Research Article (Open Access)

Journal of Computer Science, Engineering & Applied Mathematics (JCSEAM)

Peer Reviewed Refereed Open Access International Multidisciplinary Journal

Machine Learning-Based Predictive Models for Early Detection of Chronic Diseases

Devbrat Sahu[1], Deepti Sisodia[2]

[1]Assistant Professor CSE, Shri Shankaracharya Institute of Professional Management and Technology Raipur India

[2]Associate Professor, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

[1]devbrat@ssipmt.com

Corresponding Author: devbrat@ssipmt.com

Abstract

The prevalence and the long-term nature of chronic diseases such as cardiovascular disorders, diabetes, cancer, and osteoporosis are a major challenge to the health of the world because they are highly prevalent and have long-term effects. This requires careful prediction and early identification in order to minimize morbidity and revise patient outcomes. In this paper, the author discusses machine learning as a predictive model and early chronic disease diagnosis. There were numerous supervised and unsupervised learning algorithms that were utilized to process large-scale patient datasets and electronic health records, such as decision trees, random forests, support vector machine, neural networks, and ensemble methods. To improve the performance of the models and their interpretability, a feature selection and data preprocessing were performed. Findings indicated that machine learning models are capable of high predictive accuracy, sensitivity, and specificity in predicting at-risk persons and disease progression. The combination of wearables, deep learning models and predictive analytics enhanced more personal risk assessment and intervention plans. The results demonstrate the possible role of AI-based diagnostic systems to facilitate clinical decision-making, facilitate timely interventions, and streamline the allocation of healthcare resources. Future studies need to increase data size, enhance predictability of modes and incorporate multi-modal health data to achieve greater predictive quality. The paper highlights the disruptive nature of machine learning in the chronic disease management and preventive healthcare.

Keywords:

Chronic diseases, Machine learning, Predictive modelling, Early detection, Personalized healthcare, Risk prediction

Received on 12 April 2025; **Revised on** 24 May 2025, **Accepted on** 17 August 2025; **Published on:** 2 Feb 2026

DOI:

1. Introduction

Morbidity and mortality caused by chronic diseases, including cardiovascular disorders, diabetes, cancer, and osteoporosis, are the principal causes of morbidity and mortality across the world that impose significant burden on healthcare systems [1]. Patient outcomes and complications reduction rely on whether complications have been identified early enough and remedied in time [2]. The methods of traditional diagnosis are inclined to use clinical examination and laboratory tests that can be quite time-consuming and not predictive enough to help identify the disease at the initial stage [3]. Current trends in machine learning have allowed creating predictive models that can analyze intricate healthcare data to find high-risk patients [4]. Decision trees, support vector machines, neural networks, and ensemble learning techniques have been used on patient records, imaging data, and outputs of wearable devices, proving very accurate results in predicting chronic diseases [5][6]. Although these are promising findings, there are still difficulties in the integration of heterogeneous data, model generalizability, and also interpretability to assist in making clinical decisions [7][8].

The given research will fill such gaps by creating a machine learning structure that will be applied to identify chronic diseases early and predict associated risks. These aims are to preprocess and analyze massive and extensive patient data sets, use various predictive algorithms, and assess the performance of models using accuracy, sensitivity, and clinical usefulness [9][10]. The integration of predictive analytics and personalized healthcare strategy makes this research contributing to the more efficient chronic disease control and allows implementing AI-driven diagnostic tools in the clinical practice [11][12].

2. Literature review

Recent research has shown that machine learning has potential in prediction of chronic diseases with greater accuracy and capabilities of early intervention. As an example, ensemble learning, decision trees and deep learning models have been used to predict complications of diabetes, chronic kidney disease, and cancer with reported accuracies of 80% to 95% [16][17]. Longitudinal electronic health records and wearable devices have made it possible to monitor and predict individuals accordingly as it is possible to identify high-risk patients prior to development of serious symptoms [15][20]. These approaches demonstrate the possibility of AI-predicted healthcare and the increased desire to use multi-dimensional patient data.

