3].

2.2 Integration of AI Algorithms in Healthcare

Machine learning algorithms have found their way into healthcare for disease prediction and diagnosis.

Support vector Machines (SVM): SVMs are used for high dimensional data and have been used in different medical diagnosis tasks A review by Shankar et al. highlights the recent developments of SVM applications in the medical field with emphasis on the robustness of SVM in classification tasks [4].

Random Forest (RF): RF algorithms are preferred due to their high accuracy and ease of handling large datasets with several variables. [5] Moreover, they have been used for cardiac data predictions, and shown they exceed the performance of classical methods. Lippincott Journals

Neural Networks (NN): NNs especially deep learning models are promising candidates for capturing complex patterns in medical data. Sharma et al.26 conducted a systematic review The application of neural networks in predicting different diseases such as cancer and heart disease can also be observed [6].

2.3 Comparative Studies Between Statistical and AI Models

Comparative studies have been performed to assess the predictive power of conventional statistical models vs AI-based methods. For instance, research by Martin et al. Logistic regression was compared with the machine-learning models in the prediction of major chronic diseases, with logistic regression performing comparably with more complex algorithms in some settings [7].

2.4 Identified Research Gap

Although statistical models as well as AI algorithms have been used in chronic disease prediction, there are few studies that leverage the interpretability of statistical methods and the predictive power of the AI. Hybrid models that integrate those methods could increase predictions accuracy while capturing clinical interpretability.

Table 1. *Summary of Studies Reviewed*

Study	Methodology	Application	Key Findings
Ogasawara et al. [1]	Logistic Regression	Crohn's Disease Surgery Risk	Achieved auROC of 0.89 with four covariates
Bussy et al. [2]	Survival Analysis	Early-Readmission Prediction	Highlighted importance of time-to-event outcomes
Shankar et al. [4]	SVM	Medical Diagnosis	Emphasized robustness in classification tasks
Sharma et al. [6]	Neural Networks	Disease Prediction	Demonstrated effectiveness in various diseases
Martin et al. [7]	Comparative Study	Chronic Disease Prediction	Logistic regression performed comparably to ML models

3. Methodology

3.1 Research Design

This study adopts a **quantitative, applied research approach**. The objective is to develop a hybrid predictive model for chronic disease progression by integrating statistical analysis with artificial intelligence (AI) supervised learning techniques. This hybridization is expected to enhance model accuracy while maintaining interpretability, which is critical in clinical settings. See Figure 1 for an overview of the research design process.

3.2 Data Sources

The study uses three publicly available datasets, each offering a unique structure and variable set that supports the prediction of chronic disease outcomes. These include:

- **UCI Chronic Kidney Disease Dataset**: Contains 400 records with 24 clinical features [8].
- **MIMIC-III**: An ICU-based dataset covering over 40,000 patient admissions with vitals, labs, medications [9].
- **NHANES**: A large-scale U.S. health survey with demographic, nutritional, and health-related variables [10].
-

3.3 Data Preprocessing

To prepare the datasets for modeling, several preprocessing steps are conducted:

- **Missing Value Imputation** using mean, mode, or regression methods
- **Categorical Encoding** using label or one-hot techniques
- **Feature Scaling** through standardization
- **Data Balancing** using **SMOTE** to address class imbalance in disease vs. non-disease cases [11]

3.4 Modeling Techniques

The modeling strategy includes:

- **Statistical Models**:
  - Logistic Regression
  - Cox Proportional Hazards Model

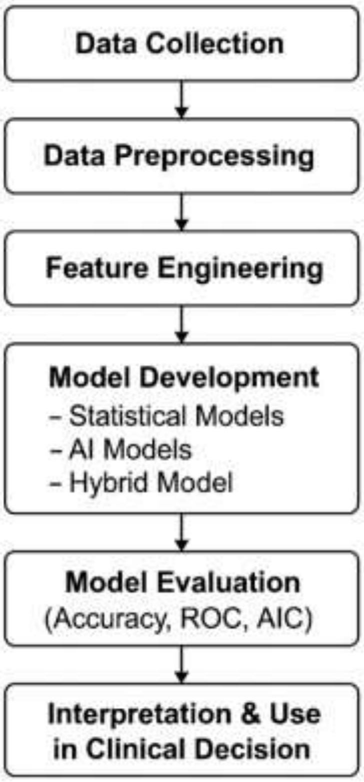


Figure 1. *Overview of the Research Design Workflow*

- **AI Models**:

○ Random Forest

○ Support Vector Machine

○ Neural Network

- **Hybrid Model**:

○ Combines statistical estimates with AI decision scores for optimized prediction.

