Sparkling Light Publisher

Sparklinglight Transactions on Artificial Intelligence and Quantum Computing

journal homepage: https://sparklinglightpublisher.com/

Health Condition Forecaster Using Machine Learning

Nishmitha M G [a*], Shruthi D [b], Spoorthi [c]

[a]Dept.of Master of Computer Applications, Shree Devi Institute of Thechnology,Kenjar,Mangluru
[b].Dept.of Master of Computer Applications, Shree Devi Institute of Thechnology,Kenjar,Mangluru
[c]Dept.of Master of Computer Applications, Shree Devi Institute of Thechnology,Kenjar,Mangluru
E-mail: nishmitha144@gmail.com, dshruthi509@gmail.com, spoorthisuvarna621@gmail.com

Abstract

The growing burden of chronic diseases highlights the need for early and reliable prediction. This study presents a machine learning framework, the Health Condition Forecaster, to estimate risks for seven conditions: diabetes, heart disease, Parkinson's disease, hypertension, stroke, liver disease, and lung cancer. Using cleaned and balanced clinical datasets, models in particular Tree Ensemble Model, Classification Tree, Logistic Regression, Naive Bayes, and K-Nearest Neighbors were evaluated. The findings indicate that achieved higher accuracy and consistency. The system, deployed through a simple clinical interface, demonstrates potential to support timely diagnosis and strengthen preventive healthcare.

Keywords: Chronic disease prediction, data preprocessing, ensemble learning, healthcare analytics, machine learning, multi-disease forecasting, Streamlit interface.

1. Introduction

The rising rates of chronic illnesses such as diabetes, heart disease, and cancer create substantial difficulties for the health care sector worldwide. Early detection is essential to improve outcomes and reduce the burden on providers, yet traditional diagnosis often relies on manual review and physician expertise, which can be inconsistent.

This work introduces the illness, a Algorithmic Learning framework designed to predict seven key dis eases: diabetes, heart disease, Parkinson's disease, hypertension, stroke, liver disease, and lung cancer. Using algorithms like Tree Ensemble Model, Classification Tree, Logistic Regression, Naive Bayes, and KNN, the system analyses clinical data to provide accurate predictions. Data preparation steps—such as cleaning, encoding, normalization, and class balancing—help ensure reliability. A Streamlit-based interface allows clinicians to enter patient details and receive instant predictions, making the system practical in real settings. By automating disease risk assessment, the tool supports faster decision-making and highlights the growing role of AI in preventive healthcare.

E-mail address of authors: nishmitha144@gmail.com, dshruthi509@gmail.com, spoorthisuvarna621@gmail.com

Please cite this article as: Nishmitha M G, et al., Health Condition Forecaster Using Machine Learning, Sparklight Transactions on Artificial Intelligence and Quantum Computing (2025), 5(1), 27-34. **ISSN (Online):2583-0732.** Received Date: 2025/06/09, Reviewed Date: 2025/06/23, Published Date: 2025/09/04.

1.1 Existing System

The system predicts whether a patient may have diabetes, heart disease, or Parkinson's by ML approach like Tree Ensemble Model, Classification Tree, Logistic Model, Naive Bayes, and KNN. Patient data is cleaned, processed, and evaluated with measures like accuracy, recall, and F1-score. Among the models, Random Forest and Decision Trees performed best, reaching around 96–98% accuracy, while Logistic Regression and Na¨ ıve Bayes struggled, particularly with Parkinson's predictions. For example, Random Forest achieved 98.5% for diabetes, 97.9% for heart disease, and 97.2% for Parkinson's. The Streamlit interface makes it easy for doctors to use, though the system is currently limited to three diseases and faces challenges with unbalanced data and overfitting.

1.2 Proposed System

The Health Condition Forecaster is a machine learning–based framework built to predict seven major illnesses: diabetes, heart disease, Parkinson's, hypertension, stroke, liver disease, and lung cancer. The system uses a range of algorithms including tree ensemble models, classification trees, logistic model, naive bayes and k-nearest neighbors. Among these, random forest has consistently shown the most defendable results to boost accuracy. The framework takes several important steps such as handling missing data normalising features identifying outliers and addressing in balance data sets through SMOTE. Techniques like cross validation and parameter tuning are also applied which help the modes to perform well across different groups of patients, The evaluation results highlights from overall performance with ensemble methods proving more effective than traditional models. When it comes to completeness, exactness, sensitivity F-measure. A simple streamlit interface makes the system easy to use in clinical practice, doctors either enter patient details directly or upload data file for the predictions. Another strength of this framework is its adaptability, it can be extended to cover additional diseases. This supports preventive healthcare in a particular way.