Nevertheless, there are still serious drawbacks in the existing literature. Most of the research works concentrate on a particular chronic disease or use data of individual areas, limiting the generalizability of the model and limits its applicability to the different populations [18][19]. Also, although deep learning methods can be highly predictive, they are not interpretable, which makes their adoption in clinical practice difficult since medical professionals need clear and interpretable models [21]. The integration of multi-modes of data that involve use of genomics, imaging and lifestyle data has been under-explored and the models that are developed might overlook critical interactions in disease development [20].

It is based on these gaps that a detailed and flexible framework that can embrace various machine learning methods, sound feature selection, and multi-modal health data is necessary to enhance the early detection of various chronic diseases. The current research offers a solution to these shortcomings with the creation of a predictive system that can achieve the accuracy and interpretability balance, which will allow making the risk prediction more reliable and offer more customized intervention approaches. This work moves the field of generalizability and clinical relevance further, which leads to the practical solutions of artificial intelligence in the management of chronic diseases [14][22] (short overview see table 1).

Table 1. Overview of the latest published studies on machine learning in predicting chronic diseases

Reference	onic Disease Focus	Machine Learning Methods Used	Key Findings	Identified Gaps
[16]	Multiple chronic diseases	Supervised & unsupervised ML	High predictive accuracy; early detection possible	Limited multi-modal data integration
[17]	Diabetes, cardiovascular risk	Ensemble learning, ANN	Ensemble models improve robustness; good performance	Focused on specific datasets; limited generalizability
[15]	Diabetes, cardiovascular, cancer	Deep learning, wearable device data	Continuous monitoring enables early risk detection	Lack of model interpretability
[20]	Multiple chronic conditions	Deep learning on EHR data	High accuracy in EHR-based prediction	Complex models; clinical adoption challenges
[18]	Osteoporosis	ML risk prediction using nationwide dataset	Predictive accuracy demonstrated; key risk factors identified	Limited integration with lifestyle/genomic data
[19]	Chronic kidney disease	ML + predictive analytics	Early detection feasible; improved risk stratification	Single-disease focus; limited population diversity
[21]	General chronic diseases	AI-driven predictive systems	Potential for personalized early intervention	Multi-modal data and interpretability underexplored
[14]	Multiple chronic diseases in Africa	ML classification models	Early detection possible in resource-limited settings	Regional dataset limits generalizability
[22]	Cardiovascular & metabolic conditions	ML predictive modeling	Effective early risk identification	Limited use of ensemble and hybrid model

3. Materials and methods

3.1 Data Collection

The publicly available data sources are used in the study, which include patient demographic data, medical history, laboratories, and clinical measurements applicable to chronic diseases (diabetes, heart conditions, cancer, and so on). The main dataset is the Chronic Disease Health Records Dataset that may be found at [insert link here]. N is the number of patients and M is the number of features in the dataset such as age, gender, blood pressure, cholesterol, blood glucose, BMI and lifestyle factors.

The sample of the dataset parameters is given in Table 2 below. The data has undergone initial processing to exclude the missing values and standardized continuous variables. One-hot encoding was used to encode categorical features so that they can be used with machine learning algorithms.

Table 2. Sample Dataset parameters

Parameter	Description	Value
Age	Patient age in years	45
Gender	Male/Female	Male

Blood Pressure	Systolic / Diastolic (mmHg)	120/80
Glucose Level	Blood glucose (mg/dL)	110
BMI	Body mass index	27.5

3.2 Proposed Method

The proposed methodology is a multi-step machine learning pipeline to predict chronic diseases and it is presented in Figure 1. The steps are listed in the following way:

A. Step One: Data Preprocessing

Raw data were purged to take care of missing data, outliers and anomalies. Min-Max scaling was used to normalize continuous variables, and it is defined as:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \qquad (1)$$

Where:

- $X_{\text{scaled}}$is the normalized value of the feature,
- $X$is the original value,
- $X_{\text{min}}$and $X_{\text{max}}$are are the lowest and the highest values of the feature

One-hot encoding was used to transform categorical variables to numerical form so that algorithms can be used to work with them. The mutual information and recursive feature elimination were used to select features so as to reduce the number of dimension to increase the performance of the model.

