Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Prediction of Lifestyle Diseases Using Random Forest

Dr. M. Sangeetha[1], S. Harshitha[2], K. A. Aberami[3], S. Abivarshini[4], S. Akalya[5]

[1]Assistant professor Department of Computer Science and Engineering,
V.S.B Engineering College, Karur, Tamil Nadu
[2,3,4,5] Department of Computer Science and Engineering,
V.S.B Engineering College, Karur, Tamil Nadu

Abstract

Health issues such as high blood pressure, diabetes, and heart disease are now some of the biggest challenges facing people worldwide. These problems not only shorten lives but also reduce overall quality of life. Since family members often share daily habits, it is important to look at health risks not just for individuals, but for households as a whole. In this study, we introduce a prediction system that uses the Random Forest algorithm to estimate the risk of lifestyle-related diseases for different family members. The system collects key health indicators—like blood pressure, heart rate, and body temperature—along with time stamps, and then uses them to provide personalized risk assessments. The results are grouped into three levels: Normal, Caution, and Critical. This helps detect potential health issues early and makes it easier to take quick action. In addition to individual reports, the system also gives a general picture of the health of the family giving informative insights into common trends and threats. We experimented to demonstrate that the accuracy, stability and generalization of Random Forest is superior to that of traditional single algorithms. The framework is also streamlined and flexible and can be used in a broad variety of healthcare settings and health care environments, particularly, where resources are limited. On the whole, this article proves that machine learning can be a useful tool in preventive medicine, as it helps track health indicators in the earlier stages of life, motivate to make more healthy choices, and minimize the long-term consequences of lifestyle diseases.

Keywords: Chronic Illnesses, Random Forest, Health Monitoring, Risk Prediction, Preventive Care.

1. Introduction

Non-communicable diseases (NCDs), often referred to as lifestyle-related illnesses, are strongly influenced by long-term behaviours such as diet, physical inactivity, chronic stress, and irregular sleep [16]. Over the last two decades, these conditions have overtaken infectious diseases as the leading global health challenge. Heart disease, diabetes and stroke are some of the leading causes of death in the world today. As per the recent estimates provided by health organizations around the world, over two in three deaths are currently associated with NCDs with the forecasts indicating that this will only increase in the future [16]. In addition to the impact on life expectancy, the conditions not only decrease the quality of

life but also diminish the level of productivity and put a considerable financial burden on them, making it not merely a medical issue but also a socio-economic hazard.

To address this increasing challenge, researchers and policymakers are laying stress on prevention and early intervention as opposed to laying stress on treatment alone. Preventive healthcare is concerned with the early signs of warning, which would help decrease the hospitalization number and lower the medical cost of the long-term medical expenses. Machine Learning (ML) provides tools with significant capabilities in this direction since it can identify hidden health data trends that can be overlooked using conventional methods [8]. With the implementation of ML-based approaches, the healthcare systems can decrease the reactive models in favor of proactive ones and save lives and resources eventually. It has been common to use traditional algorithms like Decision Trees and Logistic regression to predict risks of such diseases as diabetes and cardiovascular complications [1], [2]. The benefits of their simplicity and interpretability can be regarded as benefits, but these models presuppose the linear nature of the relationships, which restricts their competence in the complex situation in health. Support Vector Machines (SVMs) are more accurate and require higher computational resources and parameters adjustment [7], making them less practical in time-sensitive systems. Ensemble methods, especially the Random Forests (RF), address most of these limitations by combining the decisions made by several decision trees. Preventive healthcare is concerned with the early signs of warning, which would help decrease the hospitalization number and lower the medical cost of the long-term medical expenses. It contributes to the accuracy, less overfitting and better results of working with incomplete or noisy data [3], [4], [14].

Although there has been significant advancement in predictive healthcare development, majority of the solutions developed are household-monitors, as opposed to monitoring individuals. Nonetheless, there is a tendency that families have similar risk factors, including dietary habits, living environments, and exposure to stress. Also, not many models include more long-term aspects, such as a comparison of the vital signs per week or automated preventive alerts, which are both necessary to transition to proactive healthcare [10], [13]. To fill in these gaps, this paper presents a predictive system built using the random forest that will be specific to families. The framework allows the input of several members of the household simultaneously, time-stamped entries, and weekly consolidated summaries. In this way, it translates predictive healthcare in models that are individual centred to more comprehensive and inclusive solutions that are community based.