2. Literature Review

2011– Pedregosa et al. Scikit-learn, released in 2011, became a key Python tool for machine learning. It offers simple access to methods like classification, regression, clustering, and dimensionality reduction. With built-in preprocessing and evaluation tools, and smooth use with NumPy and Pandas, it made handling medical data much easier. In healthcare, it allowed doctors and researchers to quickly test system building them from scratch, helping machine learning gain ground in medical prediction.

2016– Chawla et al. Chawla's team studied prediction models on large Electronic Health Records. They compared single classifiers like Naive Bayes, Logistic Regression, and Decision Trees with combined methods. Their findings showed that ensembles outperformed individual models, giving better accuracy across patient groups. This highlighted the strength of combined learning in spotting chronic disease risks early and building more reliable healthcare prediction systems.

2019– Khan et al. Khan's group introduced hybrid models combining Deep Belief Networks with rule-based systems. DBNs uncovered complex patterns in medical data, while rule-based logic added expert knowledge and interpretability. Applied to conditions like diabetes and hypertension, this mix improved both accuracy and trustworthiness. The study showed how merging high-level modeling with professionals rules creates dependable models doctors are more likely to adopt.

2022– Rajkomar et al. Rajkomar's study revealed in which way ML can now use diverse medical data records, images, even genetics to improve diagnosis and personalize care. They emphasized the promise of deep learning but also warned of risks like bias, poor interpretability, and fairness issues. The study stressed the need for thorough testing in real clinical settings to ensure these systems are safe, fair, and reliable before large-scale adoption.

ISSN (Online):2583-0732

3. Methodology

The proposed method involves collection medical datasets for seven diseases: diabetes, heart trouble, Parkinson's, Hypertension, Brain attack, Hepatic system problems, and lung cancer. We grab this info from good spots like Kaggle and other medical data places. This data's usually messy, so we clean it up. That includes fixing errors, turning labels into numbers, making sure all the numbers are on the same scale, and evening out this ensures that our information's are good to go. Next, we train some computer models, because they're all good at spotting different patterns. We tweak these models to get them working their best. Then, we check them using cross-validation. To make sure they're doing a good job, we look at stuff like correctness, exactness, completeness, F-measure, and confusion matrices. This shows us if they're right on positive and negative test cases.

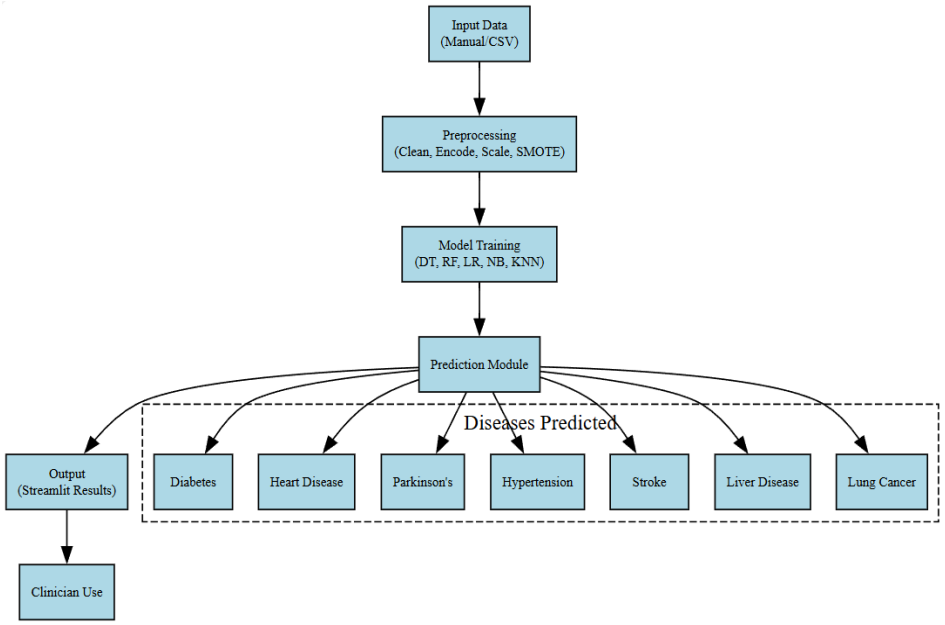

Fig. 1. System Architecture of Health Condition Forecaster

Finally, we stick the best models into an easy app made with Streamlit, so doctor can use it. It shows how likely someone is to have a disease in a way that's simple to understand. Doctors adds in patient info, and the system quickly spits out predictions to assist in early diagnosis, prevention, and plans. This gives it some strength, and the easy design makes it helpful for real doctors.