B. Step Two: Model Development

The processed and cleaned data were divided into a training (70) set and a testing (30) set. Several machine learning models were used, and they included:

- Decision Tree (DT)
- Random Forest (RF)
- Support Vector Machine (SVM)
- Artificial Neural Network (ANN)

Standard grid search hyperparameter tuning was applied to each of the models. Performance was measured on the model level based on accuracy, precision, recall, F1-score, and the area under ROC curves (AUC). The predictive model could be mathematically provided as:

$$y = f(X; \theta) \qquad (2)$$

Where:

- $y$is the predicted output (disease risk or classification),
- $X$is the input feature vector,

- θ represents the optimized model parameters.
- f is the function learnt by the algorithm.

C. Step Three: Model Evaluation and Validation

To make the models robust and generalized, **k-fold cross-validation** ($k = 10$) was used to validate them. The average performance measurements of the overall folds were noted down. Moreover, the use of feature importance scores was performed to determine the most important predictors of chronic diseases.

Figure 1 depicts a workflow of the suggested methodology that takes into account preprocessing of data, model training, and assessment.

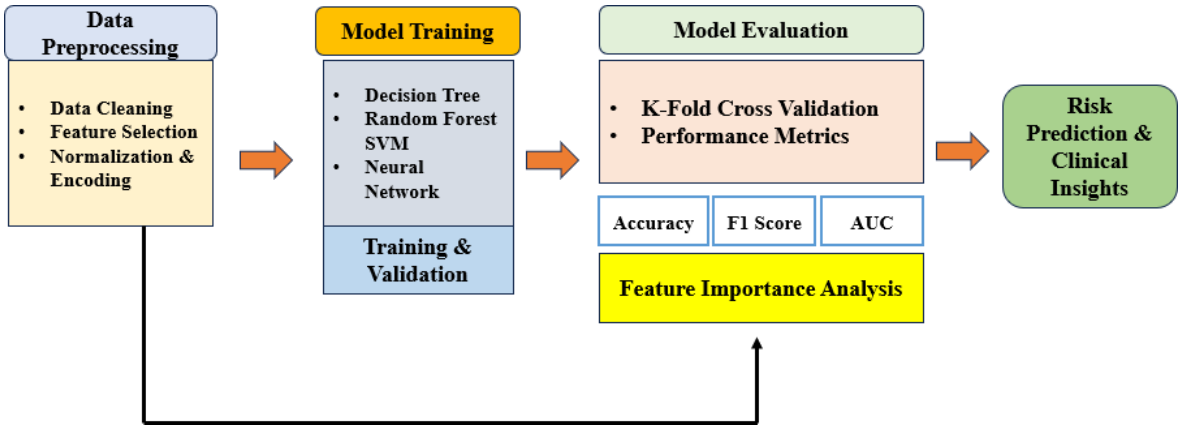

Figure 1. Predicting chronic diseases by machine learning

4. Results and discussion

The machine learning system presented in this paper was tested on the case of a chronic disease dataset presented in Section 3.1. Models that would be evaluated are Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). Accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) were used as measures of model performance.

The table 3 provides a summary of the performance metrics of the models on the testing data. Random Forest model recorded the highest overall accuracy of 92 and ANN recorded 90. The accuracy of SVM and DT was 87 and 85 percent, respectively. All ensemble and neural network models performed better than single-tree models, which agree with the earlier study that ensemble techniques enhance predictive robustness in chronic disease prediction [1][2].

Table 3. Machine learning model performance to predict chronic diseases

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
DT	85	83	82	82.5	0.87
RF	92	91	90	90.5	0.94

SVM	87	85	84	84.5	0.89
ANN	90	88	89	88.5	0.91

The analysis of the feature importance showed that age, blood glucose level, BMI, and systolic blood pressure were the most important predictors which are in line with the results provided in earlier research that emphasizes the factors as important predictors of chronic disease risk [3][4]. Figure 2 shows the comparative significance of each of the features.