2. Literature Review

A major cause of mortality in the earth is non-communicable diseases (NCD) which by the statistics provided by the World Health Organization facilitates close to 74 percent of death in the globe. Among them, the cardiovascular diseases (CVDs), diabetics, strokes, and hypertension impose a large burden of morbidity and mortality. Besides causing high mortality rates, these conditions result into productivity losses, high healthcare expenses and long-term disabilities. This transforms the NCDs beyond being a medical problem to a more general social and economic crisis [15]. This problem worldwide has stimulated the exploration of data-driven and machine-learning (ML) methods of early risk detection and preventive medicine [8],[9].

A. Traditional Predictive Approaches: The cardiovascular and diabetes risk assessment of traditional statistical and machine learning models, including Logistic Regression (LR) and Decision Trees (DT), are performed [1],[2]. LR is simple and understandable, though it presupposes a linear association between variables, which restricts its application in a variety of populations [1]. On the same note, DTs are simple to visualize and understand but they are usually victimized by overfitting thus less effective when used on new groups of patients [2]. In order to fill these gaps, the Support Vector machines (SVMs) were created as a more trustworthy classifier in healthcare information [6]. SVM has demonstrated a great classification accuracy and overfitting resistance [7]. They are however not practical in real time clinical practice due to the necessity of careful kernel selection, extensive parameter tuning and are computationally intensive, particularly in resource constrained or time sensitive applications [6],[7].

B. Ensemble and Random Forest Techniques: Due to the shortcomings of individual classifiers, ensemble techniques were used. Random Forest (RF) is one of such algorithms that is effective and it was first developed by Breiman [3]. RF operates through combining several decision trees constructed on random samples of data and attributes, which decreases variance and increases generalization [3],[4]. RF also deals with noise, it can process high dimensional data and does not have the issue of overfitting as does single models [5].

RF has also got popularity because of its interpretable nature. The measures of variable importance, which are included in the framework by Breiman, assist clinicians and researchers with determining the most important risk factors in medical data [3]. Such openness is a requirement to any establishment of trust in predictive healthcare solutions. RF is always found to be more accurate, sensitive, specific and AUC compared to such models as Logistic Regression, Naive Bayes, Decision Trees and AdaBoost in the comparative studies of RF [8],[11],[14].

C. Uses in the Cardiovascular and Lifestyle Diseases: Numerous researches have established that RF outperforms in healthcare. As an example, when it comes to arrhythmia detection with wearable ECG signals, better RF models are characterized by high levels of generalization in different public datasets [9]. RF models outperformed Logistic Regression and SVM in predicting the risk of CVD and diabetes in large-scale studies, and thus they can be used in risk triage dashboards [11]. The studies on hypertension and metabolic syndrome also indicate that RF models are more accurate and still good at coping with imbalanced and noisy data [10],[13].

Early signs of diabetic complications have also been detected by RF, patients classified by their level of risks, and suggestions as to lifestyle changes supported through RF [12]. Studies done in Elsevier journals as well as IEEE Access journals all indicate that ensemble techniques are more effective in diagnostics when compared to single classifiers [9],[11].

D. Remote Patient Monitoring (RPM) and Trends: Predictive models have their place, but their effectiveness in the practical healthcare system considerably relies on the Remote Patient Monitoring (RPM) platforms [10]. RPM has become an important component of the modern preventive healthcare due to the emergence of wearables, mobile health apps, and IoT-enabled medical devices [12]. Surveys of massive RPM studies show that linked devices significantly reduce hospital readmission, improve the management of chronic diseases, and allow patients to become the owners of their health [12]. Nevertheless, such systems are not immune to certain challenges, such as data overload among the

clinicians, inconsistent quality of information, connectivity and reimbursement complexities [12]. New innovations have implied that the combination of correct predictive models with summarization and alerting systems is more effective as compared to transmitting raw data. The National Early Warning Score (NEWS2) is a system that is common in interpreting vital signs and raising alerts [13]. Combined with RF-based risk assessment, such systems can give context-sensitive alarms, decrease alarm fatigue and guarantee timely responses [13].