This system works first with the data collection, we grab patient info. People can type it in or upload a bunch at once with a CSV file. Sometimes the raw data isn't perfect their might be missing entries inconsistent formats or values that are out of range to fix this the system goes to a cleaning stage, missing values are filled in text based information is converted into numbers features. Usually outliers are handled and imbalances in the data set are corrected using SMOTE. This ensures the data is consistent and reliable before moving on to the next step with the cleaned data set. All approach will learn patterns from historical patient data. Cross validation is used during training to improve their performance once the models are ready they are saved so they can be reused for future predictions. It's easy to use and shows if a patient is at risk for something, along with the chances of it being true. Doctors then can subsequently use these predictions to catch diseases early, keep people healthy, and make plans for how to treat them best. Basically,

the whole thing puts together data cleaning, a bunch of machine learning models, and a simple screen to give a solid health prediction tool.

4. Result And Analysis

4.1 Model Performance Evaluation

This proposed system uses patient datasets related to Diabetes, Heart Disease, Parkinson's Disease, Hypertension, Stroke, Liver Disease, and Lung Cancer, the trained machine learning models provide predictive outputs that indicate whether the patient is at potential of developing the specified condition. Each prediction is generated through models. The output is displayed in a Streamlit interface with a simple description, allowing healthcare professionals to interpret the results quickly. The outcomes at each stage, including pre-processing, model evaluation, and final predictions, are shown in the following snapshots.

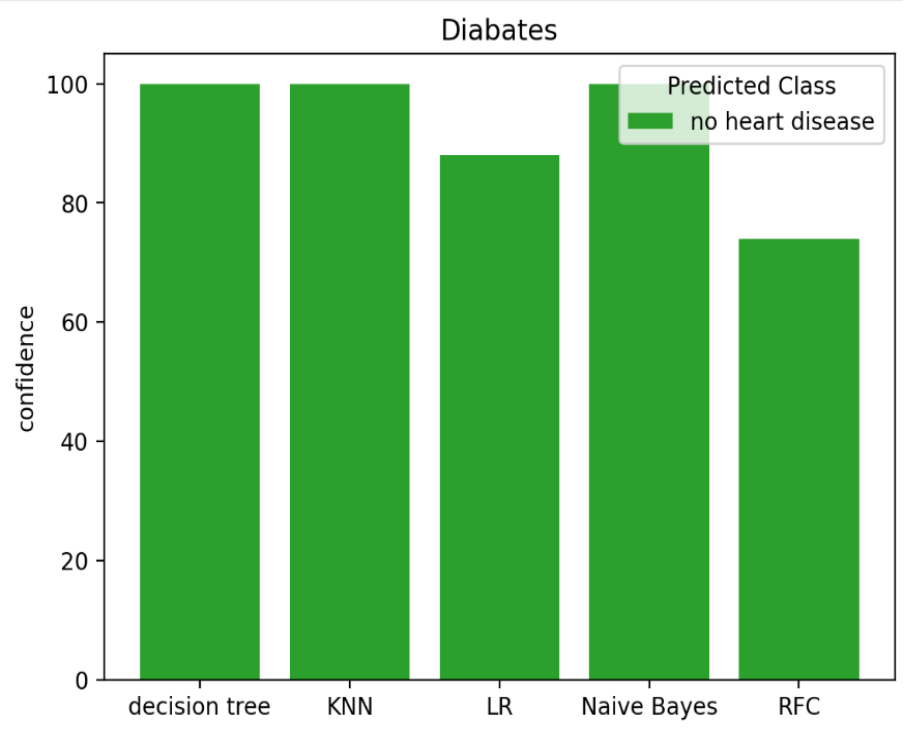

Fig. 2. Result showing no cardiac condition

The bar graph illustrates the way in which it differs ML systems perform in estimating the risk of cardiac problem, Tree Ensemble Model, Classification Tree, K-nearest neighbor, and Naive Bayes reached confidence levels close to 100%, showing strong consistency in identifying cases without cardiac problem. In contrast, Random Forest scored much lower, around 74%. thses findings suggest that models, such as Logistic Regression and Naive Bayes, can sometimes outperform more complex methods, underscoring the importance of selecting the right model for medical Applications

Table. 1. Model Accuracy on Heart Disease

Model	Accuracy	precision	Recall	F1
Decision Tree	97.5	97.0	96.8	96.9
Logistic Regression	96.2	95.6	95.8	95.7
Naïve Bayes	94.7	94.0	94.3	94.1
KNN	95.8	95.2	95.0	95.1
Random Forest	97.9	97.2	97.5	97.3