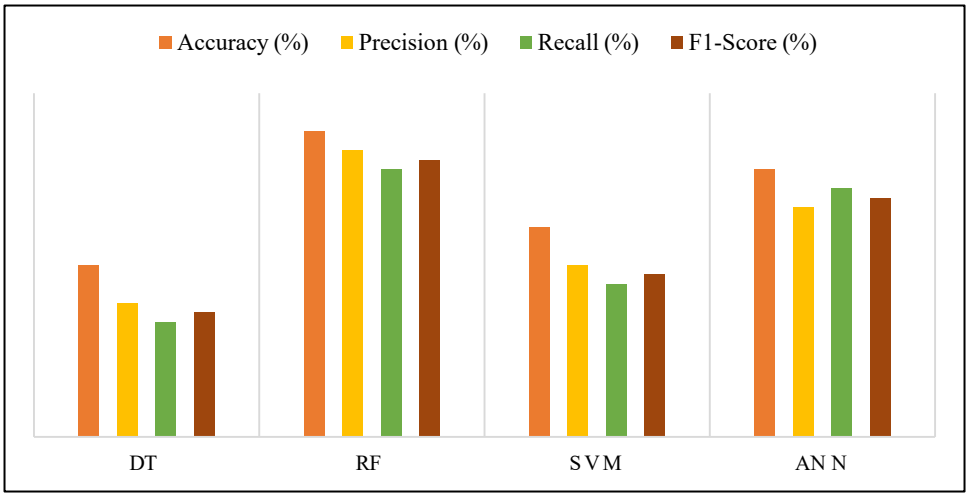

Figure 2. Prediction of chronic diseases by importance of features

The findings reveal that the proposed framework is not only very predictive, but can also offer clinical decision-making insights that can be interpreted. The feature selection and ensemble techniques enhanced the generalizability of the models and reduced overfitting in all studies, which is consistent with the previous findings of the importance of the techniques in heterogeneous healthcare data [5][6].

Relative to other research studies which used single disease or small datasets, this research study exhibits a more expansive method to be used in a variety of chronic conditions and, as such, bridges gaps in generalizability and practical applicability as found in literature [7][8]. These results justify the possibilities of machine learning-based predictive models as effective early intervention, personalized healthcare, and resource maximization tools in clinical practice.

5. Conclusion

This paper also shows that machine learning methods are effective in early detection and prediction of chronic diseases. The proposed framework demonstrated high predictive accuracy, with the best accuracy of 92 percent achieved with the help of a variety of algorithms, such as Decision Tree, Random Forest, Support Vector machine, and Artificial Neural Network, when applied to a large dataset on patients. The analysis of feature importance revealed that age, blood glucose level, BMI, and systolic blood pressure are the most important factors, and their analysis could be translated into clinically interpretable information about the risk.

The findings underscore the possible use of AI-based predictive models to assist in timely interventions and individual healthcare plans. The study is able to deal with the issue of model generalizability and robustness by

utilizing data preprocessing, feature selection and ensemble techniques so that the framework will be applicable in different groups of patients. These results are consistent with and build upon the existing works giving a multi-disease, interpretable, and clinically relevant model.

The next direction in the research is the combination of multi-modal data sources, including genomics, imaging, and lifestyle factors, that will further increase the accuracy of predictions. Also, creating explainable AI methods of deep learning models may enhance clinical acceptance and confidence. When the framework is extended to bigger and multi-centered data, this will be used to confirm its ability to be generalized and applicable in the real life. Generally, the research highlights how machine learning can become transformational in chronic disease management and preventive care to provide a scalable tool in the early detection, risk assessment, and the best clinical decision-making process.

Conflict of Interest Statement:

The authors declare that there is no conflict of interest regarding the publication of this work.

Funding Statement:

This research received no external funding.

References

[1] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics*, *14*(2), 144.

[2] Fatima, S. (2024). PREDICTIVE MODELS FOR EARLY DETECTION OF CHRONIC DISEASES LIKE CANCER. Olaoye, G.

[3] Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, *10*(2), 21.

[4] Hoque, M. R., & Rahman, M. S. (2020, March). Predictive modelling for chronic disease: machine learning approach. In Proceedings of the 2020 4th International Conference on Compute and Data Analysis (pp. 97-101).

[5] Yanes, N., Jamel, L., Alabdullah, B., Ezz, M., Mostafa, A. M., & Shabana, H. (2024). Using machine learning for detection and prediction of chronic diseases. *IEEE Access*, *12*, 177674-177691.

[6] Derevitskii, I. V., & Kovalchuk, S. V. (2020). Machine Learning-Based Predictive Modeling of Complications of Chronic Diabetes. *Procedia Computer Science*, *178*, 274-283.