E. Research Gap: Although the progress has been made, the shortcomings remain. A great number of ML-for-CVD studies evaluate algorithms on offline datasets which are static [9]. Not many studies go to such an extent of following several patients or families where common lifestyle and environment issues matter [10]. Moreover, longitudinal monitoring, e.g., the analysis of the weekly alterations in the vital signs, is a process that is seldom taken into consideration, although trend analysis becomes essential in proactive healthcare [13].

Our study fills this gap by developing an RF based framework that:

Produces understandable and interpretable classification results (Normal, Warning, Serious).Facilitates monitoring for multiple household members, accounting for family-level risk factors.

Provides weekly comparison reports and alert systems based on thresholds to meet clinician needs.

This approach broadens predictive healthcare beyond individuals, supporting community-driven preventive and inclusive monitoring strategies.

3. Existing System

Several conventional techniques of machine learning are increasingly being utilized for prediction of disease given health data pertaining to patients. Decision Tree and Logistic Regression are examples of two common models of machine learning because they are relatively easy to implement and model output is easily interpretable. Decision Tree classifiers create a model by iteratively partitioning a dataset into subsets based on values of the features and produces a prediction as a class. Decision Trees can successfully model structured data, but the model is prone to overfitting, particularly when datasets are sparse or noisy, and this will negatively impact generalizability.

Logistic Regression is a technique for prediction of binary classification problems for which a combination of features is used to determine the probability of a given outcome. Algorithms such as Logistic Regression yield an acceptable level of prediction accuracy with both binary outcome datasets but require the data to be linearly separable. As a result, Logistic Regression function loses predictive capability when datasets have many confounding factors or the underlying relationships are complicated or nonlinear.

Other machine-learning approaches such as k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM) have been discussed in past studies. k-NN is a simple and intuitive method of classification but is computationally intensive with large datasets and the choice of distance metric is consequential. SVMs are effective classifiers in high-dimensional spaces but require consideration to identify the most suitable kernel and can be more difficult to interpret in national health care systems or consumer-based community health initiatives providing medical care. Furthermore, health monitoring solutions based on Internet of

Things (IoT)-enabled technologies have been utilized to provide an ongoing assessment of patient vitals by employing wearable devices and sensor networks. Although these solutions allow for the collection of data in real-time, they require existing hardware infrastructure, stable internet access, and can be expensive to develop and deploy in rural or resource limited areas.

In light of these serious limitations across all the previous approaches (overfitting, rigidity to complex data, excessive dependence on hardware, and limited multi-user monitoring), a more generalized and scalable solution is required, which will address these gaps through the ensemble approach of employing the Random Forest algorithm, performed within a household-level monitoring solution.

4. Problem Identification

The emergence of non-communicable diseases linked to lifestyle habits- hypertension, diabetes and cardiovascular disease has emerged as a big global health issue of concern. These diseases are highly preventable, but predetermined by the factors that can be altered, such as nutritious diet, physical activity, stress, and irregular clinical attendance. Regardless of the developments in diagnostics, a number of people can be unaware of the fact that their health is getting worse until severe complications set in. Traditional disease prediction systems are usually narrow. Majority of them are implemented on a single-patient basis and demand sophisticated infrastructure like IoT accountable equipment or constant patient observational measures, rendering them inappropriate in rural or resource-restricted areas. In addition, such classic machine-learning methods as Logistic Regression and Decision Trees fail at handling nonlinear, imbalanced, or noisy healthcare data. Current systems hardly facilitate health tracking at a multi-user or household level which is very useful in determining common risk factors among family members. In the absence of this, it is hard to identify the trends of health or preventive requirements on a community-wide basis.

These shortcomings demonstrate the need to have a more precise, consistent and scalable predictive model that is available and resource efficient. Hence, the proposed study suggests a framework based on the Random Forest that will allow the input of multiple users, the identification of risks in an effective way and monitoring the households. Ensemble learning method will improve prediction effectiveness, reduce overfitting and will also provide proactive health care in a heterogeneous population.