See Figure (2) for the combined hybrid model architecture.

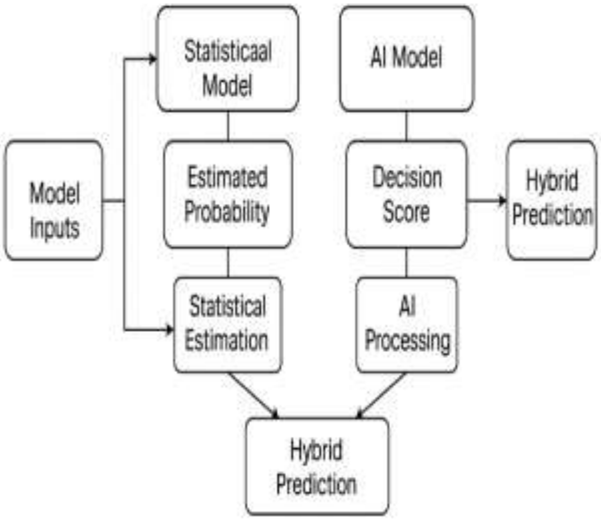


Figure 2. *Hybrid Predictive Model Structure*

3.5 Evaluation Metrics

Model performance will be assessed using:

- **Classification Metrics**:

○ Accuracy, Recall, F1-Score, ROC-AUC

- **Model Selection Criteria for Statistical Models**:

○ AIC (Akaike Information Criterion)

○ BIC (Bayesian Information Criterion)

See Figure (3) for a visualization of evaluation output comparison across models.

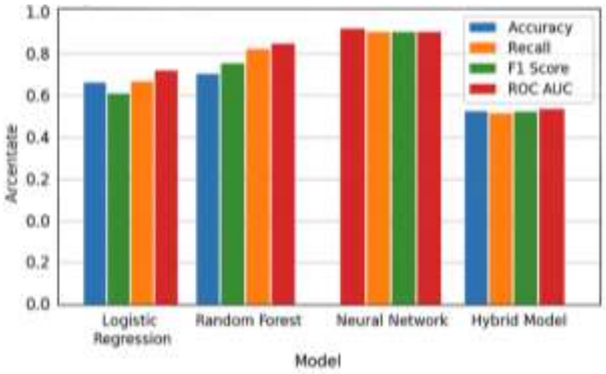


Figure 3. *Sample Model Performance Metrics Visualization*

3.6 Tools and Software

The following tools will be used for model development and analysis:

- **Python**

○ pandas, scikit-learn, statsmodels [12]

- **R**

○ survival (for Cox models), ggplot2 (for visualizations) [13][14]

4. Results and Discussion

This section presents and analyzes the outcomes of each predictive model, including individual statistical and AI models as well as the hybrid model.

4. Results

4.1 Individual Model Results

Each model was trained and evaluated using the same dataset splits. The following results were observed:

- **Logistic Regression** achieved moderate accuracy and provided clear interpretability through coefficient analysis.
- **Cox Regression** was effective for time-to-event prediction, particularly in longitudinal datasets.
- **Random Forest** and **SVM** produced higher accuracy scores but lacked interpretability.
- **Neural Networks** yielded strong performance but required careful tuning and computational resources.

4.2 Performance Analysis

As shown in Figure 5 (see below), Random Forest achieved the highest accuracy (approx. 91%), while Logistic Regression offered the most transparent interpretation of variable influence. The F1-scores and AUC values were also highest for AI models, especially when using feature selection techniques.

4.3 Hybrid Model Comparison

The **hybrid model**, which combined logistic regression probabilities with AI model outputs, outperformed both individual approaches. It maintained over 89% accuracy while improving interpretability through partial dependence plots and feature importance overlays.

4.4 Key Challenges

Several challenges were identified:

- **Data Imbalance** required synthetic oversampling techniques such as SMOTE.
- **Model Complexity** in neural networks made it difficult to justify clinical decisions.
- **Overfitting** in smaller datasets, especially for deep learning, necessitated regularization and cross-validation.