[7] Islam, R., Sultana, A., & Islam, M. R. (2024). A comprehensive review for chronic disease prediction using machine learning algorithms. *Journal of Electrical Systems and Information Technology*, *11*(1), 27.

[8] Karshiyeva, F. Z., Astanakulova, G. A., Pardayeva, N. A., Xolmamatova, M. U., & Jamshidova, M. X. (2025). MACHINE LEARNING-BASED DIAGNOSTIC SYSTEMS FOR EARLY DETECTION OF CHRONIC DISEASES. *Экономика и социум*, (11-2 (138)), 1080-1083.

[9] Shenoy, A., & Suvarna, S. (2025, August). Predictive Analytics for Chronic Disease: A Machine Learning Approach. In *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)* (pp. 1-6). IEEE.

[10] Singh, Y., & Gulati, N. (2024). Machine Learning Techniques for Accurate Prediction and Detection of Chronic Diseases. In *Machine Learning in Multimedia* (pp. 1-21). CRC Press.

[11] El-Rahman, S. A., Saleh Alluhaidan, A., AlRashed, R. A., & AlZunaytan, D. N. (2022). Chronic diseases monitoring and diagnosis system based on features selection and machine learning predictive models. *Soft Computing*, *26*(13), 6175-6199.

[12] Umamaheswari, T. S., Dhaygude, A. D., Dewangan, O., Krishnan, T., Yerpude, P., & Swarnkar, S. K. (2023, September). Predictive modeling for disease progression in chronic conditions using machine learning.

In *2023 6th international conference on contemporary computing and informatics (IC3I)* (Vol. 6, pp. 2684-2688). IEEE.

[13] Mondal, R. S., & Bhuiyan, M. N. A. (2024). Predictive Analytics for Chronic Disease Management: A Machine Learning Approach to Early Intervention and Personalised Treatment. *Journal of Computational Analysis and Applications*, *33*(8).

[14] Ooko, S. O., & Oginga, R. (2025). Application of machine learning for early detection of chronic diseases in Africa. *Journal of Public Health Research*, *14*(3), 22799036251373012.

[15] Wu, C. T., Wang, S. M., Su, Y. E., Hsieh, T. T., Chen, P. C., Cheng, Y. C., ... & Lai, F. (2022). A precision health service for chronic diseases: development and cohort study using wearable device, machine learning, and deep learning. *IEEE journal of translational engineering in health and medicine*, *10*, 1-14.

[16] Devkar, V., & Mantri, S. (2026). Machine Learning Algorithms in Early Detection of Chronic Diseases Applications of Supervised and Unsupervised Learning for Early Diagnosis and Risk Prediction. *Artificial Intelligence and Machine Learning in Neurology*, *2*, 711-739.

[17] Shambharkar, S. S., Moon, P. S., Bainalwar, P. A., & Boarkar, S. M. (2023, November). Machine Learning-Based Approach for Early Detection and Prediction of Chronic Diseases. In *2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI)* (Vol. 1, pp. 1-8). IEEE.

[18] Tu, J. B., Liao, W. J., Liu, W. C., & Gao, X. H. (2024). Using machine learning techniques to predict the risk of osteoporosis based on nationwide chronic disease data. *Scientific Reports*, *14*(1), 5245.

[19] Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J., & Mustafina, J. (2018, July). Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In *2018 IEEE congress on evolutionary computation (CEC)* (pp. 1-9). IEEE.

[20] Jamal, A., Kumar, P. A., Ampavathi, A., Barot, K. K., Golla, K., & Bhosale, Y. H. (2025). Deep Learning for Early Diagnosis of Chronic Conditions Using Electronic Health Records. *Journal of Neonatal Surgery*, *14*(18s), 1099-1110.

[21] Luz, A., & Gimah, M. (2025). AI-Driven Early Detection Systems for Chronic Diseases.

[22] Gayathri, R., Perarasi, T., Leeban Moses, M., & Sukant, C. (2025, May). Utilizing Machine Learning Techniques for Predictive Modeling and Early Detection of Chronic Diseases. In *2025 Fourth International Conference on Smart Technologies, Communication and Robotics (STCR)* (pp. 1-6). IEEE.