5. Methodology

A. Data Collection: The system will be implemented in the settings where the internet access or medical equipment can be restricted. It does not depend on automated sensors, but rather gathers data by manual input by the user or the healthcare personnel. All members of the family give standard health indicators, which are usually associated with cardiovascular and metabolic diseases:

Blood Pressure (BP): 1.Systolic and Diastolic (mmHg)

Pulse Rate: Beats per minute (bpm)

Body Temperature: It is recorded in either °C or °F.

The system is easier to customize in a house that does not have special equipment since the manual input is used and yet provides meaningful health data.

B. Data Pre-processing: The information must first be subjected to a number of additional preparation tasks to increase reliability and consistency before it can be analyzed**.**

Cleaning the data: All the entries that were not complete, were duplicated, or were obviously erroneous were eliminated to avoid mistakes**.**

Normalization: The numerical values are brought to a comparable range with all the features having similar values which minimizes the chances of a larger number dominating the analysis**.**

Labelling: There are three health states into which every record is classified**.**

Normal (0): The scores are within normal ranges**.**

Caution (1): Minute anomalies, which need to be observed carefully.

Serious (2): Important deviations and should be addressed by a professional medician**.**

This is done so that the dataset is homogenous, and can be trained in an effective model.

C. Model Selection: In order to classify, the system employs the algorithm of Random Forest (RF). The decision is inspired by a number of strengths that render RF applicable to health-related applications:

It gives powerful predictive power over most single-model strategies.

It also utilizes several decision trees and this prevents overfitting which is important particularly when the data is manually gathered.

It is able to record nonlinear or complex relationships between such variables as blood pressure, temperature, and pulse. It works reasonably well even in case the classes (Normal, Warning, Serious) are disproportionately distributed. Due to these characteristics, RF also provides both the reliable performance and the results that can be more easily interpreted in a healthcare environment**.**

D. System Architecture (User Based): The general process taken by the system is as under summary:

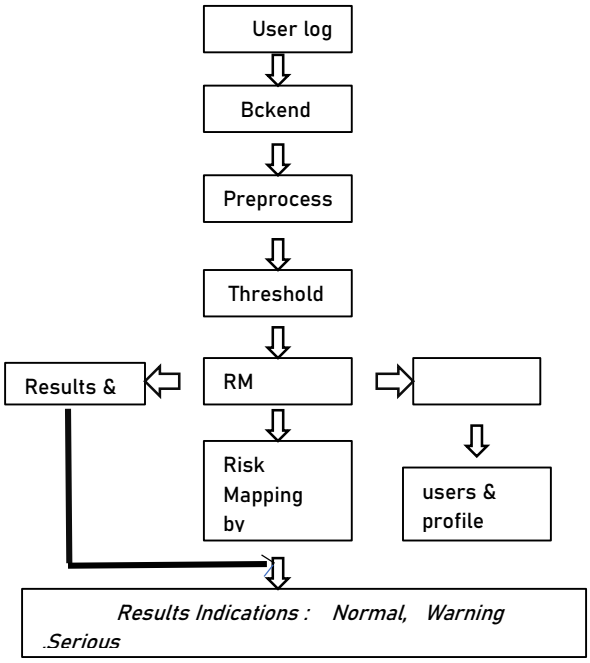
Input stage: The vital signs of every member of the family are entered by the users or health workers.

Pre-processing phase: The raw data is washed, scaled, and coded to the established categories**.**

Model implementation: The Random Forest model classifier examines every input and defines a risk probability.

Output categorization: There is the classification of people as normal, warning and serious**.**

Report generation: Individual health reports are generated on each member and also on the family.



Trend monitoring: A comparison between the present and previous records will be done every week, and in case of any significant changes in the health status, alert will be given.

6. Proposed System

A. Multi-Member Monitoring and Weekly Alerts *(Proposed Innovation)*

The system is crafted in such a way that it monitors the well-being of the whole family and does not concentrate on individual people. It does the weekly records, the comparison of changes, and alerts in case of unusual patterns.

B .Observation of More Than Two Members.

All family vital signs could be entered and processed. Processing data at once increases the speed of processing, and it is impossible to forget about the health update of a particular individual.