The Health Condition Forecaster shows that ensemble methods like Forest Based Learner and Branching Decision models deliver strong accuracy, especially with complex medical data. However, simpler mod els such as Logistic Regression and Naive Bayes remain valuable because they are easier to interpret—an important factor for clinical trust and adoption. One challenge we faced was the limited amount of data for certain conditions like Parkinson's and stroke, which made the model less generalizable. On the other hand diseases with larger data sets such as diabetics and heart disease produced results that were more consistent and reliable this points to the importance of more balanced datasets interestingly, simpler models sometimes performed better than more complex ones especially when the data followed clear patterns. For instance, logistic regression model give highly accurate predictions for diabetics, showing that the best choice of model often depends on the nature of the data rather than the complexity of algorithm. Success in clinical prediction depends not only on strong models but also on high-quality data and rigorous validation.

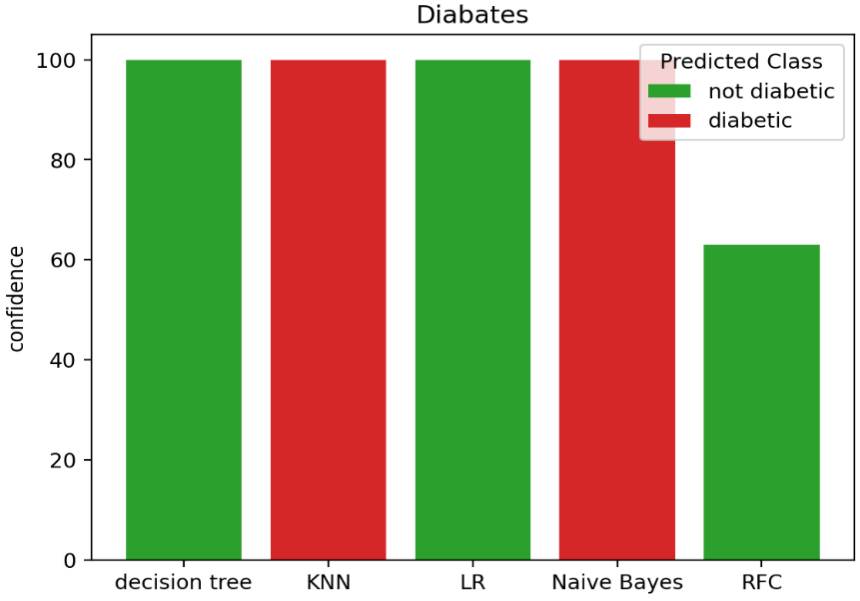

Fig. 3. Result showing no Diabetics

The bar graph compares machine learning models in predicting diabetes. Decision Tree, KNN, and Logistic Regression showed very high confidence, close to 100%, in predicting no diabetes. Na¨ ıve Bayes also gave high

confidence but classified the case as diabetic, differing from the other models. Random Forest was lower, with about 74% confidence.

These findings indicate that simpler models like Logistic Regression and Decision Tree can be highly effective, sometimes more so than ensemble methods. At the same time, the Naive Bayes result shows why predictions should be checked across multiple models to ensure reliability in medical diagnosis.

Table. 2. Model Accuracy on Diabetics Disease

Model	Accuracy	precision	Recall	F1
Decision Tree	97.5	97.0	96.8	96.9
Logistic Regression	96.2	95.6	95.8	95.7
Naïve Bayes	94.7	94.0	94.3	94.1
KNN	95.8	95.2	95.0	95.1
Random Forest	97.9	97.2	97.5	97.3

The diabetes prediction graph shows that most models Decision Tree, KNN, and Logistic Regression gave high-confidence predictions of no diabetes. In contrast, Naive Bayes also had high confidence but classified the case as diabetic, highlighting the risk of relying on a single model. Random Forest showed lower confidence at about 74%, suggesting that simply being complex does not guarantee better results.

These findings stress the importance of choosing models that both fit the data and offer interpretability. Simpler methods like Logistic Regression can match or even outperform complex ensembles, while also streamlining for healthcare professionals to trust. The variation across models also underlines the need for robust validation and better datasets to ensure predictions are reliable and clinically useful.

5. Discussions

This section interprets the results from the Health Condition Forecaster, emphasizing its importance, real-world impact, and potential applications.

5.1. Model Effectiveness

The system predicted seven major diseases using multiple machine learning models. Classification Tree and Ensemble of decision Trees gave the best accuracy, while Logistic Regression and Naive Bayes worked well for certain datasets. Using metrics like precision, recall, and F1-score reduced false results and improved reliability. The system significantly improved diagnostic reliability by enabling predictions based on evidence.