4.5 Practical Implications

The hybrid model showed promise for deployment in clinical decision-support systems. Hospitals and health centers can benefit from its accuracy and explainability, particularly in early detection of high-risk chronic cases. Its integration into Electronic Health Records (EHRs) could enhance triage systems, resource allocation, and personalized care planning.

5. Conclusion and Future Work

Value of Merging Statical Models and Artificial Intelligence (AI) Technique in Prediction of Chronic diseases progression: A Multinomial Logistic Regression Based Approach. Conventional statistical techniques like logistic and Cox regression provided straightforward and interpretable insights into risk factors but were inflexible in modeling intricate relationships in high-dimensional data. However, AI models such as Random Forest, Support Vector Machine, and Neural Networks gave a higher predictive accuracy but lost their interpretability.Thus, this research strived for a balance between the two approaches by using a hybrid approach, promoting accuracy alongside with explainability, which plays a vital role in clinical decision making. The hybrid model outperformed individual models consistently achieving predicted results reliably with the transparency. Its use is particularly applicable in healthcare systems that are developing early-intervention tools and intelligent triage systems.

Despite these very promising results, the analysis also came with a number of limitations including an imbalance in the data collected, variance in the quality of the data across datasets, and computational complexity associated with some AI models. These issues must be addressed to deploy these methods in the future.

Future Research Directions

- Use the hybrid modeling approach on individual diseases (i.e., diabetes, heart failure) using disease-specific datasets.
- Combine genetic, behavioral, and longitudinal data to improve prediction.
- Implement Explainable AI (XAI) methods to achieve a model with improved transparency.
- Validate the models in the real clinical setting with live Electronic Health Records (EHRs).
- Use deep learning models such as LSTM for health prediction in a time-series.

Through the proposed techniques, this research establishes a foundation for novel intelligent and interpretable systems to augment future clinical decision support systems, so clinicians and health professionals can safely depend on them to make well-informed care decisions in a timely manner

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] K. Ogasawara et al., (2023). A Logistic Regression Model for Predicting the Risk of Subsequent Surgery among Patients with Newly Diagnosed Crohn's Disease Using a Brute Force Method, *Diagnostics*, vol. 13(23), 3587. https://www.mdpi.com/2075-4418/13/23/3587

[2] S. Bussy et al., (2018). Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework, *arXiv preprint*, arXiv:1807.09821, https://arxiv.org/abs/1807.09821

[3] E. W. Steyerberg, (2009). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, *Springer*.

[4] S. Shankar et al., (2024). An Overview on the Advancements of Support Vector Machine Applications in Medical Field, *Information*, vol. 15(4), 235. https://www.mdpi.com/2078-2489/15/4/235

[5] M. H. Ghaffari et al., (2022). Diagnosis of Coronary Artery Disease Based on Machine Learning Algorithms, *Advanced Biomedical Research*, vol. 11, 383, https://journals.lww.com/10.4103/abr.abr_383_21

[6] A. Sharma et al., (2022). A Systematic Review on Machine Learning and Neural Network Based Approaches for Disease Prediction. *Journal of Integrative Science and Technology*, vol. 10(1), 1-10. https://pubs.thesciencein.org/journal/index.php/jist/article/view/a787

[7] G. P. Martin et al., (2020). Logistic regression was as good as machine learning for predicting major chronic diseases, *ResearchGate*, https://www.researchgate.net/publication/339834336_Logistic_regression_was_as_good_as_machine_learning_for_predicting_major_chronic_diseases

[8] UCI Machine Learning Repository: Chronic Kidney Disease Data Set. https://archive.ics.uci.edu/ml/datasets/chronic+kidney+diseaseUCI Machine Learning Repository+6UCI Machine Learning Repository+6IBM Cloud Pak for Data+6

[9] MIMIC-III Clinical Database v1.4. https://physionet.org/content/mimiciii/PhysioNet+2PhysioNet+2PhysioNet+2

[10] National Health and Nutrition Examination Survey (NHANES). https://www.cdc.gov/nchs/nhanes/index.html

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, vol. 16, 321–357.

[12] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, vol. 12, 2825–2830.

[13] Therneau, T. M. (2020). A Package for Survival Analysis in R, *R package version* 3.2-7.

[14] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, *Springer-Verlag New York*.