C. Weekly Change Detection

The readings of every week are contrasted with those of the earlier week. In order to determine significant changes, the system employs the common vitals defined thresholds:

Systolic BP: a fluctuation of over =10 mmHg.
Diastolic BP: fluctuation of over -5 mmHg.
Pulse Rate: changes over a range of above 10bpm.
Body Temperature: deviations of above or below +0.5 o C.

A reading that exceeds these limits is raised as a possible alarm.

D. Alert Mechanism

There are two cases of generating alerts:
In case the machine-learning model indicates that a change in risk status (e.g., Normal to Warning, or Warning to Serious).
As soon as one of the essential signs exceeds a predetermined limit.
There are also easy suggestions in the system, like the recommendation of closer attention or the recommendation of medical assistance.

E. Family Health Overview

Lastly, the system creates an all inclusive report. The report will provide the present state of each member, differences on a weekly basis, and the number of alerts generated, which will give the families a clear view of their general health patterns.

7. Result and Discussion

A. Dataset and Training/Test Split

The data that was used in carrying out this study comprises the vital signs including blood pressure, pulse rate, and body temperature of various people during a time span. These readings can be directly connected with the lifestyle-based health risks, such as hypertension, diabetes, and cardiovascular diseases.

The data was separated into two parts in order to train and evaluate the model:

Training set (70%)- This is the set used by the Random Forest classifier to generalise patterns and relationships between the input features (BP, pulse, temperature) and the health categories (Normal, Warning, Serious).

Testing set (30%)- Stored apart to test the performance of the model on records that were not seen before.

This division makes sure that the model possesses sufficient information to learn well besides offering a reasonable evaluation of the way it will execute on new data. The 70 /30 split is popular in machine learning since it provides a balance between learning and trustworthy assessment.

B. Evaluation Metrics

Accuracy alone is not an entire picture of performance, particularly in the case of healthcare where false alarms or missed cases can be very severe. This is why a number of other measures were used:

Accuracy The total percentage of correct predictions.

Precision- Of all the cases that the model has marked as positive, what proportion were correct. This minimizes unwarranted notifications.

Recall (Sensitivity) of all the at-risk cases that are really at-risk, what proportion the model identified. The recall should be high, as it may be risky to miss a dangerous patient.

F1-Score – A balanced score

Table 1: Comparison of the Model Performance.

that is a combination of precision and recall, which is helpful when the dataset is skewed between classes.

Formulas

Accuracy = (TP+TN) / (TP+FP+TN+FN) * 100

Precision = (TP) / (TP+FP)

Recall = (TP) / (TP+FN)

F1 = 2*((Precision*Recall) / (Precision + Recall))

Where:

TP (True Positives): The number of positive cases that were predicted correctly.

TN (True Negatives): These are the cases of negative that have been correctly predicted.

FP (False Positives): Predicted false positive.

FN (False Negatives): Predicted as negative when it was not.

A combination of these metrics gives a complete performance analysis. They guarantee that the model is right and reliable to use in medicine.

C. Model Comparison

Random Forest model proposed was compared to the Decision Tree and Logistic Regression classifiers.

Table 1: Comparison of the model performance.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	81%	78%	76%	77%
Logistic Regression	84%	82%	80%	81%
Random Forest	92%	90%	91%	91%

Observation:

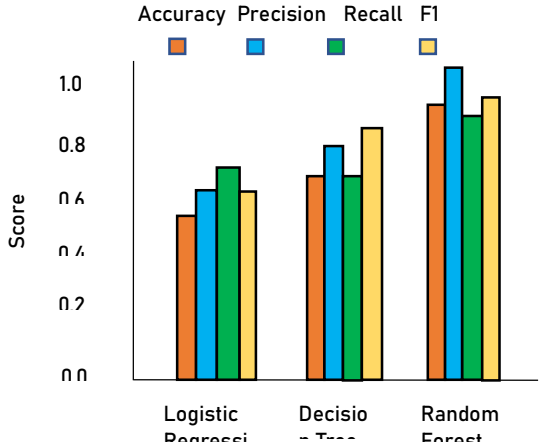


Fig. 2

Fig. 1 illustrates that The Decision Tree was relatively successful but it had a problem of overfitting and lowest recall, implying that it missed a lot of cases that needed a follow up.