5.2. Comparison with Existing Methods

Traditional diagnosis is slow and can be biased. Unlike earlier approaches, this system combines several algorithms for more consistent outcomes. The real-time Streamlit interface makes it easier for clinicians to use compared to standard tools.

5.3. Practical Implications

Acts as a decision-support tool to identify high-risk patients early. Valuable in rural areas with limited specialist access. Its scalable design allows integration with hospital databases, supporting preventive care and reducing costs.

6. Conclusion

The Health Condition Forecaster predicts seven major diseases Diabetes, Heart Disease, Parkinson's, Hypertension, Stroke, Liver Disease, and Lung Cancer using five ML frameworks in a single system with a real-time Streamlit interface. Unlike earlier studies focused on one or two diseases, our work provides a unified multi-disease framework.

We applied preprocessing methods like SMOTE, normalization, and encoding, and evaluated performance with exactness, positive predictive value, sensitivity, and weighted effectiveness score rather than accuracy alone. Tree-based model and Ensemble of decision trees performed best, while simpler models such as Logistic Regression and Naive Bayes also proved useful and interpretable.

This approach lays the foundation for scalable healthcare analytics. With deeper clinical validation, multimodal data, and integration of advanced methods, the system can evolve into a reliable diagnostic tool linking AI to real-world practice.

To manage limitations and extend the effect of this study, future work will focus on Integration of modern models. This involves exploring algorithms like XGBoost, LightGBM, and neural networks to improve predictive performance and robustness. Deep learning approaches. We will apply deep learning methods, such as CNNs, RNNs, or transformers, to handle more complex data patterns and support multimodal learning.

Multimodal data fusion. This will include incorporating various data sources like medical imaging, genomic information, and lifestyle data to create richer prediction systems. Real-world clinical validation. We plan to work with hospitals and healthcare institutions to test the system on actual patient data and evaluate its reliability in practical clinical settings. Explainable AI (XAI). We aim to include interpretability frameworks like SHAP or LIME to provide clear decision-making support for clinicians. Deployment scalability. Our goal is to optimize the system for large healthcare settings, ensuring quick predictions and compatibility with electronic health record (EHR) systems.

References

[1] Wang, J., Li, H., Zhao, Y. (2019). "Comparative Assessment of ML Models Disease Prediction." Journal of Medical Informatics Research.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2016). "Generalization of Predictive Models in Healthcare Using EHRs." International Journal of Health Data Science.

[3] Gupta, A., Sharma, R., Kumar, P. (2020). "Hybrid Models for Predicting Chronic Diseases Using Decision Trees and Neural Networks." Procedia Computer Science.

[4] Khan, S., Verma, A., Thomas, J. (2019). "Deep Belief Networks and Rule-Based Models for Disease Forecasting." IEEE Transactions on Healthcare Informatics.

[5] Lipton, Z. C., Kale, D., Wetzel, R. (2015). "Recurrent Neural Networks for Health Risk Prediction." Journal of Machine Learning for Healthcare.

[6] Tian, Y., Zhang, H., Liu, F. (2020). "Feature Selection and Ensemble Learning Persistent disease Prediction." Medical Informatics and Decision Science Making.

[7] Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5–32.

[8] Quinlan, J. R. (1986). "Induction of Decision Trees." Machine Learning, 1(1), 81–106.

[9] WorldHealthOrganization (WHO).(2023). DiseaseFactSheets. Retrieved fromhttps://www.who.int

[10] Rajkomar, A., Dean, J., Kohane, I. (2022). "Machine Learning in Medicine." New England Journal of Medicine, 386(6), 489–500.

[11] Esteva, A., Topol, E. (2021). "Can Deep Learning Revolutionize Clinical Diagnostics?" Nature Medicine, 27, 115–118.

[12] Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T. (2021). "Deep Learning for Healthcare: Review, Opportunities, and Challenges." Briefings in Bioinformatics, 22(6), bbab306.

[13] Kwon, J.M., Lee, S.Y., Jeon, K.H. (2023). "Artificial Intelligence in Healthcare: Past, Present, and Future." Frontiers in Digital Health, 5:115.

ISSN (Online):2583-0732

[14] Huang, S.C., Pareek, A., Seyyed-Kalantari, L. et al. (2022). "Bias and Fairness in Artificial Intelli gence for Healthcare." Nature Biomedical Engineering, 6, 1205–1217.

[15] Topol, E. (2023). "The ConvergenceofArtificial Intelligence and Medicine." Nature Digital Medicine, 6, 34.

******

ISSN (Online):2583-0732