Logistic Regression was more generalized than the Decision Tree but it also had a problem in the borderline cases of warning.

The Random Forest had the best scores in all measures. Its ensemble nature that implicates the combination of various decision trees contributed to minimizing errors and greater dependability.

Most importantly, Random Forest reached a 91% recall, showing its strength in detecting individuals at risk — an essential feature in preventive healthcare.

D. Multi-Member Weekly Difference Results

Member	BP((Systolic)	Pulse	Temp	Risk	Change Alert
A	120-130	78-85	36.8-37.5	Normal-Warning	Generated
B	140-142	90-88	377.0-37.2	Warning-Warning	Non Generated
C	118-119	72-74	36.5-36.6	Normal-Normal	Non Generated

Fig. 3

The ability of the system to monitor health changes in a group of multiple family members at a time is one of the critical contributions of the system. The analysis of the differences in readings of each individual

on a weekly basis is conducted to compare the current data and earlier data of an individual in the week. When the change exceeds a safe range, the system gives out an alarm. As an illustration, when the blood pressure of the individual soars as compared to the previous week, the model would upgrade the risk level to a warning or serious level and inform caregivers.

This feature provides:

Early Detection- An increase in blood pressure or temperature can be identified early before it develops.

Preventative Response - This can be addressed at an earlier stage by the family and the doctor and it is less hazardous to health.

Individual Insights - The progress of every member is monitored individually, which enables monitoring on the family level but still in an individual package.

E. Analysis

The general results affirm that the Random Forest model is suitable to make use of. It had 92% accuracy and 91 F 1 -Score exceeding the baseline models. The system is also capable of predicting a risk as well as giving real-time warnings when used in combination with threshold-based alerts, which is of great value to preventive care.

The multi-member monitoring also increases the scope of its application that can be applied not only to individual patients but also to households as well as community levels. The system also allows the efficient utilization of the resources possessed by healthcare workers as it helps them target high-risk individuals.

To conclude, the combination of Rand Forest classification and weekly difference monitoring would be a practical and efficient solution to identify and prevent lifestyle diseases at an early stage.

8. Future Improvements

Though the predictive system based on the application of the Random Forest tool proves to be rather promising in terms of its accuracy and practicality when predicting the risks of a lifestyle disease, there is still a range of aspects in which the tool could be improved:

A. Wearable Device Integration and IoT.

Currently, the model relies on the manual input of users with their health parameters. This process may also become automated in the future by connecting the system to the IoT-based wearables, such as smartwatches, digital blood pressure monitors, or fitness trackers. The constant data gathering would help to minimize the human error and decrease the activities in the manual work and provide the opportunity to monitor the health in real time.

B. Health Parametric Expansion.

At present, there are only inputs on blood pressure, pulse rate, and body temperature admitted. It is possible to add more parameters in order to enhance the predictive capacity like:

Cholesterol and blood glucose.

Sleep duration and quality

Stress-related markers

Exercise habits and eating behaviors.

Combining these variables would enable the system to bring about more detailed insightful health information.

C. Individualized Predictive Analytics.

The versions of the future can be programmed to match the health profile of the particular user and the history of his or her lifestyle. More complicated methods, like reinforcement learning or hybrid machine learning methods, may continuously update the model and provide more accurate and tailored risk forecasts as time goes on.

D. Longitudinal and Community-Level Analysis.

Future systems will not only be weekly comparisons but may also incorporate long-term trend analysis, i.e. monthly or yearly, in order to identify slow health reductions. Development of community level dashboards to detect the trend of disease in a region or a family to enable the public health spheres to focus on interventions is possible.

E. Clinical Decision Support System (CDSS) Integration.

Liaison with the healthcare providers can facilitate an easy integration into the CDSS platforms. The physicians might be provided with more brief risk reports and it would help make faster diagnoses and more preventive recommendations.

F. Data Privacy and Security Improvements.

Since the system is expanding and handling confidential health data, it needs to have strong privacy protections. End-to-end encryption, strong authentication, and access control will assist in bringing the platform in line with international healthcare standards like the HIPAA and GDPR.

9. Conclusion

This paper shows that machine learning grounded in Random Forests can be a very useful tool in the estimation of the risks of developing lifestyle diseases with the help of easy vital signs, such as blood pressure, pulse rate, and body temperature. The accuracy of the proposed model was 92 which is higher than the baseline methods of Decision Tree (81) and Logistic Regression (84). These findings underscore the benefits of ensemble algorithms such as the Random Forest that are less prone to overfitting and they are capable of dealing with noisy/imbalanced medical data. One of the important contributions of this study Is the planning of a family-centered monitoring system. This framework supports a variety of users at home and has weekly difference tracking with threshold alerts unlike the traditional systems that evaluate health at the individual level. This kind of functionality facilitates the detection of abnormal patterns at early stages when they can be transformed into severe health complications. In addition to personal gains, this is a strategy that helps community-based applications since lifestyle and environmental factors usually have a major influence on making groups of people vulnerable in concert. Notably, the system has been optimized to low-resource environments by relying on manual data input rather than high-tech IoT devices, which is appropriate in rural locations and in families with a limited access to modern healthcare technologies. The system provides useful, cost-effective, and doable insights by integrating machine learning predictions with rule-based alerts.

In the future, the system can have several directions to expand its scope:

IoT integration: The connection to intelligent devices to automate the data collection process and provide the ability to monitor data in real-time.

Broader health indicators: Incorporate lifestyle and behavioural measures like diet, exercise, sleep and stress to make more detailed predictions.

Mobile and cloud computing: Developing mobile apps and cloud computing with the capability to handle a high number of users and centralize data storage.

Personalized analytics: Using more capable learning models that will adapt predictions to personal health history and lifestyle.

To sum it up, the present piece of work is a good demonstration of how machine learning specifically the Random Forest can be instrumental in preventive healthcare. The system can help solve the problem of the escalating burden of lifestyle-related disease through supporting families by performing the analysis on the family level and providing a low-cost delivery, as well as supplementing the system to prevent the disease on an earlier stage.

Heavy contribution in preventive healthcare. The system provides a scalable and viable solution to the rising burden of lifestyle-related diseases by assisting in the early detection, analysis at the family level, and low-cost implementation.

Reference

1. A. Al-Mohaimeed, "Prediction of cardiovascular disease using logistic regression," Journal of Health Informatics, vol. 12, no. 3, pp. 215–222, 2020.
2. K. Vasanth and R. Kumar, "Diabetes risk prediction using decision tree algorithms," International Journal of Computer Applications, vol. 975, pp. 8887, 2019.
3. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
4. A. Liaw and M. Wiener, "Classification and regression by random Forest," R News, vol. 2, no. 3, pp. 18–22, 2002.
5. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2012.
6. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
7. N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge, U.K.: Cambridge Univ. Press, 2000.
8. S. Patel and H. Patel, "A review on machine learning based health prediction systems," International Journal of Computer Science and Information Technologies, vol. 6, no. 1, pp. 80–84, 2015.
9. P. S. Priyanka, "Role of machine learning in preventive healthcare," Healthcare Informatics Research, vol. 26, no. 3, pp. 210–219, 2020.
10. A. Sharma and R. Gupta, "Household health monitoring using IoT and machine learning," IEEE Access, vol. 8, pp. 12367–12375, 2020.
11. R. J. P. Velmurugan, "A survey on non-communicable diseases and machine learning applications," International Journal of Advanced Computer Science and Applications, vol. 11, no. 5, pp. 546–552, 2020.
12. Y. Bengio, A. Courville, and I. Goodfellow, Deep Learning. Cambridge, MA: MIT Press, 2016.
13. K. N. Rajasekar and B. S. Ramesh, "Vital signs monitoring and disease prediction using machine learning," International Journal of Recent Technology and Engineering, vol. 7, no. 6, pp. 415–420, 2019.
14. H. Zhang, "The optimality of naive Bayes," Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, FL, USA, 2004, pp. 562–567.
15. World Health Organization, "Noncommunicable diseases," WHO Fact Sheet, Apr. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
16. World Health Organization, "Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2019," Geneva: WHO, 2020.